LEONARDO LOPES BHERING

MAPEAMENTO GENÉTICO EM FAMÍLAS SIMULADAS DE IRMÃOS COMPLETOS

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Genética e Melhoramento, para obtenção do título de *Doctor Scientiae.*

VIÇOSA
MINAS GERAIS – BRASIL
2008

Ficha catalográfica preparada pela Seção de Catalogação e Classificação da Biblioteca Central da UFV

T
B575m
2008

Bhering, Leonardo Lopes, 1980-
Mapeamento genético em famílias simuladas de irmãos completos / Leonardo Lopes Bhering. – Viçosa, MG, 2008.
viii, 150f. : il. ; 29cm.

Orientador: Cosme Damião Cruz.
Tese (doutorado) - Universidade Federal de Viçosa.
Inclui bibliografia.

1. Genética molecular - Métodos de simulação. 2. Amostragem (Estatística). 3. Melhoramento genético. 4. Genética molecular. 5. Genética. I. Universidade Federal de Viçosa. II.Título.

CDD 22.ed. 572.8

LEONARDO LOPES BHERING

MAPEAMENTO GENÉTICO EM FAMÍLAS SIMULADAS DE IRMÃOS COMPLETOS

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Genética e Melhoramento, para obtenção do título de *Doctor Scientiae*.

APROVADA: 28 de fevereiro de 2008.

Prof. José Marcelo Soriano Viana (Co-orientador)	Prof. Pedro Crescêncio Souza Carneiro (Co-orientador)
Prof. Adésio Ferreira	Prof. Márcia Flores da Silva Ferreira

Prof. Cosme Damião Cruz
(Orientador)

AGRADECIMENTOS

À Deus, por sempre se fazer presente me iluminando todos os dias;

Aos meus pais, José Antonio Bhering e Maria do Carmo, e meu irmão Elder, pelo apoio, amizade e compreensão em todos os momentos da minha vida;

Ao professor Cosme Damião Cruz, pela paciência, amizade, confiança, incentivo e pela orientação ao longo de toda minha vida universitária; além do exemplo de dedicação aos estudos, na busca incansável de novos conhecimentos científicos.

À Universidade Federal de Viçosa, pela oportunidade de realização deste curso.

À Coordenação de Aperfeiçoamento do Pessoal de Nível Superior (CAPES) e ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), pelas bolsas de estudo concedidas.

Aos professores de graduação e de pós-graduação, pela atenção, pela disponibilidade e pelos ensinamentos transmitidos.

Aos professores Marcelo Soriano e Pedro Crescêncio, pela co-orientação, pelos ensinamentos transmitidos em suas disciplinas, e pela enorme contribuição e sugestão a serem adicionadas a tese.

Ao amigo e professor Adésio Ferreira, e sua esposa professora Márcia Ferreira, pelas contribuições e participação na banca de Tese.

Aos grandes amigos do laboratório de Bioinformática, em especial ao Edmar, Caio, Márcio, Tati, Willian, Adésio e aos pilantras, pela amizade, apoio e companheirismo.

Às secretárias do curso de pós-graduação em genética e melhoramento, Rita e Rose, pelo apoio, dedicação, atenção e amizade.

A todos que, direta ou indiretamente, contribuíram para a realização deste trabalho.

Finalmente, a quem estiver valorizando este trabalho através de sua leitura e utilização de alguma forma.

BIOGRAFIA

Leonardo Lopes Bhering filho de José Antonio Bhering e Maria do Carmo Lopes Bhering, nasceu em Viçosa, Minas Gerais, em 30 de Janeiro de 1980.

Iniciou-se o ensino fundamental em 1987 na Escola Estadual Coronel Antônio da Silva Bernardes, em Viçosa, Minas Gerais. No ano de 1991 transferiu-se para o Colégio Equipe, em Viçosa, Minas Gerais, onde concluiu o ensino fundamental e também o ensino médio no ano de 1997.

Em Fevereiro de 1999, ingressou no curso de Agronomia na Universidade Federal de Viçosa, em Viçosa, Minas Gerais, onde se graduou em 30 de Janeiro de 2004, obtendo o título de Engenheiro Agrônomo.

Em seguida ingressou no Programa de Pós-Graduação em Genética e Melhoramento de Plantas, na Universidade Federal de Lavras, em Lavras, Minas Gerais, onde no dia 23 de Fevereiro de 2006 obteve o Título de Mestre em Ciência.

No dia 16 de Fevereiro de 2006, ingressou no Programa de Pós Graduação em Genética e Melhoramento na Universidade Federal de Viçosa, em Viçosa, Minas Gerais, onde no dia 28 de Fevereiro de 2008 obteve o Título de Doctor Scientiae.

SUMÁRIO

RESUMO

BHERING, Leonardo Lopes, D.Sc., Universidade Federal de Viçosa, fevereiro de 2008. **Mapeamento genético em famílias simuladas de irmãos completos.** Orientador: Cosme Damião Cruz. Co-orientadores: José Marcelo Soriano Viana e Pedro Crescêncio Souza Carneiro.

O mapeamento genético facilita o trabalho de melhoramento, uma vez que uma ou mais marcas podem estar associadas a genes controladores de características qualitativas e quantitativas (QTL), podendo ser utilizada na seleção assistida. O estabelecimento de mapas genéticos em populações exogâmicas apresenta determinadas complicações não encontradas ao utilizar populações endogâmicas. Dentre elas tem-se variação no número de alelos segregando por loco e, geralmente, há desconhecimento da fase de ligação. Em famílias de irmãos completos diferentes graus de informações são observados na progênie, podendo esta ser do tipo completamente informativa ou não completamente informativa. Outro fator de fundamental importância para se obter dados consistentes que resultem em mapas genéticos fidedignos é o tamanho da amostra ou da população de mapeamento. Assim, objetivou-se com este trabalho fornecer subisídios para melhor entender o processo de estimação da freqüência de recombinação entre diferentes configurações gênicas presentes em família de irmãos completos, além de estimar o tamanho ideal de população para a obtenção de mapas de ligação confiáveis neste tipo de delineamento genético. Para isto foi simulado um genoma constituído de três grupos de ligação, sendo cada um constituído de 11 marcas moleculares multialélicas, codominantes, eqüidistantes, com saturação de 10 cM. A partir deste genoma foram simulados dois cenários, sendo um com genitores completamente informativos e outro com genitores formados aleatoriamente. Foi ainda considerado diferente tamanhos de população para cada cenário. Sendo estas populações constituídas de 100, 200, 300 e 400 indivíduos e 100 repetições de cada tamanho amostral. Após a obtenção de todas as estimativas da freqüência de recombinação concluiu-se que para locos completamente informativos pode-se calcular a freqüência de recombinação entre pares de locos tanto a partir da freqüência gamética de cada genitor quanto a partir da freqüência genotípica conjunta da progênie. Para locos parcialmente

informativos a obtenção da freqüência de recombinação a partir da freqüência genotípica conjunta foi mais apropriada. Concluiu-se ainda que, para populações completamente informativas um tamanho populacional de 200 indivíduos seria o suficiente para resgatar as informações originais. Contudo, para a população não completamente informativa seria necessária a utilização de uma população maior, constituída de 600 indivíduos.

ABSTRACT

BHERING, Leonardo Lopes, D.Sc., Universidade Federal de Viçosa, February, 2008. **Genetic mapping in simulate full siblings family.** Adviser: Cosme Damião Cruz. Co-Advisers: José Marcelo Soriano Viana and Pedro Crescêncio Souza Carneiro.

The genetic mapping facilitates the breeding work once one or more marks of the genotype can be associated to controlling genes of qualitative and quantitative traits (QTL). The establishment of genetic maps in exogamic populations presents certain complications not found when endogamic populations are used. One of these complications is the variation in the numbers of alleles segregating per locus and the linkage phase which is usually unknown. In full siblings families it is possible to find different degrees of information in the progeny, which can be completely informative or not informative. The objective of this work was to discuss the differences in the mapping involving populations with different degrees of information, through simulation, and to generate and to analyze data starting from simulation of the genome and of populations, and based in these simulated data to evaluate the optimum size of populations for study of genetic mapping of full siblings' families. It was simulated a genome with three linkage groups, each one with 11 molecular marks which were multi-allelic, codominants, equidistants, with saturation of 10 cM. Starting from this genome, two situations were simulated, one with informative genitors, and other with genitors formed randomly. Together was generated samples had 100, 200, 400 and 600 individuals with 100 repetitions were accomplished by sample. Once all the recombination frequencies were estimated it was found out that in completely informative locus, the recombination frequency can be calculated among equal pairs of locus starting from the gametic frequency of each genitor or starting from the genotipic frequency of the progeny. For partially informative locus, the recombination frequency was more appropriate estimated starting from the united genotipic frequency. In completely informative populations, an optimum size of 200 individuals would be enough to rescue the original information, however, for the population not completely informative it would be necessary a larger population, constituted of 600 individuals.

1. INTRODUÇÃO GERAL

Com o desenvolvimento da biologia molecular surgiram novas ferramentas que possibilitaram a construção de mapas genéticos mais acurados. Os marcadores de DNA têm permitido a construção de mapas genéticos para várias espécies vegetais e animais. Tais mapas podem atingir alto grau de saturação devido a disponibilidade de grande número de marcas genéticas, as quais têm a vantagem de não serem influenciadas pelo ambiente, além de serem altamente polimórficas.

No estudo genético de determinada característica é de interesse do pesquisador conhecer o número de genes e alelos envolvidos no controle da sua expressão, a localização e posição relativa desses genes nos cromossomos, assim com a sua relação com outros genes. Nesse contexto, os mapas genéticos são de fundamental importância, uma vez que permitem a visualização, mesmo que de forma relativa, da organização dos genes nos cromossomos.

No mapeamento genético podem ser usados dois tipos básicos de populações segregantes, as oriundas de cruzamentos controlados e as exogâmicas. O primeiro tipo engloba as populações F_2, retrocruzamentos, F_1"pseudo-testcross", duplo-haplóides e linhas endogâmicas recombinantes "RILs – Recombinat Inbred Lines", e no segundo tipo as famílias de irmãos completos (FIC), meio-irmãos (FMI) e populações oriundas de intercruzamentos.

O estabelecimento de mapas genéticos em populações exogâmicas apresenta determinadas complicações não encontradas ao utilizar populações endogâmicas. Dentre elas tem-se variação no número de alelos segregando por loco e, geralmente, há desconhecimento da fase de ligação. Em famílias de irmãos completos, diferentes graus de informações são observados na progênie, podendo esta ser do tipo completamente informativa ou não completamente informativa. Uma vez obtida as freqüências de recombinações entre as marcas, inicia-se a construção do mapa genético.

Na construção de mapas genéticos são usados modelos estatísticos para descrever sistemas genéticos e biológicos reais. No entanto, esses sistemas são complexos, impossibilitando a inclusão de todas as variáveis nos modelos utilizados. O mapeamento genético facilita o trabalho de melhoramento, uma vez que uma ou mais marcas podem estar associadas a genes controladores de características qualitativas e quantitativas (QTL), podendo ser utilizada na seleção assistida. Porém, um dos fatores de fundamental

importância para se obter dados consistentes que resultem em mapas genéticos fidedignos é o tamanho da amostra ou da população de mapeamento.

Assim, objetivou-se com este trabalho fornecer subsídios para melhor entender o processo de estimação da freqüência de recombinação entre diferentes configurações gênicas presentes em família de irmãos completos, além de estimar o tamanho ideal de população para a obtenção de mapas de ligação confiáveis neste tipo de delineamento genético.

CAPÍTULO 1

MAPEAMENTO GENÉTICO EM FAMÍLIAS DE IRMÃOS COMPLETOS – ESTUDO DE CASO

VIÇOSA
MINAS GERAIS – BRASIL
2008

RESUMO

BHERING, Leonardo Lopes, D.Sc., Universidade Federal de Viçosa, fevereiro de 2008. **Mapeamento genético em famílias simuladas de irmãos completos (1): Estudo de caso.** Orientador: Cosme Damião Cruz. Co-orientadores: José Marcelo Soriano Viana e Pedro Crescêncio Souza Carneiro.

O estabelecimento de mapas genéticos em populações exogâmicas apresenta determinadas complicações não encontradas ao utilizar populações endogâmicas. Dentre elas tem-se variação no número de alelos segregando por loco e, geralmente, há desconhecimento da fase de ligação. Em famílias de irmãos completos diferentes graus de informações são observados na progênie, podendo esta ser do tipo completamente informativa ou não completamente informativa. O objetivo deste trabalho foi, por meio de simulação, discutir as diferenças existentes no mapeamento envolvendo populações com diferentes graus de informação. Para isto foi simulado um genoma constituído de três grupos de ligação, sendo cada um constituído de 11 marcas moleculares multialélicas, codominantes, eqüidistantes, com saturação de 10 cM. A partir deste genoma foram simulados dois cenários, sendo um com genitores completamente informativos e outro com genitores formados aleatoriamente. Após a obtenção de todas as estimativas da freqüência de recombinação concluiu-se que para locos completamente informativos pode-se calcular a freqüência de recombinação entre pares de locos tanto a partir da freqüência gamética de cada genitor quanto a partir da freqüência genotípica conjunta da progênie. Para locos parcialmente informativos a obtenção da freqüência de recombinação a partir da freqüência genotípica conjunta foi mais apropriada.

Termos de indexação: populações exogâmicas, genômica, Máxima Verossimilhança.

ABSTRACT

BHERING, Leonardo Lopes, D.Sc., Universidade Federal de Viçosa. **Genetic mapping in simulated full siblings family (1): The special study.** Adviser: Cosme Damião Cruz Co-Advisers: José Marcelo Soriano Viana and Pedro Crescêncio Souza Carneiro.

The establishment of genetic maps in exogamic populations presents certain complications not found when endogamic populations are used. One of these complications is the variation in the numbers of alleles segregating per locus and the linkage phase which is usually unknown. The method most used to estimate the recombination percentage is the method of the maxim likelihood. In full siblings families it is possible to find different degrees of information in the progeny, which can be completely informative or not informative. The objective of this work was to discuss the differences in the mapping involving populations with different degrees of information, through simulation. It was simulated a genome with three linkage groups, each one with 11 molecular marks which were multi-allelic, codominants, equidistants, with saturation of 10 cM. Starting from this genome, two situations were simulated, one with informative genitors, and other with genitors formed randomly. Once all the recombination frequencies were estimated it was found out that in completely informative locus, the recombination frequency can be calculated among equal pairs of locus starting from the gametic frequency of each genitor or starting from the genotipic frequency of the progeny. For partially informative locus, the recombination frequency was more appropriate estimated starting from the united genotipic frequency.

Indexation terms: Exogamic Populations, Genomics, Maxim likelihood

1. Introdução

O estabelecimento de mapas genéticos em populações exogâmicas apresenta certas dificuldades que não são encontradas quando utilizados os delineamentos genéticos estabelecidos a partir de linhagens endogâmicas, como populações F_2, RILs (Recombinant Inbred Lines), Duplo-Haplóides, Retrocruzamentos, dentre outros. Em populações segregantes derivadas de linhagens endogâmicas todos os locos estarão segregando para apenas dois alelos. Em adição, a fase de ligação do duplo heterozigoto pode ser claramente determinada com base na análise da segregação dos gametas recombinantes da população (Liu, 1998).

Contrariamente, na descendência de cruzamentos entre dois indivíduos não idênticos de uma população exogâmica o número de alelos segregando por loco marcador poderá variar em até quatro, para uma espécie diplóide, variando também entre locos. Ademais, usualmente a fase de ligação é desconhecida, sendo possível determiná-la apenas de forma estocástica.

Em determinadas espécies de plantas não é possível obter populações segregantes derivadas de linhagens endogâmicas, devido à auto-incompatibilidade, depressão endogâmica ou longo período juvenil. Dessa forma, em tais espécies é preciso empregar delineamentos genéticos de populações exogâmicas como Famílias de Meio Irmãos (FMI) e Famílias de Irmãos Completos (FIC).

Alguns artigos na literatura relatam sobre a utilização de marcadores moleculares em populações exogâmicas, mostrando algumas expressões para a estimação da freqüência de recombinação, sem abranger todas as situações possíveis (Ritter et al., 1990; Arus et al., 1994; Ritter e Salamini, 1996; Maliepaard et al., 1997). Entretanto, na maioria deles, encontra-se apenas a aplicação de algum software para a obtenção das estimativas de distância não discutindo a metodologia utilizada.

Neste trabalho, foi discutido, com base em um exemplo simulado, a análise de ligação entre pares de marcas para diferentes tipos de acasalamentos em FIC. O objetivo foi contrastar dois cenários de simulação, um que emprega genitores completamente informativos para todos os locos (Cenário 1); outro que emprega genitores de forma aleatória (Cenário 2) quanto ao conteúdo de informação. Foi considerada a utilização de marcadores codominantes, distribuídos de forma eqüidistante ao longo de um genoma hipotético, de forma o que permitiu uma discussão a respeito de como determinar a fase de

ligação e os problemas gerados pela falta de informação para determinar ligação entre certas combinações de locos.

2. Revisão de Literatura

2.1. Mapeamento Genético

Na construção de mapas genéticos deve-se, preliminarmente, proceder à triagem de marcadores genéticos, preocupando-se com o número e o tipo de marcadores a serem utilizados. Obtidos os escores dos marcadores na população de mapeamento, efetuam-se testes de segregação para a análise individual das marcas (Schuster e Cruz, 2004). O objetivo desta etapa é selecionar adequadamente aqueles marcadores que apresentem as freqüências esperadas para a segregação de um loco único em uma população. A ocorrência de segregação distorcida significativa indica que o modelo genético adotado pode ser inapropriado, que os dados são de baixa qualidade ou que o processo de amostragem não foi aleatório (Liu, 1998).

Na etapa de construção do mapa genético, o primeiro passo é a análise de ligação para pares de locos. Assim, inicialmente é necessário que todos os marcadores sejam analisados par a par, verificando se existe ligação entre eles. O teste apropriado para esta situação é o teste χ^2. Entretanto, este teste é apenas qualitativo e quando detectada a evidência de ligação deve-se obter a porcentagem de recombinação entre os pares de marcadores. O método mais utilizado para a estimação da porcentagem de recombinação é o método da máxima verossimilhança (MV). No mapeamento genético, este método é empregado tanto na obtenção das estimativas das freqüências de recombinação quanto na estimação de parâmetros no mapeamento de QTL (Schuster e Cruz, 2004).

Após a estimação da porcentagem de recombinação é necessário definir a freqüência máxima de recombinação e o LOD mínimo para inferir se dois locos estão ligados. O objetivo desta verificação é estabelecer critérios para serem utilizados na formação dos grupos de ligação. Quanto mais informativo for o conjunto de dados, maior será a aproximação do número dos grupos de ligação em relação ao número haplóide de cromossomos da espécie (Liu, 1998). Um grande número de marcadores não ligados é sinal de baixa qualidade dos dados ou amostragem insuficiente de marcadores ou indivíduos (Schuster e Cruz, 2004).

Um ponto importante a destacar no processo de agrupamento é a falta de aditividade das distâncias entre pares de marcas quando expressas pela porcentagem de recombinação. Para facilitar o agrupamento e o ordenamento dos locos em um grupo de ligação o critério de otimização a ser utilizado são as funções de mapeamento genético. As duas principais funções de mapeamento utilizadas são a de Haldane (1919) e a de Kosambi (1944). O objetivo das funções de mapeamento é estabelecer a relação entre distância de mapa e freqüência de recombinação entre os pares de marca, resolvendo ou minimizando o problema da aditividade. Cabe ressaltar que diferentes funções de mapeamento correspondem a diferentes graus de interferência assumidos na permuta entre regiões adjacentes (Schuster e Cruz, 2004).

Depois de definidos os grupos de ligação, uma última abordagem é estimar as freqüências de recombinação *multipoint* ou multilocos entre pares de marcas. A análise multilocos considera todos os marcadores ligados em um grupo de ligação simultaneamente, resultando em uma análise única para cada grupo de ligação (Lynch e Walsh, 1998). De forma geral a análise multilocos está condicionada à função de mapeamento utilizada. A escolha de uma, dentre as diferentes funções disponíveis, depende das pressuposições a respeito das distribuições de permuta, do grau de interferência e do comprimento do segmento cromossômico analisado (Schuster e Cruz, 2004).

Matter (1951), Allard (1956) e Weber e Wricke (1994) desenvolveram estimadores de máxima verossimilhança para obter as freqüências de recombinação para Retrocruzamentos e populações F_2. Ritter et al. (1990) desenvolveram estimadores para situações entre cruzamentos entre pais heterozigotos. Arus et al. (1994) contribuíram adicionando estimadores para mais duas situações, mesma coisa realizada por Ritter e Salamini (1996) e Maliepaard et al. (1997) completaram os estudos anteriormente citados, mostrando todas as combinações possíveis com dois ou quatro alelos, desconsiderando o efeito da epistasia, para famílias de irmãos completos. Este estudo foi complementado por Wu e Ma (2002) ao adicionar epistasia em seus modelos.

2.2. Conteúdo de Informação e Fase de Ligação em Mapeamento com Famílias de Irmãos Completos

Diferentes configurações de marcadores podem estar segregando em Famílias de Irmãos Completos originadas a partir do cruzamento entre parentais derivados de uma

população exogâmica. De acordo com Haseman e Elston (1972), para o caso geral de sistemas multialélicos com quatro ou mais alelos, haverá, basicamente, três categorias e sete tipos distintos de acasalamentos ou diferentes tipos de pares de irmãos que caracterizam a herança de marcas individuais. Assim, considerando-se um loco A com alelos i, j, k e l, ter-se-ão os seguintes tipos:

1) Cruzamento entre genitores homozigotos:

I. A_iA_i-A_iA_i

II. A_iA_i-A_jA_j

2) Cruzamento entre genitores homozigotos e heterozigotos:

III. A_iA_i-A_iA_j

IV. A_iA_i-A_jA_k

3) Cruzamento entre genitores heterozigotos:

V. A_iA_j-A_iA_j

VI. A_iA_j-A_iA_k

VII. A_iA_j-A_kA_l

Observa-se que o tipo I envolve apenas um alelo, os tipos II, III e V dois alelos, os tipos IV e VI três alelos, e o tipo VII quatro alelos. Independente do número total de alelos em um loco, no máximo quatro alelos poderão estar segregando em determinado acasalamento, considerando uma espécie diplóide. Verifica-se ainda que A_iA_i-A_iA_i e A_jA_j-A_jA_j são dois acasalamentos genotipicamente diferentes, mas ambos pertencendo ao tipo I. Da mesma forma, A_iA_j-A_iA_j, A_iA_k-A_iA_k, A_iA_l-A_iA_l, A_jA_k-A_jA_k, A_jA_l-A_jA_l e A_kA_l-A_kA_l são todos pertencentes ao tipo V de acasalamento.

Na análise de segregação, deve-se considerar que somente acasalamentos que envolvam pelo menos um dos genitores heterozigoto sejam informativos para fins de mapeamento. Dessa forma, apenas cruzamentos dos tipos III, IV (cruzamentos entre genitores homozigotos e heterozigotos) e V, VI, e VII (cruzamentos entre genitores heterozigotos) são informativos. Adicionalmente, deve ser observada a existência de indivíduos informativos na prole, que são aqueles nos quais é possível identificar a origem de seus alelos em relação aos parentais. De forma resumida, é possível classificar os seguintes tipos de acasalamentos quanto ao grau de informação da progênie (Lynch e Walsh, 1998): (1) famílias derivadas de cruzamentos completamente informativos; (2)

famílias derivadas de “retrocruzamentos” e (3) famílias derivadas de “intercruzamentos” (Tabela 1).

Tabela 1. Tipos de famílias quanto ao grau de informação.

Categoria	Cruzamento	Característica
Completamente Informativo	A_iA_j x A_kA_l	Alelos de cada genitor são distinguidos
Retrocruzamento	A_iA_j x A_kA_k	Alelos de apenas um genitor são distinguidos
Intercruzamento	A_iA_j x A_iA_j	Só descendentes homozigotos são informativos

Fonte: Lynch e Walsh (1998)

Uma família com loco completamente informativo é derivada do cruzamento A_iA_j x A_kA_l. Nesse tipo de cruzamento, os genitores são heterozigotos para o marcador, possibilitando que toda a progênie seja informativa na distinção entre os alelos derivados de cada genitor. A segregação esperada é de 1:1:1:1.

Uma família com loco informativo, é aquela proveniente de “retrocruzamento” A_iA_j x A_kA_k (em que um genitor é heterozigoto para o marcador enquanto o outro é homozigoto). Dessa forma, apenas os alelos do genitor informativo (heterozigoto) podem ser distinguidos na progênie, de maneira que a segregação esperada é de 1:1.

No último caso, em famílias de “intercruzamentos” ou famílias parcialmente informativas (A_iA_j x. A_iA_j), apenas a progênie homozigota é informativa, pois não é possível distinguir nos indivíduos heterozigotos a procedência dos alelos de cada genitor. A segregação esperada neste caso é de 1:2:1.

Considerando-se dois locos segregando e marcas codominantes, 81 configurações podem surgir da combinação dos diferentes tipos de acasalamentos citados acima. Destas configurações, 17 proporcionam informação sobre ligação e o restante não contém informação de ligação. Para marcadores dominantes, 7 de 9 configurações proporcionam informação de ligação (Liu, 1998).

Verifica-se, portanto, que na análise de ligação em Famílias de Irmãos Completos (FIC) existirão diferentes funções de verossimilhança para cada configuração empregada. Diferentes conteúdos de informação serão, portanto, obtidos com o uso de FIC no mapeamento genético, envolvendo cruzamentos completamente informativos, informativos e parcialmente informativos e suas combinações.

Um fator complicador na análise de ligação em cruzamentos evolvendo genitores exogâmicos é com relação à determinação da fase de ligação, que em geral não é conhecida a *priori*. Ressalta-se que o conhecimento da fase de ligação é requerido para a detecção de eventos de recombinação. A fase de ligação define a configuração dos alelos de um par de locos heterozigotos em cromossomos homólogos de um parental. Em populações segregantes derivadas de linhagens endogâmicas, a fase de ligação é a mesma nos parentais, ao passo que em FIC as seguintes combinações de fase de ligação podem ser encontradas: (1) acoplamento (ou atração) no primeiro parental e indefinido no segundo, ou vice-versa; (2) repulsão no primeiro parental e indefinido no segundo, ou vice-versa; (3) acoplamento em ambos os parentais; (4) repulsão em ambos parentais; (5) acoplamento no primeiro parental e repulsão no segundo, ou vice-versa (Maliepaard et al., 1997).

2.3. Mapeamento em Família de Irmãos Completos

A utilização de família de irmãos completos para avaliações genéticas, pode contribuir muito para o avanço nos estudos de mapeamento genético e detecção de QTLs em diferentes culturas. Um exemplo disto acontece em culturas perenes nas quais o cálculo da freqüência de recombinação entre pares de locos é mais complexo do que em espécies anuais, para as quais é factível a obtenção de linhagens endogâmicas. Isto se deve ao fato que as espécies perenes possuem longo ciclo de vida e, geralmente, sofrem alta depressão em cruzamentos endogâmicos, dificultando ou até impossibilitando a obtenção de linhagens homozigóticas e, portanto, o emprego de delineamentos clássicos de mapeamento, tais como populações F_2, retrocruzamentos e linhagens recombinantes endogâmicas (RILs). Dessa maneira, as famílias disponíveis para mapeamento em perenes, como pode ser observado em *Eucalyptus*, são derivadas do cruzamento de parentais heterozigóticos. Com isso, pode-se ter na população de mapeamento até quatro alelos segregando para cada loco, tornando complicadas as análises de ligação por causa da existência de muitos tipos de razões de segregação entre pares de marcas (Maliepaard et al., 1997). Com a estratégia de pseudo cruzamento-teste (Grattapaglia e Sederoff, 1994), na qual os marcadores segregantes são analisados separadamente para cada parental, a utilização de genitores heterozigóticos para a construção de mapas de ligação tornou-se possível. A limitação desta estratégia é que ela apenas faz uso de uma porção dos marcadores moleculares, uma vez que todos os marcadores não informativos devem ser eliminados das análises. Ao utilizar marcadores dominantes esta é uma estratégia muito útil, porém com a maior utilização atual

de marcadores codominantes, novas estratégias devem ser empregadas afim de que possa aproveitar também estes marcadores não informativos pra aumentar o número de informações e a acurácia dos mapas gerados.

Ritter et al. (1990) e Ritter e Salamini (1996) propuseram um método estatístico para estimar a fração de recombinação entre diferentes tipos de segregação entre marcas. Através de métodos analíticos e simulação, Maliepaard et al.(1997) discutiu sobre o poder e precisão das estimativas de freqüência de recombinação obtidas entre pares de marcas. Em 2002, Wu et al., utilizando análise multiloco, definiram expressões possibilitando a obtenção da freqüência de recombinação e da fase de ligação para múltiplos marcadores. Trabalhos semelhantes foram realizados por Lander e Green (1987), George et al. (1999), Guilherme et al. (2002), Lu et al. (2004), Ma et al. (2004) e Jansen (2005).

Novaes (2006), utilizando uma família de irmãos completos de Eucalyptus constituída de 188 indivíduos, obtidas através do cruzamento entre *E. grandis* x *E. urophylla*, realizou o mapeamento genético e detecção de QTL para qualidade de madeira. Em seu estudo foi utilizada a estratégia de pseudocruzamento teste e posterior integração de mapas. Este autor encontrou QTL associados a todas as características fenotípicas avaliadas. Outros autores também utilizaram uma família de irmãos completos, proveniente de cruzamento entre *E. grandis* x *E. urophylla*, para construção de mapas pela estratégia de pseudocruzamento teste (Grattapaglia e Sederoff, 1994; Verhaegen e Plomion, 1996). Diversos outros trabalhos, como os de Ukrainetz et al. (2008) e Doligez et al. (2006), podem ser encontrados na literatura utilizando famílias de irmãos completos, em *Eucalyptus* ou outras culturas, para mapeamento e detecção de QTL. Porém na maioria encontra-se apenas a aplicação de algum software para a obtenção das estimativas de freqüência de recombinação, não discutindo a metodologia utilizada.

No mapeamento de irmãos completos muitas vezes, é necessário lançar mão de estratégias de integração de mapas. O processo de integração de mapas pode também ser requerido em análise de mapeamento, avaliando famílias exogâmicas para a obtenção das estimativas de freqüência de recombinação entre locos marcadores que não permitem tal obtenção de forma direta. Este processo é utilizado para estimar valores de porcentagem de recombinação entre locos não informativos a partir das informações das distâncias entre locos adjacentes às marcas cuja distância não pode ser determinada por impossibilidade de detecção de classes recombinantes. Sendo assim, para estes locos não informativos tem que se lançar mão do uso de marcadores âncoras, os quais possibilitam que a estimação da

freqüência de recombinação entre estes locos não informativos seja obtida de forma indireta, permitindo a alocação de tais marcadores no mapa genético. Como exemplo, pode-se considerar o cruzamento $A_1A_2B_1B_2C_1C_1$ x $A_1A_1B_3B_4C_1C_2$. Veja que a distância entre os locos A e C é indeterminada.

		½ A_1C_1	½ A_1C_2
½	A_1C_1	r *= ?	
½	A_2C_1		

(*) Apesar de ambos os locos segregarem, o valor da porcentagem de recombinação (r) é indeterminado.

Porém é possível obter valores de distância entre A e B e entre B e C. Desta forma, a âncora B possibilitará a integração das informações e posterior estimação da distância entre A e C.

No mapeamento de irmãos completos duas situações podem ser observadas. Em uma o cruzamento avaliado é completamente informativo e na outra nem toda progênie do cruzamento é informativa. No segundo caso, em certas ocasiões, deve ser usada a estratégia de integração de mapas a fim de obter a estimativa da freqüência de recombinação de forma indireta.

No caso em que os locos são completamente informativos, podem ser obtidas três estimativas da porcentagem de recombinação. Nas duas primeiras situações podem ser obtidas estimativas da freqüência de recombinação com base nas freqüências gaméticas marginais de cada genitor. Alternativamente, pode-se utilizar a informação da freqüência genotípica conjunta da progênie, sendo possível, para este tipo de combinação, a reconstrução completa dos haplótipos parentais na prole. No caso em que os locos não são completamente informativos usualmente tem se utilizado a técnica do Pseudotestecross (Gratttapaglia e Sederoff, 1994), onde faz um mapa para cada genitor e depois integra os dois mapas a fim de obter um mapa único. Porém ao usar esta estratégia, os marcadores não informativos devem ser excluídos da análise havendo, portanto, perda de informações de alguns indivíduos da população. Uma solução para evitar esta perda seria a utilização da informação da freqüência genotípica conjunta da progênie.

Como exemplo de cálculo da porcentagem de recombinação para os diferentes tipos de acasalamento será abordado a situação em que os locos são completamente informativos

e, portanto, com as combinações do tipo (1:1:1:1)(1:1:1:1). Pode-se escrever de forma generalizada o seguinte cruzamento:

$$P_1 : A_1A_2B_1B_2 \qquad \times \qquad P_2 : A_3A_4B_3B_4$$

Considerando os haplótipos para os marcadores no parental 1 em fase de acoplamento, os gametas produzidos e suas freqüências serão do tipo:

$$f(A_1B_1) = (1-r)/2 = P$$
$$f(A_1B_2) = r/2 = R$$
$$f(A_2B_1) = r/2 = R$$
$$f(A_2B_2) = (1-r)/2 = P$$

Da mesma forma para o segundo parental, ter-se-á:

$$f(A_3B_3) = (1-r)/2 = P$$
$$f(A_3B_4) = r/2 = R$$
$$f(A_4B_3) = r/2 = R$$
$$f(A_4B_4) = (1-r)/2 = P$$

Combinando-se os gametas dos dois parentais teremos 16 classes genotípicas, conforme esquema a seguir:

Gametas		A_3B_3	A_3B_4	A_4B_3	A_4B_4	Freqüência	Total de
		P	R	R	P	Gamética	indivíduos
A_1B_1	P	$P^2_{13\text{-}13}$	$PR_{13\text{-}14}$	$PR_{14\text{-}13}$	$P^2_{14\text{-}14}$	(1-r)/2	$n_{1.}$
A_1B_2	R	$PR_{13\text{-}23}$	$R^2_{13\text{-}24}$	$R^2_{14\text{-}23}$	$PR_{14\text{-}24}$	r/2	$n_{2.}$
A_2B_1	R	$PR_{23\text{-}13}$	$R^2_{23\text{-}14}$	$R^2_{24\text{-}13}$	$PR_{24\text{-}14}$	r/2	$n_{3.}$
A_2B_2	P	$P^2_{23\text{-}23}$	$PR_{23\text{-}24}$	$PR_{24\text{-}23}$	$P^2_{24\text{-}24}$	(1-r)/2	$n_{4.}$
Freq. Gamética		(1-r)/2	r/2	r/2	(1-r)/2		
Total		$n_{.1}$	$n_{.2}$	$n_{.3}$	$n_{.4}$		N

Como mencionado, o cálculo da porcentagem de recombinação (r_G) pode ser feito de três maneiras:

a) Com base nas freqüências gaméticas marginais do genitor 1, de modo que:

O genitor 1 estará em fase de acoplamento (atração) se for observado que $(n_{1.} + n_{4.}) > (n_{2.} + n_{3.})$. Nesta situação, a porcentagem de recombinação será calculada por:

$$r_{G1CIS} = (n_{2.} + n_{3.}) / N$$

Caso seja observado que $(n_{1.} + n_{4.}) < (n_{2.} + n_{3.})$, o genitor 1 estará em fase de repulsão de modo que:

$$r_{G1TRANS} = (n_{1.} + n_{4.}) / N$$

Uma vez que,

$r_{G1CIS} = (n_{2.} + n_{3.}) / N$, $r_{G1TRANS} = (n_{1.} + n_{4.}) / N$ e $(n_{1.} + n_{2.} + n_{3.} + n_{4.}) / N = 1$, tem-se:

$r_{G1CIS} + r_{G1TRANS} = 1$ então,

$r_{G1TRANS} = 1 - r_{G1CIS}$

b) Com base nas freqüências gaméticas marginais do genitor 2, de modo que:

O genitor 2 estará em fase de acoplamento se for observado que $(n_{.1} + n_{.4}) > (n_{.2} + n_{.3})$. Nesta situação, a porcentagem de recombinação será calculada por:

$$r_{G2CIS} = (n_{.2} + n_{.3}) / N$$

Caso seja observado que $(n_{.1} + n_{.4}) < (n_{.2} + n_{.3})$, o genitor 1 estará em fase de repulsão de modo que:

$$r_{G2TRANS} = (n_{.1} + n_{.4}) / N$$

Como observado anteriormente de maneira similar, a seguinte relação é verdadeira:

$$r_{G2TRANS} = 1 - r_{G2CIS}$$

c) Com base na freqüência genotípica conjunta da progênie, por meio do método da máxima verossimilhança:

Tendo estabelecido adequadamente a fase de ligação, pode-se estimar a porcentagem de recombinação por meio do Método da Máxima Verossimilhança, com base nas freqüências genotípicas da progênie. Assim, a função de verossimilhança pode ser descrita para o exemplo, como apresentada a seguir:

$$L(r;n_i) = \frac{N!}{n_{11}!...n_{44}!}\left(P^2\right)^{n_{11}+n_{14}+n_{41}+n_{44}}\left(R^2\right)^{n_{22}+n_{23}+n_{32}+n_{33}}(PR)^{n_{12}+n_{13}+n_{21}+n_{24}+n_{31}+n_{34}+n_{42}+n_{43}}$$

Maximizando o logaritmo natural da função de verossimilhança, é obtido:

$$r = \frac{S_R}{n} + \frac{1}{2}\frac{S_{PR}}{n} = \frac{r_{G1} + r_{G2}}{2}$$

em que: S_R é a soma do número de indivíduos originados de duas cromátides recombinantes, cuja freqüência esperada é R^2; S_{PR} é a soma do número de indivíduos originados de uma cromátide recombinante e outra paternal, cuja freqüência esperada é PR.

Na situação onde a progênie não é completamente informativa, apenas a estimativa obtida por meio da função de máxima verossimilhança deve ser adotada, uma vez que leva em consideração todos os indivíduos da população, fornecendo resultados mais coerentes, conforme será discutido com maior riqueza de detalhes mais adiante.

3. Material e Métodos

Um genoma hipotético foi simulado utili\zando o módulo de simulação do aplicativo computacional GQMOL (Cruz, 2005). Foram simulados genomas parentais e amostras de populações de famílias de irmãos completos.

Foi tomada como referência uma espécie diplóide fictícia com 2n = 2x = 6 cromossos, composto de três grupos de ligação (GL). Cada grupo de ligação formado possuía tamanho de 100 cM, com 11 marcas moleculares co-dominantes eqüidistantes, com saturação de 10 cM . As marcas foram consideradas codominantes. Foram gerados por meio de simulação dois cenários de mapeamento genético em Famílias de Irmãos Completos, como descrito a seguir:

Cenário 1. Os genitores foram gerados de maneira a serem completamente informativos para todos os locos sendo, portanto, estabelecidos os cruzamentos do tipo A_iA_j x A_kA_l, nos quais os pais possuem alelos diferentes e são heterozigotos para cada loco. Deste modo, todas as progênies são informativas em distinção dos alelos alternativos provenientes de ambos os pais, de forma que, os alelos parentais podem ser distintos e ambos os pais podem ser examinados pela comparação de valores da característica nas proles $A_i\_$ vs $A_j\_$ e $A_k\_$ vs $A_l\_$. A partir do cruzamento foi obtida uma família de irmãos completos com 200 indivíduos.

Cenário 2. Os genitores foram gerados de forma aleatória, de modo que o conteúdo de informação seja variável para cada loco. A partir dos genitores foi obtida uma família de irmãos completos com 200 indivíduos

3.1. Cenário 1

O cenário 1 foi estabelecido para realizar o mapeamento genético com base em informações de famílias de irmãos completos derivado de genitores completamente informativos.

O esquema do genoma de cada genitor, composto por três grupos de ligação, está representado na Figura 1. Neste esquema pode perceber que o Pai 1 sempre possui alelos do tipo 12 (A_1A_2), já o Pai 2 possui alelos do tipo 34 (A_3A_4), de forma que a descendência produzida deste cruzamento seja completamente informativa. Pode observar ainda a distância média de 10 cM entre marcas moleculares.

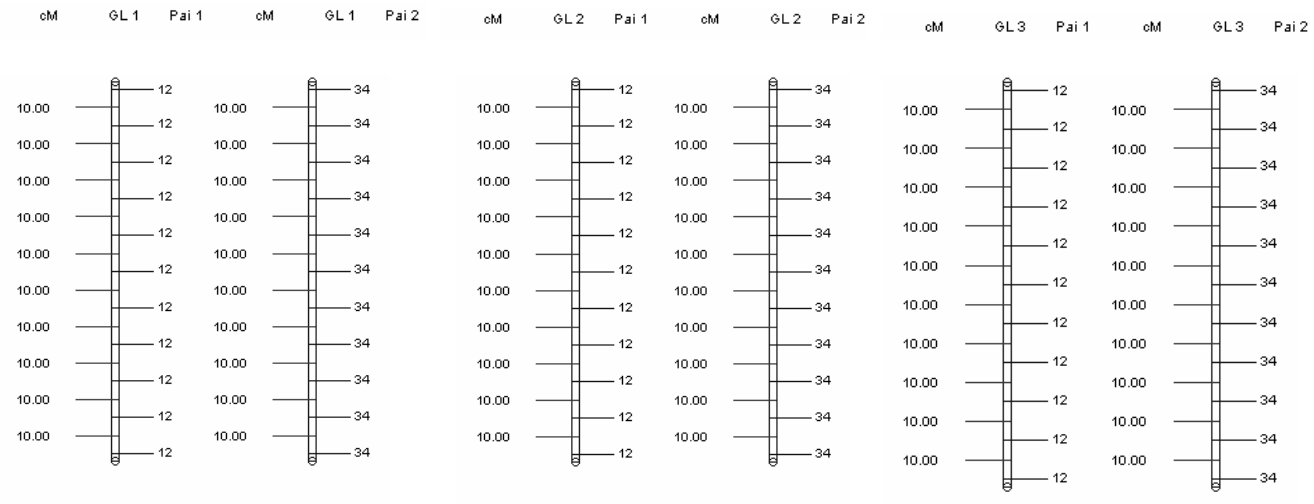

Figura 1. Representação do genoma dos genitores utilizados para gerar a população de mapeamento a ser analisada no cenário 1.

De posse do genoma simulado dos genitores realizou-se a simulação de uma população de mapeamento sendo família de irmãos completos (FIC) constituída de 200 indivíduos. O arquivo com os dados ficou disposto de forma resumida como apresentado na Tabela 2. Como foi utilizado marcadores completamente informativos do tipo A_iA_j x A_kA_l a segregação esperada a ser testada é a seguinte: 1 A_iA_k : 1 A_iA_l : 1 A_jA_k : 1 A_jA_l .

Os 200 indivíduos provenientes da simulação são agrupados em classes, de acordo com o seu genótipo, a partir daí foi realizado o teste de segregação, comparando a proporção observada, com a esperada por meio do teste qui-quadrado (χ^2).

Tabela 2. Esquema resumido dos dados gerados pela simulação do Cenário 1.

Indivíduos	Marcadores							
	M_1	M_2	M_3	M_4	…	M_{31}	M_{32}	M_{33}
P_1	12	12	12	12	...	12	12	12
P_2	34	34	34	34	...	34	34	34
ind_1	14	14	14	14	...	13	14	14
ind_2	14	14	14	14	...	23	23	23
ind_3	13	13	13	13	...	13	13	23
ind_4	13	23	23	24	...	13	13	13
	...	...	...	...		...	...	...
ind_{199}	24	24	24	24	...	13	13	13
ind_{200}	23	23	23	13	...	13	13	13

M_1 a M_{33}: marcadores codominantes multialélicos. P_1 e P_2: Genitores simulados. Ind_i: Indivíduos formados do cruzamento entre P_1 e P_2.

3.2. Cenário 2

Este cenário foi estabelecido para realizar o mapeamento genético com base em informações de famílias de irmãos completos derivadas de genitores ao acaso. O esquema do genoma dos genitores utilizado para a simulação da população de mapeamento (FIC) está representado na Figura 3. Neste esquema pode-se observar a distância de 10 cM entre marcas adjacentes. Observa-se ainda que a constituição genotípica de cada marcador foi obtida de forma aleatória, não seguindo o padrão observado no cenário 1, em que o Pai 1 era sempre 12 e o Pai 2 sempre do tipo 34.

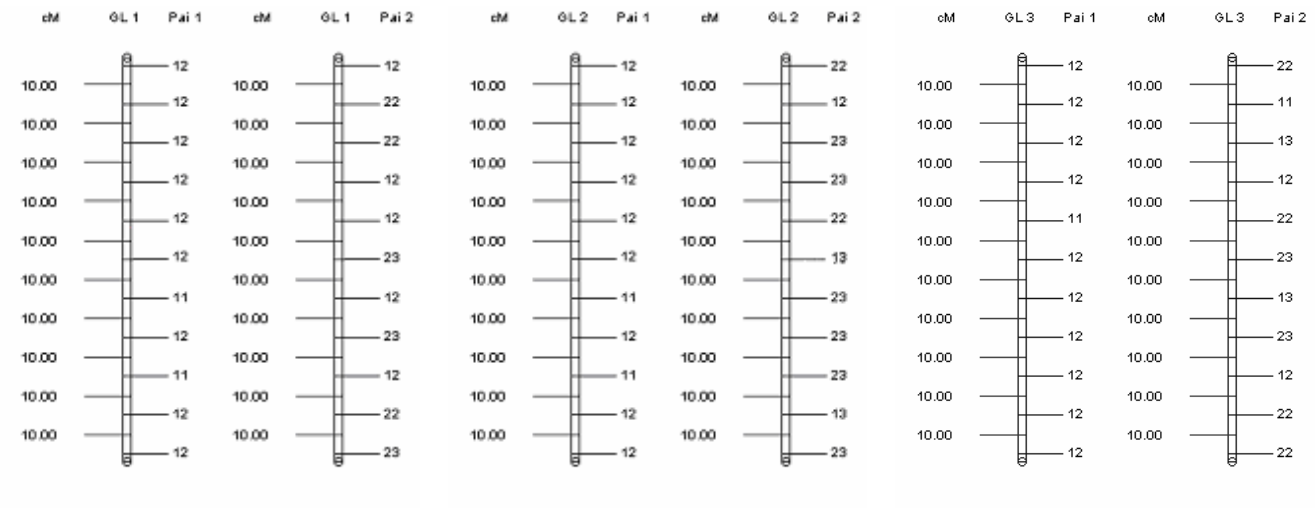

Figura 3. Mapa genético com representação genotípica dos genitores utilizados para gerar a população de mapeamento a ser analisada no cenário 2.

De posse do genoma simulado e dos genitores realizou a simulação da população de mapeamento, que no caso foi uma família de irmãos completos (FIC) constituída de 200 indivíduos. O arquivo com os dados ficou disposto de forma resumida como demonstrado na Tabela 3.

Tabela 3. Esquema resumido dos dados gerados pela simulação do Cenário 2.

	M_1	M_2	M_3	M_4	…	M_{31}	M_{32}	M_{33}
P_1	12	12	12	12	···	12	12	12
P_2	12	22	22	12	···	12	22	22
ind_1	22	22	22	12	···	12	22	12
ind_2	12	22	22	22	···	22	22	22
ind_3	12	12	12	12	···	11	22	12
ind_4	12	12	22	22	···	22	22	12
	...	...	...	...		...	...	...
ind_{199}	22	22	22	22	···	12	12	12
ind_{200}	11	12	12	11	···	11	12	22

M_1 a M_{33}: marcadores codominantes multialélicos. P_1 e P_2: Genitores simulados. Ind_i: Indivíduos formados do cruzamento entre P_1 e P_2.

4. Resultados e Discussão

4.1. Cenário 1

4.1.1. Teste de Segregação

Na Tabela 4 está representada a análise de segregação para marcas individuais com base no teste Qui-quadrado (χ^2) realizado com o auxílio do software Gqmol (Cruz, 2005). Como pode ser observado, todas as marcas analisadas segregam na proporção de 1:1:1:1, de modo que a combinação entre todos os locos será do tipo (1:1:1:1)(1:1:1:1). Como discutido anteriormente, se quatro alelos distintos estão presentes na prole de um cruzamento, os haplótipos que geraram a prole podem ser reconstruídos completamente, sendo possível distinguir, para todos os indivíduos, os gametas de origem parental e aqueles oriundos de recombinação. Dessa forma, a combinação de dois locos completamente informativos produz uma progênie totalmente informativa, o que permite, determinar a freqüência de recombinação entre os dois locos quaisquer.

Tabela 4. Teste de segregação dos 33 marcadores ao nível de 5% de probabilidade pelo Teste de χ^2.

		Classes Genotípicas							
Marcador	Pais	13	14	23	24	Hipótese	GL	Qui-quadrado	Probabilidade(%)
C11	12x34	56	49	42	53	1:1:1:1	3	2,20	53,1948 ns
C12	12x34	52	47	43	58	1:1:1:1	3	2,52	47,1688 ns
C13	12x34	48	48	50	54	1:1:1:1	3	0,48	92,3263 ns
C14	12x34	52	50	43	55	1:1:1:1	3	1,56	66,8493 ns
C15	12x34	53	50	39	58	1:1:1:1	3	3,88	27,4717 ns
C16	12x34	57	49	39	55	1:1:1:1	3	3,92	27,0233 ns
C17	12x34	52	52	54	42	1:1:1:1	3	1,76	62,3678 ns
C18	12x34	53	50	58	39	1:1:1:1	3	3,88	27,4717 ns
C19	12x34	53	50	56	41	1:1:1:1	3	2,52	47,1688 ns
C110	12x34	54	45	58	43	1:1:1:1	3	3,08	37,9454 ns
C111	12x34	41	44	58	57	1:1:1:1	3	4,60	20,3542 ns
C21	12x34	49	47	49	55	1:1:1:1	3	0,72	86,8490 ns
C22	12x34	48	52	52	48	1:1:1:1	3	0,32	95,6224 ns
C23	12x34	45	52	53	50	1:1:1:1	3	0,76	85,9009 ns
C24	12x34	42	55	49	54	1:1:1:1	3	2,12	54,7877 ns
C25	12x34	43	56	49	52	1:1:1:1	3	1,80	61,4935 ns
C26	12x34	41	58	54	47	1:1:1:1	3	3,40	33,3965 ns
C27	12x34	46	52	60	42	1:1:1:1	3	3,68	29,8156 ns
C28	12x34	49	50	60	41	1:1:1:1	3	3,64	30,3053 ns
C29	12x34	46	44	58	52	1:1:1:1	3	2,40	49,3635 ns

C210	12x34	47	48	50	55	1:1:1:1	3	0,76	85,9009 ns
C211	12x34	53	43	45	59	1:1:1:1	3	3,28	35,0436 ns
C31	12x34	56	48	38	58	1:1:1:1	3	4,96	17,4750 ns
C32	12x34	54	46	41	59	1:1:1:1	3	3,88	27,4717 ns
C33	12x34	50	56	46	48	1:1:1:1	3	1,12	77,2248 ns
C34	12x34	52	55	41	52	1:1:1:1	3	2,28	51,6363 ns
C35	12x34	45	54	45	56	1:1:1:1	3	2,04	56,4146 ns
C36	12x34	47	47	52	54	1:1:1:1	3	0,76	85,9009 ns
C37	12x34	51	49	53	47	1:1:1:1	3	0,40	94,0242 ns
C38	12x34	56	47	44	53	1:1:1:1	3	1,80	61,4935 ns
C39	12x34	47	54	49	50	1:1:1:1	3	0,52	91,4476 ns
C310	12x34	57	49	39	55	1:1:1:1	3	3,92	27,0233 ns
C311	12x34	43	56	49	52	1:1:1:1	3	1,80	61,4935 ns

ns: Não-significativo admitindo nível crítico igual a 5% de probabilidade

4.1.2. Porcentagem de recombinação entre pares de marcadores

No caso em que os locos são completamente informativos, podem ser obtidas três estimativas da porcentagem de recombinação. As duas primeiras estimativas podem ser obtidas com base nas freqüências gaméticas marginais de cada genitor. Na terceira pode-se utilizar a informação da freqüência genotípica conjunta da progênie, sendo possível, para este tipo de combinação a reconstrução completa dos haplótipos parentais na prole.

Para exemplificar o cálculo da porcentagem de recombinação, foram tomadas as marcas C_{11} (12 x 34) e C_{12} (12 x 34). Para o cálculo da porcentagem de recombinação nos diferentes tipos de acasalamento será abordada, como ilustração, a situação na qual os locos são completamente informativos – combinações do tipo (1:1:1:1)(1:1:1:1), este será apresentado com maior riqueza de detalhes.

Tomando as marcas C_{11} (12 x 34) e C_{12} (12 x 34), pode-se escrever de forma generalizada, o seguinte cruzamento:

$$P_1 : A_1A_2B_1B_2 \qquad \times \qquad P_2 : A_3A_4B_3B_4$$

Considerando os haplótipos para os marcadores no genitor 1 em fase de acoplamento, os gametas produzidos e suas freqüências são do tipo:

$$f(A_1B_1) = (1-r)/2 = P$$

$$f(A_1B_2) = r/2 = R$$

$$f(A_2B_1) = r/2 = R$$

$$f(A_2B_2) = (1-r)/2 = P$$

Da mesma forma para o segundo genitor, ter-se-á:

$$f(A_3B_3) = (1-r)/2 = P$$

$$f(A_3B_4) = r/2 = R$$

$$f(A_4B_3) = r/2 = R$$

$$f(A_4B_4) = (1-r)/2 = P$$

Combinando-se os gametas dos dois genitores tem-se 16 classes genotípicas, conforme mostrado na Tabela 5.

Tabela 5. Freqüências gaméticas dos genitores e as freqüências genotípicas da progênie para cruzamento entre dois locos completamente informativos (1:1:1:1)(1:1:1:1).

Gametas		A_3B_3	A_3B_4	A_4B_3	A_4B_4	Freqüência	Total de
		P	R	R	P	Gamética	indivíduos
A_1B_1	P	$P^2_{13\text{-}13}$	$PR_{13\text{-}14}$	$PR_{14\text{-}13}$	$P^2_{14\text{-}14}$	(1-r)/2	$n_{1.}$
A_1B_2	R	$PR_{13\text{-}23}$	$R^2_{13\text{-}24}$	$R^2_{14\text{-}23}$	$PR_{14\text{-}24}$	r/2	$n_{2.}$
A_2B_1	R	$PR_{23\text{-}13}$	$R^2_{23\text{-}14}$	$R^2_{24\text{-}13}$	$PR_{24\text{-}14}$	r/2	$n_{3.}$
A_2B_2	P	$P^2_{23\text{-}23}$	$PR_{23\text{-}24}$	$PR_{24\text{-}23}$	$P^2_{24\text{-}24}$	(1-r)/2	$n_{4.}$
Freq. Gamética		(1-r)/2	r/2	r/2	(1-r)/2		
Total		$n_{.1}$	$n_{.2}$	$n_{.3}$	$n_{.4}$		N

Como mencionado, o cálculo da porcentagem de recombinação pode ser feito de três maneiras:

a) Com base nos freqüências gaméticas marginais do genitor 1

O genitor 1 estará em fase de acoplamento (atração) se for observado que $(n_{1.} + n_{4.}) > (n_{2.} + n_{3.})$. Nesta situação, a porcentagem de recombinação será calculada por:

$$r_{G1CIS} = (n_{2.} + n_{3.})/N$$

Caso seja observado que $(n_{1.} + n_{4.}) < (n_{2.} + n_{3.})$, o genitor 1 estará em fase de repulsão de modo que:

$$r_{G1TRANS} = (n_{1.} + n_{4.})/N$$

Uma vez que,

$r_{G1CIS} = (n_{2.} + n_{3.})/N$, $r_{G1TRANS} = (n_{1.} + n_{4.})/N$ e $(n_{1.} + n_{2.} + n_{3.} + n_{4.})/N = 1$, tem-se:

$r_{G1CIS} + r_{G1TRANS} = 1$ então,

$r_{G1TRANS} = 1 - r_{G1CIS}$

Este estimador é, também, de máxima verossimilhança.

b) Com base nos freqüências gaméticas marginais do genitor 2

O genitor 2 estará em fase de acoplamento se for observado que $(n_{.1} + n_{.4}) > (n_{.2} + n_{.3})$. Nesta situação, a porcentagem de recombinação será calculada por:

$$r_{G2CIS} = (n_{.2} + n_{.3})/N$$

Caso seja observado que $(n_{.1} + n_{.4}) < (n_{.2} + n_{.3})$, o genitor 1 estará em fase de repulsão de modo que:

$$r_{G2TRANS} = (n_{.1} + n_{.4})/N$$

Como observado anteriormente de maneira similar, a seguinte relação é verdadeira:

$$r_{G2TRANS} = 1 - r_{G2CIS}$$

Este estimador é, também, de máxima verossimilhança.

c) Com base na freqüência genotípica conjunta da progênie, por meio do método da máxima verossimilhança envolvendo todas as classes genotípicas

Tendo estabelecido adequadamente a fase de ligação, pode-se estimar a porcentagem de recombinação por meio do Método da Máxima Verossimilhança, com base nas freqüências genotípicas da progênie. Assim, a função de verossimilhança pode ser descrita para o exemplo, como apresentada a seguir:

$$L(r;n_i) = \frac{N!}{n_{11}!...n_{44}!}\left(P^2\right)^{n_{11}+n_{14}+n_{41}+n_{44}}\left(R^2\right)^{n_{22}+n_{23}+n_{32}+n_{33}}(PR)^{n_{12}+n_{13}+n_{21}+n_{24}+n_{31}+n_{34}+n_{42}+n_{43}}$$

Maximizando o logaritmo natural da função de verossimilhança, é obtido:

$$r = \frac{S_R}{n} + \frac{1}{2}\frac{S_{PR}}{n} = \frac{r_{G1} + r_{G2}}{2}$$

em que: S_R é soma do número de indivíduos originados de duas cromátides recombinantes, cuja freqüência esperada é R^2; S_{PR} é a soma do número de indivíduos originados de uma cromátide recombinante e outra paternal, cuja freqüência esperada é PR.

Com base na Tabela 5, pode-se alocar os valores das diferentes classes genotípicas a partir dos dados dos marcadores C_{11} e C_{12} que segregam nos 200 indivíduos da FIC (Tabela 6). Por meio da Tabela 7, foi possível relacionar as classes genotípicas de marcadores segregando na progênie às suas respectivas freqüências genotípicas esperadas . Dessa forma, pode-se construir funções de verossimilhanças para estimar os valores de porcentagem de recombinação para os diferentes tipos de combinações de acasalamento.

Calculando-se a porcentagem de recombinação com base nas freqüências marginais gaméticas de cada genitor, teremos:

$$r_{G1CIS} = (n_{2.} + n_{3.})/N = (9+3)/200 = 0{,}06\text{, pois } (n_{1.} + n_{4.}) > (n_{2.} + n_{3.})$$

$$r_{G2CIS} = (n_{.2} + n_{.3})/N = (9+6)/200 = 0{,}075\text{, pois } (n_{.1} + n_{.4}) > (n_{.2} + n_{.3})$$

Tabela 6. Representação de cruzamento do tipo: $A_1A_2B_1B_2$ x $A_3A_4B_3B_4$, exemplificado pelas marcas C_{11} e C_{12} pertencentes ao caso: (1:1:1:1) (1:1:1:1).

Gametas		A_3B_3 P	A_3B_4 R	A_4B_3 R	A_4B_4 P	Freqüência Gamética	Total de indivíduos
A_1B_1	P	$P^2_{13\text{-}13}$ 48	$PR_{13\text{-}14}$ 2	$PR_{14\text{-}13}$ 1	$P^2_{14\text{-}14}$ 45	(1-r)/2	96
A_1B_2	R	$PR_{13\text{-}23}$ 5	$R^2_{13\text{-}24}$ 1	$R^2_{14\text{-}23}$ 0	$PR_{14\text{-}24}$ 3	r/2	9
A_2B_1	R	$PR_{23\text{-}13}$ 3	$R^2_{23\text{-}14}$ 0	$R^2_{24\text{-}13}$ 0	$PR_{24\text{-}14}$ 0	r/2	3
A_2B_2	P	$P^2_{23\text{-}23}$ 33	$PR_{23\text{-}24}$ 6	$PR_{24\text{-}23}$ 5	$P^2_{24\text{-}24}$ 48	(1-r)/2	92
Freq. Gamética		(1-r)/2	r/2	r/2	(1-r)/2		
Total		89	9	6	96		200

Como mencionado, empregando-se cruzamentos com dois locos completamente informativos foi possível reconstruir todas as classes genotípicas e relacioná-las diretamente com suas freqüências esperadas como pode ser visto na Tabela 7. Dessa forma, pode-se calcular a porcentagem de recombinação com base nas freqüências genotípicas conjuntas da progênie. Como demonstrado:

$$r = \frac{S_R}{n} + \frac{1}{2}\frac{S_{PR}}{n} = \frac{1}{200} + \frac{1}{2}\frac{25}{200} = 0{,}0675$$

De forma que:

$$r = \frac{r_{G1} + r_{G2}}{2} = (0{,}06 + 0{,}075)\ /2 = 0{,}0675$$

Por meio da função de verossimilhança tem-se:

$$L(r; n_i) = \frac{200!}{48!2!\ldots5!48!}\left(P^2\right)^{48+45+33+48}\left(R^2\right)^{1+0+0+0}(PR)^{2+1+5+3+3+0+6+5},$$

fazendo $\frac{200!}{48!2!\ldots5!48!} = \lambda$ tem-se

$$L(r;n_i) = \lambda\left(P^2\right)^{174}\left(R^2\right)^{1}(PR)^{25}$$

$$L(r;n_i) = \lambda\left(\frac{1}{4}(1-r)^2\right)^{174}\left(\frac{1}{4}r^2\right)^{1}(\frac{1}{4}r(1-r))^{25}$$

A função suporte para estes dados pode, então ser resumidas em:

$$\ell(r,n_i) = \ln\lambda + \left[2x174\ln(1-r)\right] + \left[2\ln r\right] + \left[25\ln(r(1-r)\right]$$

A partir da primeira derivada de $\ell(r,n_i)$ em relação à r, obtém-se a função escore:

$$\frac{\partial\ell(r;n_i)}{\partial r} = \frac{-348}{1-r} + \frac{2}{r} + \frac{25(1-2r)}{r(1-r)}$$

$$\frac{\partial\ell(r;n_i)}{\partial r} = \frac{-348r + 2(1-r) + 25(1-2r)}{r(1-r)}$$

Fazendo as simplificações necessárias e igualando a função escore à zero, obtém-se a equação a seguir:

$$\frac{-348r + 2(1-r) + 25(1-2r)}{r(1-r)} = 0$$

$$-400r + 27 = 0$$

$$r = \frac{27}{400} = 0{,}0675$$

Desta forma encontra-se que a solução é r = 0,0675. Sendo assim, verifica-se que na construção de um mapa a partir de marcadores completamente informativos, pode usar toda a informação existente e aplicar a função de máxima verossimilhança, ou obter as estimativas da freqüência de recombinação das marginais e, posteriormente obter a média delas.

De forma geral para os locos em fase de aproximação nos dois pais tem-se:

$$L(r;n_i) = \lambda\left(P^2\right)^{n1}(PR)^{n2}\left(R^2\right)^{n3}$$

$$L(r;n_i) = \lambda\left(\frac{1}{4}(1-r)^2\right)^{n1}\left[\frac{1}{4}r(1-r)\right]^{n2}\left(\frac{1}{4}r^2\right)^{n3}$$

$$L(r;n_i) = \lambda\left((1-r)^2\right)^{n1}\left[r(1-r)\right]^{n2}\left(r^2\right)^{n3}$$

$$\ell(r,n_i) = \ln\lambda + \left[2n_1\ln(1-r)\right] + n_2\ln(1-r) + n_2\ln(r) + 2n_3\ln(r)$$

$$\ell(r,n_i) = \ln\lambda + (2n_1 + n_2)\ln(1-r) + (2n_3 + n_2)\ln(r)$$

$$\frac{\partial\ell(r;n_i)}{\partial r} = \frac{-(2n_1 + n_2)}{1-r} + \frac{(2n_3 + n_2)}{r}$$

$$\frac{-(2n_1 + n_2)}{1-r} + \frac{(2n_3 + n_2)}{r} = 0$$

$$-r(2n_1 + n_2) + (1 - r)(2n_3 + n_2) = 0$$

$$-2r(N) + (2n_3 + n_2) = 0$$

$$r = \frac{2n_3 + n_2}{2N}$$

Aplicando a equação geral obtida tem-se:

$$r = \frac{2(1) + 25}{2(200)} = \frac{27}{400} = 0{,}0675$$

Verifica-se, portanto, que a estimativa da porcentagem de recombinação dada pelas freqüências genotípicas conjuntas observadas na progênie foi igual à média das estimativas de porcentagem de recombinação com base nas freqüências gaméticas marginais dos genitores. Assim, em famílias obtidas do cruzamento de dois locos completamente informativos, a estimação da porcentagem de recombinação pode ser feita tanto com base nas freqüências marginais dos parentais ou com base na informação conjunta da progênie. Entretanto, será mostrado mais adiante – nas situações que envolvem diferentes conteúdos de informação para os locos envolvidos nos acasalamentos – que diferenças podem surgir nas estimativas calculadas por meio de freqüências marginais ou conjuntas. Neste caso onde existem diferentes tipos de informações, que a estratégia de obtenção de mapa único através da freqüência genotípica se torna extremamente eficiente. Isto porque, esta estratégia faz uso de todos os indivíduos da população, ao contrário da estratégia de realização de mapas através das freqüências marginais dos parentais, que leva em consideração apenas as informações oriundos da progênie completamente informativa para a realização de dois mapas de ligação, sendo um para cada parental, e depois realiza-se a integração dos mapas.

O mapa integrado dos genitores obtido da população de 200 indivíduos da FIC, após a análise dos pares de marcas está representado na Figura 4.

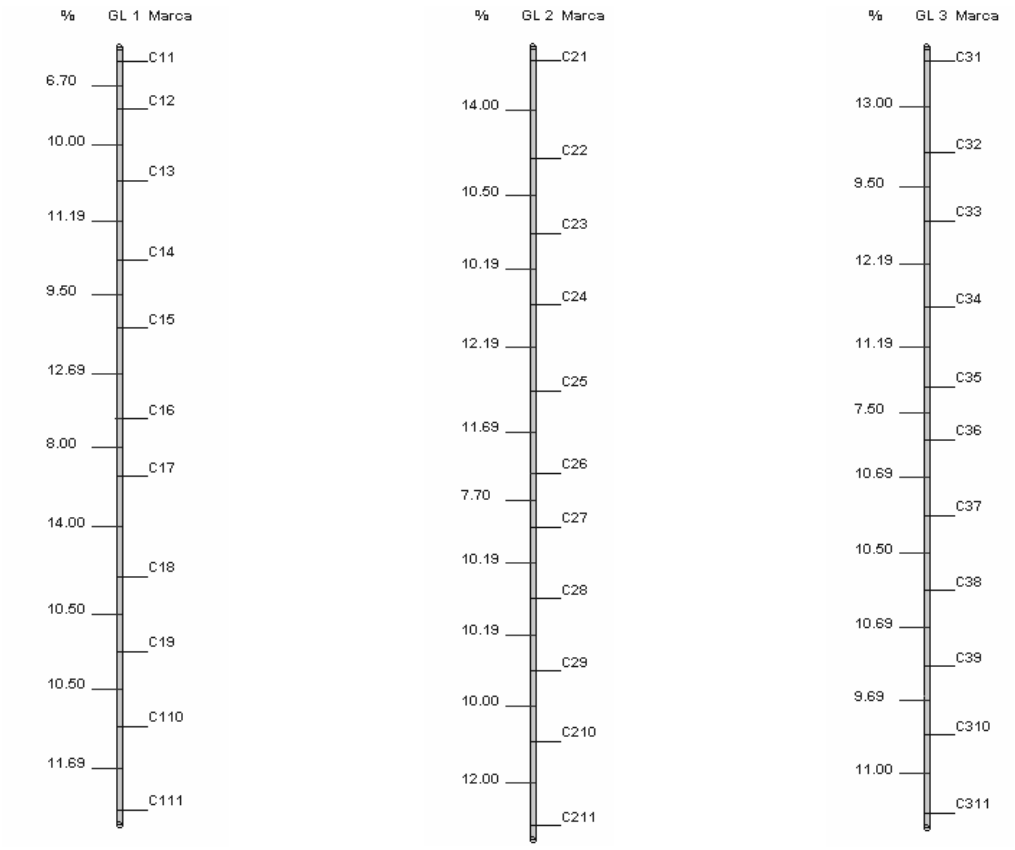

Figura 4. Mapa integrado dos genitores utilizados para gerar a população de mapeamento a ser analisada no cenário 1

Para testar se um par de marcadores está ligado, ou seja, se a porcentagem de recombinação paramétrica é igual ou inferior a 0,5, o LOD escore pode ser usado como teste estatístico. O LOD escore é o logaritmo de base 10 da razão entre a estimativa de máxima verossimilhança determinada ($r = \hat{r}$) e a estimativa de verossimilhança sob a hipótese de nulidade (r = 0,5, considerando ausência de ligação). Um LOD escore de 3 é comumente utilizado e significa dizer que a estimativa da porcentagem de recombinação é mil vezes mais provável que a hipótese de nulidade. Esta alta estringência é necessária, pois são realizados múltiplos testes para os diferentes pares de marcadores (Malieppard et. al., 1997).

Para o exemplo anterior tem-se;

$$\mathrm{LOD} = \mathrm{Log}_{10}\left[\frac{L(r;n_i)}{L(r=0{,}5;n_i)}\right]$$

$$LOD = \text{Log}_{10}\left[\frac{L(r = 0{,}0675; n_i)}{L(r = 0{,}5; n_i)}\right]$$

$$LOD = \text{Log}_{10}\left[\frac{\lambda\left(\frac{(1-0{,}0675)^2}{4}\right)^{174}\left(\frac{(0{,}0675)^2}{4}\right)^{1}\left(\frac{0{,}0675(1-0{,}0675)}{4}\right)^{25}}{\lambda\left(\frac{(1-0{,}5)^2}{4}\right)^{174}\left(\frac{(0{,}5)^2}{4}\right)^{1}\left(\frac{0{,}5(1-0{,}5)}{4}\right)^{25}}\right]$$

$$LOD = \text{Log}_{10}\left[\frac{(0{,}21739)^{174}\left(1{,}1390x10^{-3}\right)^{1}(0{,}01573)^{25}}{(0{,}0625)^{174}(0{,}0625)^{1}(0{,}0625)^{25}}\right]$$

$$LOD = \text{Log}_{10}\left[\frac{4{,}5125x10^{-164}}{1{,}4997x10^{-241}}\right]$$

$$LOD = \text{Log}_{10}(3x10^{77})$$

$$LOD = 77{,}48$$

No exemplo apresentado, foi possível deduzir sobre a fase de ligação para ambos os genitores analisados. Entretanto, em algumas situações isto não é possível, uma vez que locos informativos ou parcialmente informativos podem estar envolvidos. Em tais situações, um procedimento recomendado é estimar a porcentagem de recombinação e os respectivos LOD escores para cada combinação de fase de ligação. Assim, a combinação que apresentar maior valor de LOD e menor estimativa de porcentagem de recombinação é deduzida como a mais provável para os locos em questão. Entretanto, deve-se ressaltar que este método é subjetivo e seu sucesso pode variar entre as diferentes configurações de acasalamento (Maliepaard et al., 1997).

Na Tabela 8 pode-se observar as diferentes estimativas de distâncias e LOD para os locos C_{11} e C_{12}. Verifica-se que a configuração Ap x Ap apresenta maior LOD escore e menor porcentagem de recombinação que as demais configurações, concordando com os resultados obtidos anteriormente.

Tabela 7. Classes genotípicas, freqüências observadas e esperadas para a progênie do cruzamento $A_1A_2B_1B_2$ x $A_3A_4B_3B_4$ considerando as marcas C_{11} e C_{12}

Classes Genotípicas	Observado	Esperado	Esperado (r=0,5)
13-13	48	P^2	12,5
13-14	2	PR	12,5
13-23	5	PR	12,5
13-24	1	R^2	12,5
14-13	1	PR	12,5
14-14	45	P^2	12,5
14-23	0	R^2	12,5
14-24	3	PR	12,5
23-13	3	PR	12,5
23-14	0	R^2	12,5
23-23	33	P^2	12,5
23-24	6	PR	12,5
24-13	0	R^2	12,5
24-14	0	PR	12,5
24-23	5	PR	12,5
24-24	48	P^2	12,5
Total	200	1	200

Tabela 8. Estimativas da porcentagem de recombinação estimadas com base nas freqüências conjuntas da progênie – para diferentes combinações de fase de ligação – e com base nas freqüências marginais dos genitores.

Marcas: C_{11} e C_{12}	Tipo de Acasalamento: (1:1:1:1)(1:1:1:1)	
	Freqüência Conjunta	
Fase de Ligação	% de Recombinação (r)	LOD escore
Ap x Ap	6,7	77,482
Ap x Re	49,3	0,02
Re x Ap	50,6	0,019
Re x Re	51,5	4,432
	Freqüência Marginal	
r_{G1} = 6 %	r_{G2} = 7,5 %	r_{med} = 6,75 %

Ap: Aproximação; Re: Repulsão

4.2. Cenário 2

4.2.1. Teste de Segregação

Na Tabela 9, verifica-se a análise de segregação para marcas individuais com base no teste qui-quadrado (χ^2) para uma progênie de 200 indivíduos obtidas de genitores gerados ao acaso. As marcas analisadas segregaram em diferentes proporções, de modo que são encontrados locos com segregação completamente informativa (1:1:1:1), informativa (1:1) ou parcialmente informativa(1:2:1). A marca C_{29} não segregou conforme o esperado, e deveria ser descartada evitando prejudicar análises futuras, uma vez que ela apresentou distorção. Porém, como o intuito deste estudo de caso é observar os vários tipos de cruzamentos e segregações existentes na população simulada, esta marca foi mantida nas análises apenas para fins didáticos.

A distorção na razão de segregação esperada pode ser atribuída a diversas causas, tais como: i) processos de seleção no estádio de gameta ou zigoto (Zamir e Tadmor, 1986); ii) locos gênicos que apresentam seleção natural, locos próximos a genes que levam a menor viabilidade de gametas, como observado em arroz por He et al. (2001) e; iii) conversão gênica.

Quando uma célula diplóide sofre meiose são produzidas quatro células haplóides, exatamente metade dos alelos nestas células deveria ser de origem materna (alelos que a célula diplóide recebeu de sua mãe) e a outra metade paterna (alelos que a célula diplóide recebeu de seu pai). Porém, estudos demonstram que em alguns organismos, como fungos,por exemplo, que o padrão de segregação esperado têm sido violado. Ocasionalmente, as meioses produzem três copias de alelos maternais e somente uma cópia do alelo paternal. Este fenômeno é conhecido com conversão gênica. A conversão gênica sempre ocorre em associação com eventos de recombinação genética de homólogos, e acredita-se ser uma conseqüência direta de mecanismos gerais de recombinação e reparo do DNA (Alberts et al., 2002).

Tabela 9. Teste de segregação dos 33 marcadores da população não- completamente informativa, ao nível de 1% de probabilidade pelo Teste de χ^2.

		Classes Genotípicas									
Marcador	Pais	11	12	13	22	23	Hipótese	GL	Qui-quadrado	Probabilidade(%)	
C11	12x12	55	88		57		1:2:1	2	2,92	23,2236	ns
C12	12x22		90		110		1:1	1	2,00	15,7299	ns
C13	12x22		92		108		1:1	1	1,28	25,7899	ns
C14	12x12	50	99		51		1:2:1	2	0,03	98,5112	ns
C15	12x12	53	97		50		1:2:1	2	0,27	87,3716	ns
C16	12x23		55	35	59	51	1:1:1:1	3	6,64	8,4302	ns
C17	12x23		49	48	53	50	1:1:1:1	3	0,28	96,3745	ns
C18	11x12	104	96				1:1	1	0,32	57,1608	ns
C19	12x22		91		109		1:1	1	1,62	20,3092	ns
C110	12x23		45	48	54	53	1:1:1:1	3	1,08	78,1904	ns
C111	12x22		101		99		1:1	1	0,02	88,7537	ns
C21	12x12	54	98		48		1:2:1	2	0,44	80,2519	ns
C22	12x23		43	52	54	51	1:1:1:1	3	1,40	70,5535	ns
C23	12x23		38	55	57	50	1:1:1:1	3	4,36	22,5123	ns
C24	12x22		93		107		1:1	1	0,98	32,2199	ns
C25	12x13	47	59	52		42	1:1:1:1	3	3,16	36,7608	ns
C26	11x23		103	97			1:1	1	0,18	67,1373	ns
C27	11x23		97	103			1:1	1	0,18	67,1373	ns
C28	12x13	57	47	34		62	1:1:1:1	3	9,16	2,7238	ns
C29	12x23		60	30	51	59	1:1:1:1	3	11,64	0,8724	*
C210	12x22		102		98		1:1	1	0,08	77,7297	ns
C211	12x11	100	100				1:1	1	0,00	100,0	ns
C31	12x13	55	56	40		49	1:1:1:1	3	3,24	35,6081	ns
C32	12x12	53	103		44		1:2:1	2	0,99	60,9571	ns
C33	11x22		200				1	-	-	-	-
C34	12x23		56	42	58	44	1:1:1:1	3	4,00	26,1464	ns
C35	12x13	52	55	44	49		1:1:1:1	3	1,32	72,4389	ns
C36	12x23		53	42	57	48	1:1:1:1	3	2,52	47,1688	ns
C37	12x12	54	92		54		1:2:1	2	1,28	52,7293	ns
C38	12x22		101		99		1:1	1	0,02	88,7537	ns
C39	12x22		103		97		1:1	1	0,18	67,1373	ns
C310	12x12	55	94		51		1:2:1	2	0,88	64,4036	ns
C311	12x23		47	52	59	42	1:1:1:1	3	3,16	36,7608	ns

*: Valor abaixo do nível crítico especificado

C_{ij}: indica se tratar de um marcador que, no genoma simulado, se encontra no grupo de ligação i, e é o marcador j, (i = 1, 2, 3 ; e j = 1, 2, ..., 11)

4.2.2. Porcentagem de recombinação entre pares de marcadores

Como discutido anteriormente, na análise de pares de marcas que envolvem combinações do tipo (1:1:1:1)(1:1:1:1), os haplótipos que geraram a prole podem ser reconstruídos completamente, sendo possível distinguir os gametas de origem parental e

aqueles oriundos de recombinação, produzindo uma progênie totalmente informativa. Entretanto, para os dados do Cenário 2, nem todas as combinações serão totalmente informativas. Diferentes combinações podem surgir ente locos completamente informativos, informativos e parcialmente informativos.

De posse da população simulada, para exemplificar o cálculo da porcentagem de recombinação, foram obtidas nove situações diferentes. Um “caso 10” foi adicionado, apesar de não ter ocorrido na população apenas para fins de ilustração e maiores esclarecimentos. Para cada caso, serão discutidas as maneiras como as estimativas de porcentagem de recombinação podem ser calculadas.

Os casos analisados são enumerados a seguir:

1) Caso (1:1:1:1) (1:1:1:1);

2) Caso (1:1:1:1) (1:2:1);

3) Caso (1:2:1)(1:1:1:1);

4) Caso (1:2:1)(1:2:1);

5) Caso (1:2:1)(1:1);

6) Caso (1:1) (1:2:1);

7) Caso (1:1) (1:1);

8) Caso (1:1) (1:1:1:1);

9) Caso (1:1:1:1) (1:1);

10) Caso (1:1)(1:1) (extra);

Na Tabela 10, está ilustrada cada segregação encontrada na população, os tipos de classes esperadas que ocorrem em cada tipo de segregação e os marcadores da população que se enquadram em cada um destes tipos de segregações.

Tabela 10. Tipos de segregação encontrados na população simulada, marcadores pertencentes a cada um destes tipos e as classes de probabilidades esperadas em cada uma das segregações.

Segregação	Marcadores	Classes
1. (1:1:1:1) (1:1:1:1)	C_{22} X C_{23} ; C_{311} X C_{25} C_{28} X C_{29} ; C_{34} X C_{35} C_{35} X C_{36}	P^2; PR; R^2
2. (1:1:1:1) (1:2:1)	C_{31} X C_{32} ; C_{36} X C_{37}	P^2; PR; R^2; 2PR; P^2+R^2
3. (1:2:1)(1:1:1:1)	C_{310} X C_{16} ; C_{21} X C_{22} C_{32} X C_{34}	P^2; PR; R^2; 2PR; P^2+R^2
4. (1:2:1)(1:2:1)	C_{14} X C_{15} ; C_{15} X C_{310}	P^2; R^2; 2PR; $2P^2$+$2R^2$
5. (1:2:1)(1:1)	C_{11} X C_{12} ; C_{37} X C_{38}	P^2 + PR; R^2 + PR; P^2 + R^2 + 2PR
6. (1:1) (1:2:1)	C_{13}X C_{14} ; C_{111} X C_{21}	P^2 + PR; R^2 + PR; P^2 + R^2 + 2PR
7. (1:1) (1:1)	C_{12} X C_{13} ; C_{210} X C_{211} C_{38} X C_{39}	$2P^2$ + 2PR; $2R^2$ + 2PR
8. (1:1) (1:1:1:1)	C_{18} X $C_{110;}$ C_{24} X C_{311} C_{27} X C_{28} ; C_{211} X C_{31}	P^2 + PR; R^2 + PR
9. (1:1:1:1) (1:1)	C_{17} X C_{18}; C_{110}X C_{19} ; C_{23}X C_{24}; C_{25} X C_{26} ;	P^2 + PR; R^2 + PR
10. (1:1)(1:1) (extra)	-	-

Notação: C_{ij} indica se tratar de um marcador que, no genoma simulado, se encontra no grupo de ligação i, e é o marcador j, (i = 1, 2, 3 ; e j = 1, 2, ..., 11)

Para facilitar o entendimento das expressões de máxima verossimilhança ao longo desta simulação foi reunida na Tabela 11 todas as classes de probabilidades esperadas e a sua estimativa. Estes valores, ali apresentados, foram os utilizados posteriormente nas funções de Verossimilhança.

Tabela 11. Classes de probabilidades e respectivos estimadores utilizados na função de Verossimilhança.

Probabilidades	Estimadores
P^2	$\left(\frac{1-r}{2}\right)\left(\frac{1-r}{2}\right)=\left(\frac{1-r}{2}\right)^2=\frac{(1-r)^2}{4}$
R^2	$\left(\frac{r}{2}\right)\left(\frac{r}{2}\right)=\frac{r^2}{4}$
PR	$\left(\frac{1-r}{2}\right)\left(\frac{r}{2}\right)=\frac{r(1-r)}{4}$
2PR	$2\frac{r(1-r)}{4}=\frac{r(1-r)}{2}$
$P^2 + R^2$	$\frac{(1-r)^2}{4}+\frac{r^2}{4}=\frac{2r^2-2r+1}{4}$
$2\ P^2 + 2\ R^2$	$2\frac{2r^2-2r+1}{4}=\frac{2r^2-2r+1}{2}$
P^2 + PR	$\frac{(1-r)^2}{4}+\frac{r(1-r)}{4}=\frac{1-r}{4}$
R^2 + PR	$\left(\frac{r^2}{4}\right)+\left(\frac{r(1-r)}{4}\right)=\frac{r}{4}$
$P^2 + R^2 + 2$ PR	$\left(\frac{1-r}{4}\right)+\left(\frac{r}{4}\right)=\frac{1}{4}$
$2\ P^2 + 2$ PR	$2\left(\frac{1-r}{4}\right)=\frac{1-r}{2}$
$2\ R^2 + 2$ PR	$2\left(\frac{r}{4}\right)=\frac{r}{2}$

A seguir será feito um estudo, caso a caso, das segregações utilizando como exemplo um par de marcadores, dos 33 existentes, que se enquadra em cada caso e algumas considerações adicionais também serão descritas:

1) Caso (1:1:1:1) (1:1:1:1): Marcas C_{22} x C_{23}

Cruzamento Tipo: $A_1A_2B_1B_2$ x $A_2A_3B_2B_3$

Deste cruzamento tem-se:

A_1A_2 x A_2A_3	B_1B_2 x B_2B_3
↓	↓
1 A_1 A_2	1 B_1 B_2
1 A_1 A_3	1 B_1 B_3
1 A_2 A_2	1 B_2 B_2
1 A_2 A_3	1 B_2 B_3

Este caso aborda a segregação entre dois locos completamente informativos. É o caso abordado no cenário 1, então será abordado aqui de forma resumida.

Para esta situação a função de Verossimilhança utilizada é a seguinte:

$$L(r;n_i) = \lambda\left(P^2\right)^{n1}(PR)^{n2}\left(R^2\right)^{n3}$$

A partir desta função pode-se estimar a freqüência de recombinação através da expressão a seguir, como já foi demonstrado no cenário 1.

$$r = \frac{2n_3 + n_2}{2N}$$

Para os marcadores C_{22} x C_{23} tem-se:

$$r = \frac{2(2) + 38}{2(200)} = 0{,}105 = 10{,}5\%$$

Na Tabela 12 está representada a freqüência de recombinação obtida entre pares de marcadores da população simulada que se enquadram no caso (1:1:1:1) (1:1:1:1).

Tabela 12. Freqüência de recombinação dos cinco pares de marcadores que possuem os dois locos completamente informativos.

Marcadores	Cruzamento	Freqüência de Recombinação
C_{22} x C_{23}	12 12 x 23 23	$r = \frac{2(2) + 38}{2(200)} = 0{,}105 = 10{,}5\%$
C_{28} x C_{29}*	12 12 x 13 23	$r = \frac{2(1) + 42}{2(200)} = 0{,}11 = 11\%$
C_{34} x C_{35}	12 12 x 23 13	$r = \frac{2(1) + 39}{2(200)} = 0{,}1025 = 10{,}25\%$
C_{35} x C_{36}	12 12 x 13 23	$r = \frac{2(1) + 38}{2(200)} = 0{,}1 = 10\%$

* Marcador C_{29}: Apresentou distorção de segregação

Esta forma de cruzamento esquematizada anteriormente pode ser explicada a seguir:

Como exemplo, tem-se o primeiro cruzamento $12^{(1)}\ 12^{(2)}$ x $23^{(1)}\ 23^{(2)}$.

$12^{(1)}$: Loco 1(marcador C_{22}) do genitor 1

$12^{(2)}$: Loco 2(marcador C_{23}) do genitor 1

$23^{(1)}$: Loco 1(marcador C_{22}) do genitor 2

$23^{(2)}$: Loco 2(marcador C_{23}) do genitor 2

Outra forma de representar este cruzamento é: $A_1A_2B_1B_2$ x $A_2A_3B_2B_3$

2) Caso (1:1:1:1)(1:2:1): Marcas C_{31} x C_{32}

Cruzamento Tipo: $A_1A_2B_1B_2$ x $A_1A_3B_1B_2$

Deste cruzamento tem-se:

A_1A_2 x A_1A_3	B_1B_2 x B_1B_2
↓	↓
1 A_1 A_1	1 B_1 B_1
1 A_1 A_3	2 B_1 B_2
1 A_1 A_2	1 B_2 B_2
1 A_2 A_3	

Este caso aborda a segregação entre um loco completamente informativo e um loco parcialmente informativo. Verifica-se na Tabela 13A que para o loco B (12 x 12) não é possível distinguir os genótipos do tipo 12, quanto à origem dos alelos dos parentais. Dessa forma, os genótipos 11-12, 12-12, 13-12 e 23-12 estão confundidos em mais de uma classe genotípica, não sendo possível distingui-las. O fato de se encontrar genótipos confundidos em mais de uma classe pode reduzir a precisão na estimativa da distância entre os dois locos, uma vez que um menor número de classes será utilizado para a obtenção das estimativas da porcentagem de recombinação. Além disso, o valor de r deixa de ser a média entre os valores marginais da tabela.

Na tabela 13B são apresentadas as 12 classes genotípicas detectáveis para este tipo de cruzamento, suas freqüências observadas e esperadas. Verifica-se, comparando-se ao cruzamento envolvendo dois locos completamente informativos, que menos classes genotípicas são utilizadas para a obtenção das estimativas de porcentagem de recombinação.

Tabela 13A. Cruzamento Tipo: $A_1A_2B_1B_2$ x $A_1A_3B_1B_2$

Gametas		A_1B_1 P	A_1B_2 R	A_3B_1 R	A_3B_2 P	Freqüência Gamética	Total de indivíduos
A_1B_1	P	P^2 11-11 45	PR 11-12 (10)*	PR 13-11 3	P^2 13-12 (35)	(1-r)/2	48
A_1B_2	R	PR 11-12 (10)	R^2 11-22 0	R^2 13-12 (35)	PR 13-22 2	r/2	2
A_2B_1	R	PR 12-11 4	R^2 12-12 (47)	R^2 23-11 1	PR 23-12 (11)	r/2	5
A_2B_2	P	P^2 12-12 (47)	PR 12-22 5	PR 23-12 (11)	P^2 23-22 37	(1-r)/2	42
Freq. Gamética		(1-r)/2	r/2	r/2	(1-r)/2		
Total		49	5	4	39		97

*Classes genotípicas entre parênteses indicam classes não distinguíveis.

Tabela 13B. Cruzamento Tipo: $A_1A_2B_1B_2$ x $A_1A_3B_1B_2$

Classes Genotípicas	Observado	Esperado	Esperado (r=0,5)
11-11	45	P^2	12,5
11-12	10	PR+PR	25
11-22	0	R^2	12,5
12-11	4	PR	12,5
12-12	47	P^2+R^2	25
12-22	5	PR	12,5
13-11	3	PR	12,5
13-12	35	P^2+R^2	25
13-22	2	PR	12,5
23-11	1	R^2	12,5
23-12	11	PR+PR	25
23-22	37	P^2	12,5
Total	200	1	200

De posse das freqüências esperadas de cada classe genotípica e do número de indivíduos observados em cada uma destas classes pode-se obter uma expressão de Máxima Verossimilhança, de forma a obter uma fórmula geral para o cálculo da porcentagem de recombinação na situação em que envolve cruzamento do tipo (1:1:1:1)(1:2:1).

Sendo assim tem-se:

$$L(r;n_i) = \lambda\left(P^2\right)^{n1}(PR)^{n2}\left(R^2\right)^{n3}(PR+PR)^{n4}(P^2+R^2)^{n5}$$

$$L(r;n_i) = \lambda\left(\frac{1}{4}(1-r)^2\right)^{n1}\left[\frac{1}{4}r(1-r)\right]^{n2}\left(\frac{1}{4}r^2\right)^{n3}\left[\frac{1}{2}r(1-r)\right]^{n4}\left[\frac{1}{4}(2r^2-2r+1)\right]^{n5}$$

$$L(r;n_i) = \lambda\left((1-r)^2\right)^{n1}\left[r(1-r)\right]^{n2}\left(r^2\right)^{n3}\left[r(1-r)\right]^{n4}\left[2r^2-2r+1\right]^{n5}$$

A função suporte será:

$$\ell(r,n_i) = \ln\lambda + \left[2n_1\ln(1-r)\right] + n_2\ln(1-r) + n_2\ln(r) + 2n_3\ln(r) + n_4\ln(1-r) + n_4\ln(r) + n_5\ln(2r^2-2r+1)$$

$$\ell(r,n_i) = \ln\lambda + (2n_1+n_2+n_4)\ln(1-r) + (n_2+2n_3+n_4)\ln(r) + n_5\ln(2r^2-2r+1)$$

O valor máximo da função é obtido por meio de:

$$\frac{\partial\ell(r;n_i)}{\partial r} = \frac{-(2n_1+n_2+n_4)}{1-r} + \frac{(n_2+2n_3+n_4)}{r} + \frac{n_5(4r-2)}{2r^2-2r+1}$$

$$\frac{-(2n_1+n_2+n_4)}{1-r} + \frac{(n_2+2n_3+n_4)}{r} + \frac{n_5(4r-2)}{2r^2-2r+1} = 0$$

$$\left[-(2n_1+n_2+n_4)(2r^3-2r^2+r)\right] + \left[(n_2+2n_3+n_4)(-2r^3+4r^2-3r+1)\right] + \left[n_5(-4r^3+6r^2-2r)\right] = 0$$

$$r^3\left[-2(2n_1+n_2+n_4)-2(n_2+2n_3+n_4)-4n_5\right] + r^2\left[2(2n_1+n_2+n_4)+4(n_2+2n_3+n_4)+6n_5\right]$$
$$+ r\left[-(2n_1+n_2+n_4)-3(n_2+2n_3+n_4)-2n_5\right] + (n_2+2n_3+n_4) = 0$$

$$r^3\left[-4n_1-4n_2-4n_3-4n_4-4n_5)\right] + r^2\left[4n_1+6n_2+8n_3+6n_4+6n_5)\right] + r\left[-2n_1-4n_2-6n_3-4n_4-2n_5)\right] + (n_2+2n_3+n_4) = 0$$

$$r^3\left[-4N\right] + r^2\left[2(2N+n_2+2n_3+n_4+n_5)\right] + r\left[-2(N+n_2+2n_3+n_4)\right] + (n_2+2n_3+n_4) = 0$$

Substituindo os valores da Tabela 13B tem-se:

$$r^3\left[-800\right] + r^2\left[2(400+14+2+21+82)\right] + r\left[-2(200+14+2+21)\right] + (14+2+21) = 0$$

$$r^3\left[-800\right] + r^2\left[1038\right] + r\left[-474\right] + (37) = 0$$

Ao obter a raiz do polinômio, real dentro do intervalo de 0 a 50%, tem-se que: r = 9,71%

Uma das formas de obtenção da raiz do polinômio de terceiro grau pode ser dada usando o Método de Tartaglia (ou Método de Cardano) para a obtenção das raízes de uma equação de 3° grau. Este Método será apresentado de forma resumida a seguir:

Uma equação geral do terceiro grau na variável x é dada por:

$$a\,x^3 + b\,x^2 + c\,x + d = 0$$

e se o coeficiente a do termo do terceiro grau é não nulo, divide-se esta equação por a para obter:

$$x^3 + (b/a)\,x^2 + (c/a)\,x + (d/a) = 0$$

e assim considera-se só as equações em que o coeficiente de x^3 seja igual a 1, isto é, equações da forma geral:

$$x^3 + A\,x^2 + B\,x + C = 0$$

onde A=b/a, B=c/a e C=d/a. Fazendo a substituição de translação: x = y-A/3

na equação acima, obtem-se:

$$y^3 + (B-A^2/3)\,y + (C-AB/3+2A^3/27) = 0$$

e tomando $p=(B-A^2/3)$ e $q=C-AB/3+(2/27)A^3$, pode-se simplificar a equação do terceiro grau na variável y, para:

$$y^3 + p\,y + q = 0$$

Como toda equação desta forma possui pelo menos uma raiz real, procura-se esta raiz na forma y=u+v. Substituindo y por u+v, na última equação, obtem-se:

$$(u+v)^3 + p(u+v) + q = 0$$

o que equivale a

$$u^3 + v^3 + 3uv(u+v) + p(u+v) + q = 0$$

ou seja

$$u^3 + v^3 + (3uv+p)(u+v) + q = 0$$

Usando esta última equação e impondo a condição para que:

$$p = -3uv \quad e \quad q= -(u^3+v^3)$$

obtem-se valores de u e v para os quais y=u+v deverá ser uma raiz da equação. Estas últimas condições implicam que:

$$u^3\,v^3=-p^3/27 \quad e \quad u^3+v^3 = -q$$

Considerando u^3 e v^3 como variáveis, o problema equivale a resolver uma equação do 2°. grau da forma:

$$z^2 - S\,z + P = 0$$

onde

S = soma das raízes = $u^3 + v^3$

P = produto das raízes = $u^3\,v^3$

Resolvendo a equação do 2o. grau:

$$z^2 + q\,z - p^3/27 = 0$$

para obter as partes u e v da primeira raiz:

$r1 = u + v$

Com o discriminante desta última equação, definido por:

$D = q^2/4 + p^3/27$

e utilizando a fórmula de Bhaskara , obtem-se:

$u^3 = -q/2 + D^{1/2}$

$v^3 = -q/2 - D^{1/2}$

A primeira raiz r_1 da equação original

$x^3 + A\,x^2 + B\,x + C = 0$

depende da translação realizada no início e será dada por:

$r_1 = u + v - A/3$

É importante enfatizar que este método continua com expressões para obter todas as três raízes do polinômio, sendo uma real e as outras duas não-reais. Porém aqui só foram ilustrados os passos pra obtenção da primeira raiz, ou raiz real, visto que as outras não teriam utilidade para fins de mapeamento.

Para ilustrar a aplicação deste método considera-se a equação apresentada anteriormente e obter a freqüência de recombinação.

Dado $r^3[-800]+r^2[1038]+r[-474]+(37)=0$ tem-se:

Passo1: Dividir toda a equação por -800

$r^3 + r^2[-1{,}2975]+r[0{,}5925]+(-0{,}04625)=0$

de forma que obtem-se uma equação do tipo $r^3 + Ar^2 + Br + C = 0$

Passo 2: Fazer a substituição de translação: $x = y-A/3$

na equação acima, obtem-se:

$y^3 + (B-A^2/3)\,y + (C-AB/3+2A^3/27) = 0$

sendo que $p =(B-A^2/3)$, e substituindo os valores temos p= 0,03133

$q= (C-AB/3+2A^3/27)$, e substituindo os valores temos que q= 0,0482

Passo 3: Obter o discriminante $D = q^2/4 + p^3/27$

$D = 5{,}819 \times 10^{-4}$

$\sqrt{D} = 0{,}0241236$

Passo 3: Obter os valores de u e v:

$u^3 = -q/2 + D^{1/2}$

substituindo tem-se que $u^3= 2{,}36\times10^{-5}$ e $u = \sqrt[3]{u^3} = 0{,}02868$

de forma similar temos que

$v^3 = -q/2 - D½$

$v^3 = -0,04822$ e $v = -0,36398$

Passo 4: Obter a raiz real que é dada por: $r_1 = u + v - A/3$

$r_1 = 0,02868 + (-0,36398) – (-1,2975 / 3) = 0,09719$ ou 9,71%

Como não se conhece a fase de ligação dos genitores, estima-se as frequências de recombinações das 4 possíveis fases e aquela que possuir o maior valor de LOD escore é considerada como sendo a verdadeira fase de ligação, como está representado na Tabela 13C, onde pode se observar que o menor valor de freqüência de recombinação (9,8 %) estava associado ao maior valor de LOD (42,561), sendo portanto, a fase de ligação em aproximação nos dois genitores.

Tabela 13C. Cruzamento Tipo: $A_1A_2B_1B_2$ x $A_1A_3B_1B_2$

Marcas: C_{31} e C_{32}	Tipo de Acasalamento: (1:1:1:1)(1:2:1)	
	Freqüência Conjunta	
Fase de Ligação	% de Recombinação	LOD escore
Ap x Ap	9,8	42,561
Ap x Re	49,4	0,011
Re x Ap	50	0,011
Re x Re	50	2,098
	Freqüência Marginal	
r_{G1} = 7,21 %	r_{G2} = 9,27 %	r_{med} = 8,24 %

Na Tabela 13D está representado as frequências de recombinação obtida entre pares de marcadores da população simulada que se enquadram no caso (1:1:1:1) (1:2:1).

Tabela 13D. Freqüência de recombinação dos dois pares de marcadores que possuem cruzamentos do tipo (1:1:1:1) (1:2:1).

Marcadores	**Cruzamento**	**Freqüência de Recombinação**
C_{31} X C_{32}	12 12 x 13 12	9,71%
C_{36} X C_{37}	12 12 x 23 12	0,0971=9,71%

3) Caso (1:2:1) (1:1:1:1): Marcas C_{21} e C_{22}

Cruzamento Tipo: $A_1A_2B_1B_2$ x $A_1A_2B_2B_3$

Deste cruzamento tem-se:

A_1A_2 x A_1A_2	B_1B_2 x B_2B_3
↓	↓
1 A_1 A_1	1 B_1 B_2
2 A_1 A_2	1 B_1 B_3
1 A_2 A_2	1 B_2 B_2
	1 B_2 B_3

Este caso aborda a segregação entre um loco completamente informativo e um loco parcialmente informativo. Verifica-se na Tabela 14A que para a marca C_{21} (12 x 12) não é possível distinguir os genótipos do tipo 12, quanto à origem dos alelos dos parentais. Dessa forma, os genótipos 12-12, 12-13, 12-22, 12-23, estão confundidos em mais de uma classe genotípica, não sendo possível distinguí-los. O fato de se encontrar genótipos confundidos em mais de uma classe pode reduzir a precisão na estimativa da distância entre os dois locos.

Na Tabela 14B são apresentadas as 12 classes genotípicas detectáveis para este tipo de cruzamento, suas freqüências observadas e esperadas. Verifica-se, comparando-se ao cruzamento envolvendo dois locos completamente informativos, que menos classes genotípicas são utilizadas para a obtenção das estimativas de porcentagem de recombinação.

Tabela 14A. Cruzamento Tipo: $A_1A_2B_1B_2$ x $A_1A_2B_2B_3$

Gametas		A_1B_2 P	A_1B_3 R	A_2B_2 R	A_2B_3 P	Freqüência Gamética	Total de indivíduos
A_1B_1	P	$P^2_{11\text{-}12}$ 39	$PR_{11\text{-}13}$ 5	$PR_{12\text{-}12}$ (4)*	$P^2_{12\text{-}13}$ (44)	(1-r)/2	44
A_1B_2	R	$PR_{11\text{-}22}$ 9	$R^2_{11\text{-}23}$ 1	$R^2_{12\text{-}22}$ (42)	$PR_{12\text{-}23}$ (8)	r/2	10
A_2B_1	R	$PR_{12\text{-}12}$ (4)	$R^2_{12\text{-}13}$ (44)	$R^2_{22\text{-}12}$ 0	$PR_{22\text{-}13}$ 3	r/2	3
A_2B_2	P	$P^2_{12\text{-}22}$ (42)	$PR_{12\text{-}23}$ (8)	$PR_{22\text{-}22}$ 3	$P^2_{22\text{-}23}$ 42	(1-r)/2	45
Freq. Gamética		(1-r)/2	r/2	r/2	(1-r)/2		
Total		48	6	3	45		102

* Classes genotípicas entre parênteses indicam classes não distinguíveis.

Tabela 14B. Cruzamento Tipo: $A_1A_2B_1B_2$ x $A_1A_2B_2B_3$

Classes Genotípicas	Observado	Esperado	Esperado (r=0,5)
11-12	39	P^2	12,5
11-13	5	PR	12,5
11-22	9	PR	12,5
11-23	1	R^2	12,5
12-12	4	PR+PR	25
12-13	44	P^2+R^2	25
12-22	42	P^2+R^2	25
12-23	8	PR+PR	25
22-12	0	R^2	12,5
22-13	3	PR	12,5
22-22	3	PR	12,5
22-23	42	P^2	12,5
Total	200	1	200

Na Tabela 14C são observados os valores de estimativas de distâncias com base no método da máxima verossimilhança, para diferentes configurações de fase de ligação, baseadas na distribuição conjunta da progênie. Neste caso, a configuração Ap x Ap apresentou maior estimativa de LOD e menor porcentagem de recombinação. Este fato pode ser confirmado pela análise das freqüências marginais dos gametas parentais. Constata-se que os dois parentais apresentam-se em configuração *cis* para os locos em análise.

Da mesma forma que na estimação com as freqüências conjuntas, a estimativa com base nas freqüências marginais também perde conteúdo de informação pela presença de classes indistinguíveis. Na Tabela 14C, são apresentadas as estimativas de porcentagem de recombinação para ambos os parentais e o seu valor médio. Ao contrário do observado para acasalamentos envolvendo locos completamente informativos, a estimativa da porcentagem de recombinação com base na freqüência conjunta não se iguala à média das estimativas das freqüências marginais.

Para obtenção dos valores da porcentagem de recombinação conjunta utilizando a função de máxima verossimilhança utiliza-se a mesma expressão apresentada para o acasalamento (1:1:1:1) (1:2:1), sendo ela escrita a seguir :

$$L(r;n_i) = \lambda\left(P^2\right)^{n1}(PR)^{n2}\left(R^2\right)^{n3}(PR+PR)^{n4}(P^2+R^2)^{n5}$$

$$r^3\left[-4N\right]+r^2\left[2(2N+n_2+2n_3+n_4+n_5)\right]+r\left[-2(N+n_2+2n_3+n_4)\right]+(n_2+2n_3+n_4)=0$$

Para as marcas C_{21} e C_{22} a freqüência de recombinação é demonstrada a seguir. A obtenção da raiz do polinômio é feita com aplicação do Método de Tartaglia apresentado anteriormente.

$$r^3\left[-800)\right]+r^2\left[2(400+20+2+12+86)\right]+r\left[-2(200+20+2+12)\right]+(20+2+12)=0$$

$$r^3\left[-800)\right]+r^2\left[1040\right]+r\left[-468\right]+34=0$$

$$r = 0{,}089 = 8{,}9\%$$

Na Tabela 14D está representada a frequência de recombinação obtida entre pares de marcadores da população simulada que se enquadram no caso (1:2:1). (1:1:1:1)

Tabela 14C. Cruzamento Tipo: $A_1A_2B_1B_2$ x $A_1A_2B_2B_3$

Marcas: C_{21} e C_{22}	Tipo de Acasalamento: (1:2:1)(1:1:1:1)	
	Freqüência Conjunta	
Fase de Ligação	% de Recombinação	LOD escore
Ap x Ap	8,9	44,342
Ap x Re	50	0,039
Re x Ap	48,9	0,04
Re x Re	50	2,075
	Freqüência Marginal	
r_{G1} = 12,74 %	r_{G2} = 8,82 %	r_{med} = 10,78 %

Tabela 14D. Freqüência de recombinação de três pares de marcadores que possuem cruzamentos do tipo (1:2:1) (1:1:1:1).

Marcadores	**Cruzamento**	**Freqüência de Recombinação**
C_{21} X C_{22}	12 12 x 12 23	8,9
C_{32} X C_{34}	12 12 x 12 23	14,30

4) Caso (1:2:1)(1:2:1): Marcas C_{14} e C_{15}

Cruzamento Tipo: $A_1A_2B_1B_2$ x $A_1A_2B_1B_2$

A_1A_2 x A_1A_2	B_1B_2 x B_1B_2
↓	↓
1 A_1 A_1	1 B_1 B_1
2 A_1 A_2	2 B_1 B_2
1 A_2 A_2	1 B_2 B_2

Neste caso é apresentada a análise entre dois locos parcialmente informativos. Nesta situação, os indivíduos com o genótipo 12 são indistinguíveis, para ambas as marcas C_{14} e C_{15}, com relação à origem de seus alelos nos parentais (Tabela 15A).

É possível estimar a porcentagem de recombinação com base na distribuição marginal de ambos os genitores, conforme mostrado na Tabela 15A, entretanto, o confundimento das classes indistinguíveis pode reduzir a precisão da estimativa. Da mesma forma, a estimativa de Máxima Verossimilhança para a distribuição conjunta na progênie apresenta classes sobrepostas. Verifica-se que as estimativas com base na freqüência marginal e conjunta diferem consideravelmente.

Tabela 15A. Cruzamento Tipo: $A_1A_2B_1B_2$ x $A_1A_2B_1B_2$

Gametas		A_1B_1 P	A_1B_2 R	A_2B_1 R	A_2B_2 P	Freqüência Gamética	Total de indivíduos
A_1B_1	P	P^2 11-11 41	PR 11-12 (9)	PR 12-11 (12)	P^2 12-12 (79)	(1-r)/2	41
A_1B_2	R	PR 11-12 (9)	R^2 11-22 0	R^2 12-12 (79)	PR 12-22 (8)	r/2	0
A_2B_1	R	PR 12-11 (12)	R^2 12-12 (79)	R^2 22-11 0	PR 22-12 (9)	r/2	0
A_2B_2	P	P^2 12-12 (79)	PR 12-22 (8)	PR 22-12 (9)	P^2 22-22 42	(1-r)/2	42
Freq. Gamética		(1-r)/2	r/2	r/2	(1-r)/2		
Total		41	0	0	42		83

Tabela 15B. Cruzamento Tipo: $A_1A_2B_1B_2$ x $A_1A_2B_1B_2$

Classes Genotípicas	Observado	Esperado	Esperado (r=0,5)
11-11	41	P^2	12,5
11-12	9	PR+RP	25
11-22	0	R^2	12,5
12-11	12	PR+RP	25
12-12	79	$P^2+R^2+R^2+P^2$	50
12-22	8	RP+PR	25
22-11	0	R^2	12,5
22-12	9	RP+PR	25
22-22	42	P^2	12,5
Total	200	1	200

De posse das classes e número de indivíduos observados em cada uma destas classes podemos obter uma expressão de Máxima Verossimilhança, de forma a obter uma fórmula geral para o cálculo da porcentagem de recombinação na situação em que envolve cruzamento do tipo (1:2:1)(1:2:1).

Sendo assim tem-se:

$$L(r;n_i) = \lambda\left(P^2\right)^{n1}(2PR)^{n2}\left(R^2\right)^{n3}(2P^2+2R^2)^{n4}$$

$$L(r;n_i) = \lambda\left(\frac{1}{4}(1-r)^2\right)^{n1}\left[\frac{1}{2}r(1-r)\right]^{n2}\left(\frac{1}{4}r^2\right)^{n3}\left[\frac{1}{2}(2r^2-2r+1)\right]^{n4}$$

$$L(r;n_i) = \lambda(1-r)^{2n1}\left[r(1-r)\right]^{n2}\left(r\right)^{2n3}\left[2r^2-2r+1\right]^{n4}$$

$$\ell(r,n_i) = \ln\lambda + \left[2n_1\ln(1-r)\right] + n_2\ln(1-r) + n_2\ln(r) + 2n_3\ln(r) + n_4\ln(2r^2-2r+1)$$

$$\ell(r,n_i) = \ln\lambda + (2n_1+n_{24})\ln(1-r) + (n_2+2n_3)\ln(r) + n_4\ln(2r^2-2r+1)$$

$$\frac{\partial\ell(r;n_i)}{\partial r} = \frac{-(2n_1+n_2)}{1-r} + \frac{(n_2+2n_3)}{r} + \frac{n_4(4r-2)}{2r^2-2r+1}$$

$$\frac{-(2n_1+n_2)}{1-r} + \frac{(n_2+2n_3)}{r} + \frac{n_4(4r-2)}{2r^2-2r+1} = 0$$

$$\left[-(2n_1+n_2)(2r^3-2r^2+r)\right] + \left[(n_2+2n_3)(-2r^3+4r^2-3r+1)\right] + \left[n_4(-4r^3+6r^2-2r)\right] = 0$$

$$r^3\left[-4n_1-4n_2-4n_3-4n_4)\right] + r^2\left[4n_1+6n_2+8n_3+6n_4)\right] + r\left[-2n_1-4n_2-6n_3-2n_4)\right] + (n_2+2n_3) = 0$$

$$r^3\left[-4N\right] + r^2\left[2(2N+n_2+2n_3+n_4)\right] + r\left[-2(N+n_2+2n_3)\right] + (n_2+2n_3) = 0$$

Substituindo os valores da Tabela 15B temos:

$$r^3\left[-800\right] + r^2\left[2(400+38+0+79)\right] + r\left[-2(200+38+0)\right] + (38+0) = 0$$

$$r^3\left[-800\right] + r^2\left[1034\right] + r\left[-476\right] + (38) = 0$$

Ao obter a raiz do polinômio tem-se que: r = 9,97%

Na Tabela 15C tem-se os valores de porcentagem de recombinação e LOD para as quatro diferentes fases de ligação possíveis entre os genitores. Note que o menor valor de r associado ao maior LOD está na situação em que os dois locos estão em aproximação, sendo esta a configuração mais provável. Além disso, pode se observar que em duas das possíveis fases de ligações não foi possível o cálculo da freqüência de recombinação.

Tabela 15C. Cruzamento Tipo: $A_1A_2B_1B_2$ x $A_1A_2B_1B_2$

Marcas: C_{14} e C_{15}	Tipo de Acasalamento: (1:2:1)(1:2:1)	
	Freqüência Conjunta	
Fase de Ligação	% de Recombinação	LOD escore
Ap x Ap	10,0	42,488
Ap x Re	-	-
Re x Ap	-	-
Re x Re	50,0	2,147
	Freqüência Marginal	
r_{G1} = 0 %	r_{G2} = 0 %	r_{med} = 0 %

Na Tabela 15D está a frequência de recombinação obtida entre pares de marcadores da população simulada que se enquadram no caso (1:2:1) (1:2:1).

Tabela 15D. Freqüência de recombinação de três pares de marcadores que possuem cruzamentos do tipo (1:2:1) (1:2:1).

Marcadores	Cruzamento	Freqüência de Recombinação
C_{14} X C_{15}	12 12 x 12 12	9,97
C_{15} X C_{310}	12 12 x 12 12	10,25

5) Caso (1:2:1): (1:1) Marcas C_{11} e C_{12}

Cruzamento Tipo: $A_1A_2B_1B_2$ x $A_1A_2B_2B_2$

A_1A_2 x A_1A_2 $\qquad$ B_1B_2 x B_2B_2

↓ $\qquad$ ↓

1 A_1 A_1 $\qquad$ 1 B_1 B_2

2 A_1 A_2 $\qquad$ 1 B_2 B_2

1 A_2 A_2

Neste caso, verifica-se a análise entre um loco informativo e outro parcialmente informativo. Como pode ser deduzido da Tabela 16A, não é possível determinar, com base na distribuição marginal entre os parentais, a estimativa da distância entre dois locos. Adicionalmente, o genitor 2 apresenta o loco C_{22} em homozigose, o que eventualmente bloqueia a contribuição do segundo parental para a análise de ligação.

Tabela 16A. Cruzamento Tipo: $A_1A_2B_1B_2$ x $A_1A_2B_2B_2$

Gametas		A_1B_2 P	A_1B_2 R	A_2B_2 R	A_2B_2 P	Freqüência Gamética	Total de indivíduos
A_1B_1	P	$P^2_{11\text{-}12}$ (45)	$PR_{11\text{-}12}$ (45)	$PR_{12\text{-}12}$ (44)	$P^2_{12\text{-}12}$ (44)	1/2	?
A_1B_2	R	$PR_{11\text{-}22}$ (10)	$R^2_{11\text{-}22}$ (10)	$R^2_{12\text{-}22}$ (44)	$PR_{12\text{-}22}$ (44)		
A_2B_1	R	$PR_{12\text{-}12}$ (44)	$R^2_{12\text{-}12}$ (44)	$R^2_{22\text{-}12}$ (1)	$PR_{22\text{-}12}$ (1)	1/2	?
A_2B_2	P	$P^2_{12\text{-}22}$ (44)	$PR_{12\text{-}22}$ (44)	$PR_{22\text{-}22}$ (56)	$P^2_{22\text{-}22}$ (56)		
Freq. Gamética		(1-r)/2	r/2	r/2	(1-r)/2		
Total		?	?	?	?		200

Na Tabela 16B estão explicitadas as classes genotípicas detectáveis para este tipo de acasalamento. Verifica-se que apenas seis classes genotípicas são distinguíveis e podem ser utilizadas na estimativa da porcentagem de recombinação com base na distribuição conjunta da progênie.

Tabela 16B. Cruzamento Tipo: $A_1A_2B_1B_2$ x $A_1A_2B_2B_2$

Classes	Observado	Esperado	Esperado (r=0,5)
11-12	45	PP+PR	25
11-22	10	RP+RR	25
12-12	44	PR+PP+RP+RR	50
12-22	44	RR+RP+PP+PR	50
22-12	1	RR+RP	25
22-22	56	PR+PP	25
Total	200	1	200

De posse das classes e do número de indivíduos observados em cada uma delas podemos obter uma expressão de Máxima Verossimilhança, de forma a obter uma fórmula geral para o cálculo da porcentagem de recombinação neste caso, sendo assim temos:

$$L(r;n_i) = \lambda\left(P^2 + PR\right)^{n1}(R^2 + PR)^{n2}\left(P^2 + R^2 + 2PR\right)^{n3}$$

$$L(r;n_i) = \lambda\left(\frac{1}{4}(1-r)\right)^{n1}\left[\frac{1}{4}r\right]^{n2}\left(\frac{1}{4}\right)^{n3}$$

$$L(r;n_i) = \lambda(1-r)^{n1}[r]^{n2}(1)^{n3}$$

$$\ell(r,n_i) = \ln\lambda + n_1\ln(1-r) + n_2\ln(r) + n_3\ln(1)$$

$$\frac{\partial\ell(r;n_i)}{\partial r} = \frac{-(n_1)}{1-r} + \frac{(n_2)}{r}$$

$$\frac{-(n_1)}{1-r} + \frac{(n_2)}{r} = 0$$

$$-r(n_1) + (1-r)(n_2) = 0$$

$$-r(n_1 + n_2) + (n_2) = 0$$

$$r = \frac{n_2}{n_1 + n_2}$$

De posse desta expressão geral, e com base nos dados relatados no exemplo acima temos:

$$r = \frac{11}{101+11} = 0{,}098 = 9{,}8\%$$

Na Tabela 16C são apresentadas as estimativas de distâncias entre os locos para as 4 possíveis fases de ligações dos genitores e os respectivos LOD escores. Como pode ser verificado, as configurações AP x AP e AP x Re apresentam as mesmas estimativas, permanecendo incerta a fase de ligação para o genitor 2. A estimativa obtida de r = 9,8 % aproxima-se da distância original entre as marcas com base no genoma simulado.

Tabela 16C. Cruzamento Tipo: $A_1A_2B_1B_2$ x $A_1A_2B_2B_2$

Marcas: C11 e C12	Tipo de Acasalamento: (1:2:1) (1:1)	
	Freqüência Conjunta	
Fase de Ligação	% de Recombinação	LOD escore
Ap x Ap	r = 9,8	18,095
Ap x Re	r = 9,8	18,095
Re x Ap	50	1,151
Re x Re	50	1,151
	Freqüência Marginal	
r_{G1} = indeterminado	r_{G2} = indeterminado	r_{med} = indeterminado

Na Tabela 16D está representada a freqüência de recombinação obtida entre pares de marcadores da população simulada que se enquadram no caso (1:2:1) (1: 1).

Tabela 16D. Freqüência de recombinação de dois pares de marcadores que possuem cruzamentos do tipo (1:2:1) (1:1).

Marcadores	Cruzamento	Freqüência de Recombinação
C_{11} X C_{12}	12 12 x 12 22	$r = \frac{11}{101+11} = 0{,}098 = 9{,}8\%$
C_{37} X C_{38}	12 12 x 12 22	$r = \frac{11}{97+11} = 0{,}1018 = 10{,}18\%$

6) Caso (1:1)(1:2:1): Marcas C_{13} e C_{14}

Cruzamento Tipo: $A_1A_2B_1B_2$ x $A_2A_2B_1B_2$

A_1A_2 x A_2A_2 → 1 A_1 A_2, 1 A_2 A_2

B_1B_2 x B_1B_2 → 1 B_1 B_1, 2 B_1 B_2, 1 B_2 B_2

Neste caso, verifica-se a análise entre um loco informativo e outro parcialmente informativo. Como pode ser deduzido da Tabela 17A, não é possível determinar a estimava da distância entre dois locos, com base na distribuição marginal entre os parentais. Adicionalmente, de forma similar ao apresentado no caso 2, o genitor 2 apresenta o loco C_{13} em homozigose, o que eventualmente bloqueia a contribuição do segundo parental para a análise de ligação.

Tabela 17A. Cruzamento Tipo: $A_1A_2B_1B_2$ x $A_2A_2B_1B_2$

Gametas		A_1B_1 P	A_1B_2 R	A_2B_1 R	A_2B_2 P	Freqüência Gamética	Total de indivíduos
A_2B_1	P	$P^2_{12\text{-}11}$ (49)	$PR_{12\text{-}12}$ (42)	$PR_{22\text{-}11}$ (1)	$P^2_{22\text{-}12}$ (57)	1/2	?
A_2B_1	R	$PR_{12\text{-}11}$ (49)	$R^2_{12\text{-}12}$ (42)	$R^2_{22\text{-}11}$ (1)	$PR_{22\text{-}12}$ (57)		
A_2B_2	R	$PR_{12\text{-}12}$ (42)	$R^2_{12\text{-}22}$ (1)	$R^2_{22\text{-}12}$ (57)	$PR_{22\text{-}22}$ (50)	1/2	?
A_2B_2	P	$P^2_{12\text{-}12}$ (42)	$PR_{12\text{-}22}$ (1)	$PR_{22\text{-}12}$ (57)	$P^2_{22\text{-}22}$ (50)		
Freq. Gamética		(1-r)/2	r/2	r/2	(1-r)/2		
Total		?	?	?	?		200

Na Tabela 17B estão explicitadas as classes genotípicas detectáveis para este tipo de acasalamento. Verifica-se que apenas seis classes genotípicas são distinguíveis e podem ser utilizadas na estimativa da porcentagem de recombinação com base na distribuição conjunta da progênie.

Tabela 17B. Cruzamento Tipo: $A_1A_2B_1B_2$ x $A_2A_2B_1B_2$

Classes	Observado	Esperado	Esperado (r=0,5)
12-11	49	P^2+PR	25
12-12	42	P^2+ R^2 + 2PR	50
12-22	1	R^2+PR	25
22-11	1	R^2+PR	25
22-12	57	P^2+ R^2 + 2PR	50
22-22	50	P^2+PR	25
Total	200	1	200

Para obtenção dos valores da porcentagem de recombinação conjunta utilizando a função de Máxima Verossimilhança fez uso da mesma expressão apresentada para o acasalamento (1:1) (1:2:1), sendo esta apresentada de forma resumida a seguir :

$$L(r;n_i) = \lambda\left(P^2 + PR\right)^{n1}(R^2 + PR)^{n2}\left(P^2 + R^2 + 2PR\right)^{n3}$$

$$r = \frac{n_2}{n_1 + n_2}$$

De posse desta expressão geral, e com base nos dados relatados no exemplo tem-se:

$$r = \frac{2}{99 + 2} = 0.0198 = 1.98\%$$

Na Tabela 17C são apresentadas as estimativas de distâncias entre os locos e os respectivos LOD escores. Como pode ser verificado, as configurações AP x AP e AP x Re apresentam as mesmas estimativas (r = 2 %) permanecendo incerta a fase de ligação para o genitor 2.

Tabela 17C. Cruzamento Tipo: $A_1A_2B_1B_2$ x $A_2A_2B_1B_2$

Marcas: C_{13} e C_{14}	Tipo de Acasalamento: (1:1)(1:2:1)	
	Freqüência Conjunta	
Fase de Ligação	% de Recombinação	LOD escore
Ap x Ap	2	26,138
Ap x Re	2	26,138
Re x Ap	50	1,245
Re x Re	50	1,245
	Freqüência Marginal	
r_{G1} = indeterminado	r_{G2} = indeterminado	r_{med} = indeterminado

Na Tabela 17D está representada a freqüência de recombinação obtida entre pares de marcadores da população simulada que se enquadram no caso (1: 1) (1:2:1).

Tabela 17D. Freqüência de recombinação de dois pares de marcadores que possuem cruzamentos do tipo (1: 1) (1:2:1).

Marcadores	**Cruzamento**	**Freqüência de Recombinação**
C_{13}X C_{14}	12 12 x 22 12	$r = \frac{2}{99 + 2} = 0{,}0198 = 1{,}98\%$
C_{111} X C_{21}	12 12 x 22 12	$r = \frac{9}{93 + 9} = 0{,}088 = 8{,}8\%$

7) Caso (1:1)(1:1): Marcas C_{12} e C_{13}

Cruzamento Tipo: $A_1A_2B_1B_2$ x $A_2A_2B_2B_2$

A_1A_2 x A_2A_2	B_1B_2 x B_2B_2
↓	↓
1 A_1 A_2	1 B_1 B_2
1 A_2 A_2	1 B_2 B_2

A análise de ligação para este caso envolve dois locos informativos. Verifica-se que a configuração assumida é idêntica àquela para Retrocruzamentos em linhagens endogâmicas (Tabela 18A). Como o genitor 2 apresenta-se em homozigose completa, apenas o parental 1 é utilizado para a determinação da distância entre os locos. Assim como o esperado para retrocruzamentos, apenas quatro classes podem ser distinguidas (Tabela 18B).

Tabela 18A. Cruzamento Tipo: $A_1A_2B_1B_2$ x $A_2A_2B_2B_2$

Gametas		A_2B_2 P	A_2B_2 R	A_2B_2 R	A_2B_2 P	Freqüência Gamética	Total de indivíduos
A_1B_1	P	$P^2_{12\text{-}12}$ (78)	$PR_{12\text{-}12}$ (78)	$PR_{12\text{-}12}$ (78)	$P^2_{12\text{-}12}$ (78)	(1-r)/2	78
A_1B_2	R	$PR_{12\text{-}22}$ (12)	$R^2_{12\text{-}22}$ (12)	$R^2_{12\text{-}22}$ (12)	$PR_{12\text{-}22}$ (12)	r/2	12
A_2B_1	R	$PR_{22\text{-}12}$ (14)	$R^2_{22\text{-}12}$ (14)	$R^2_{22\text{-}12}$ (14)	$PR_{22\text{-}12}$ (14)	r/2	14
A_2B_2	P	$P^2_{22\text{-}22}$ (96)	$PR_{22\text{-}22}$ (96)	$PR_{22\text{-}22}$ (96)	$P^2_{22\text{-}22}$ (96)	(1-r)/2	96
Freq. Gamética				1			
Total				200			200

Tabela 18B. Cruzamento Tipo: $A_1A_2B_1B_2$ x $A_2A_2B_2B_2$

Classes	Observado	Esperado	Esperado (r=0,5)
12-12	78	$2P^2$+2PR	50
12-22	12	$2R^2$+ 2PR	50
22-12	14	$2R^2$+ 2PR	50
22-22	96	$2P^2$+2PR	50
Total	200	1	200

De posse das classes e número de indivíduos observados em cada uma delas podemos obter uma expressão de Máxima Verossimilhança, de forma a obter uma fórmula geral para o cálculo da porcentagem de recombinação neste caso, sendo assim temos:

$$L(r;n_i) = \lambda\left(2P^2 + 2PR\right)^{n1}(2R^2 + 2PR)^{n2}$$

$$L(r;n_i) = \lambda\left(\frac{1}{2}(1-r)\right)^{n1}\left[\frac{1}{2}r\right]^{n2}$$

$$L(r;n_i) = \lambda(1-r)^{n1}[r]^{n2}$$

$$\ell(r,n_i) = \ln\lambda + n_1\ln(1-r) + n_2\ln(r)$$

$$\frac{\partial\ell(r;n_i)}{\partial r} = \frac{-(n_1)}{1-r} + \frac{(n_2)}{r}$$

$$\frac{-(n_1)}{1-r} + \frac{(n_2)}{r} = 0$$

$$-r(n_1) + (1-r)(n_2) = 0$$

$$-r(n_1 + n_2) + (n_2) = 0$$

$$r = \frac{n_2}{n_1 + n_2}$$

$$r = \frac{n_2}{N}$$

De posse desta expressão geral, e com base nos dados relatados no exemplo acima temos:

$$r = \frac{26}{200} = 0{,}13 = 13\%$$

Na Tabela 18C são apresentadas as estimativas de distâncias entre os locos C_{12} e C_{13}, com base na distribuição conjunta e marginal. Verifica-se que as estimativas se

igualam. A estimativa obtida foi de r = 13%. Nota-se que a fase de ligação do genitor1 é aproximação, porém a fase de ligação do genitor 2 continua indefinida.

Tabela 18C. Cruzamento Tipo: $A_1A_2B_1B_2$ x $A_2A_2B_2B_2$

Marcas: C_{12} e C_{13}	Tipo de Acasalamento: (1:1)(1:1)	
	Freqüência Conjunta	
Fase de Ligação	% de Recombinação	LOD escore
Ap x Ap	13	26,645
Ap x Re	13	26,645
Re x Ap	50	1,89
Re x Re	50	1,89
	Freqüência Marginal	
r_{G1} = 13%	r_{G2} = indeterminado	r_{med} = 13%

Na Tabela 18D está representada a freqüência de recombinação obtida entre pares de marcadores da população simulada que se enquadram no caso (1: 1) (1:1).

Tabela 18D. Freqüência de recombinação de três pares de marcadores que possuem cruzamentos do tipo (1:1) (1:1).

Marcadores	**Cruzamento**	**Freqüência de Recombinação**
C_{12} X C_{13}	12 12 x 22 22	$r = \frac{26}{200} = 0{,}13 = 13\%$
C_{210} X C_{211}	12 12 x 22 11	$r = \frac{22}{200} = 0{,}11 = 11\%$
C_{38} X C_{39}	12 12 x 22 22	$r = \frac{14}{200} = 0{,}07 = 7\%$

8) Caso (1:1)(1:1:1:1): Marcas C_{18} e C_{110}

Cruzamento Tipo: $A_1A_1B_1B_2$ x $A_1A_2B_2B_3$

A_1A_1 x A_1A_2	B_1B_2 x B_2B_3
↓	↓
1 A_1 A_1	1 B_1 B_2
1 A_1 A_2	1 B_1 B_3
	1 B_2 B_2
	1 B_2 B_3

Este caso envolve a análise entre um loco informativo e o outro completamente informativo. Nota-se, de acordo com as Tabelas 19A e 19B, que somente 8 classes genotípicas são distinguíveis, de forma que as freqüências genotípicas estão confundidas entre estas classes. Comparando-se com a análise para locos completamente informativos, verifica-se que pode haver uma considerável redução no conteúdo de informação, dada o confundimento de classes.

Adicionalmente, apenas para o genitor 2 (12-23) é possível determinar, com base na distribuição marginal, os gametas parentais e recombinantes e assim estimar a porcentagem de recombinação. Para o genitor 1 (11-12), apenas o loco C_{110} é heterozigoto, de forma que os gametas segregam na proporção de 1/2:1/2 (Tabela 19A), não sendo possível estimar com base na distribuição marginal do genitor 1 à distância entre os locos.

Tabela 19A. Cruzamento Tipo: $A_1A_1B_1B_2$ x $A_1A_2B_2B_3$

Gametas		A_1B_1 P	A_1B_1 R	A_1B_2 R	A_1B_2 P	Freq. Gamética	Total
A_1B_2	P	$P^2_{11\text{-}12}$ (37)*	$PR_{11\text{-}12}$ (37)	$PR_{11\text{-}22}$ (46)	$P^2_{11\text{-}22}$ (46)	(1-r)/2	83
A_1B_3	R	$PR_{11\text{-}13}$ (12)	$R^2_{11\text{-}13}$ (12)	$R^2_{11\text{-}23}$ (9)	$PR_{11\text{-}23}$ (9)	r/2	21
A_2B_2	R	$PR_{12\text{-}12}$ (8)	$R^2_{12\text{-}12}$ (8)	$R^2_{12\text{-}22}$ (8)	$PR_{12\text{-}22}$ (8)	r/2	16
A_2B_3	P	$P^2_{12\text{-}13}$ (36)	$PR_{12\text{-}13}$ (36)	$PR_{12\text{-}23}$ (44)	$P^2_{12\text{-}23}$ (44)	(1-r)/2	80
Freq. Gamética		1/2		1/2			
Total		93		107			200

* Classes genotípicas entre parênteses indicam classes não distinguíveis.

Tabela 19B. Cruzamento Tipo: $A_1A_1B_1B_2$ x $A_1A_2B_2B_3$

Classes Genotípicas	Observado	Esperado	Esperado (r=0,5)
11 12	37	P^2+PR	25
11 13	12	R^2+PR	25
11 22	46	P^2+PR	25
11 23	9	R^2+PR	25
12 12	8	R^2+PR	25
12 13	36	P^2+PR	25
12 22	8	R^2+PR	25
12 23	44	P^2+PR	25
Total	200	1	200

A função de Máxima Verossimilhança utilizada com a finalidade de obter a freqüência de recombinação conjunta é apresentada a seguir:

$$L(r;n_i) = \lambda\left(P^2 + PR\right)^{n1}(R^2 + PR)^{n2}$$

$$L(r;n_i) = \lambda\left(\frac{1}{4}(1-r)\right)^{n1}\left[\frac{1}{4}r\right]^{n2}$$

$$L(r;n_i) = \lambda(1-r)^{n1}[r]^{n2}$$

$$\ell(r,n_i) = \ln\lambda + n_1\ln(1-r) + n_2\ln(r)$$

$$\frac{\partial\ell(r;n_i)}{\partial r} = \frac{-(n_1)}{1-r} + \frac{(n_2)}{r}$$

$$\frac{-(n_1)}{1-r} + \frac{(n_2)}{r} = 0$$

$$-r(n_1) + (1-r)(n_2) = 0$$

$$-r(n_1+n_2) + (n_2) = 0$$

$$r = \frac{n_2}{n_1+n_2}$$

$$r = \frac{n_2}{N}$$

De posse desta expressão geral, e com base nos dados relatados no exemplo acima temos:

$$r = \frac{37}{200} = 0{,}185 = 18{,}5\%$$

Verifica-se na Tabela 19C que a estimativa da distância com base na distribuição conjunta iguala-se à estimativa com base no genitor 2. Entretanto, a fase de ligação permanece incerta – entre Ap x Ap e Re x Ap – pois, como comentado, não é possível distinguir no genitor 1 os gametas parentais e recombinantes.

Tabela 19C. Cruzamento Tipo: $A_1A_1B_1B_2$ x $A_1A_2B_2B_3$

Marcas: C_{18} e C_{110}	Tipo de Acasalamento: (1:1)(1:1:1:1)	
	Freqüência Conjunta	
Fase de Ligação	% de Recombinação	LOD escore
Ap x Ap	18,5	18,61
Ap x Re	50	1,603
Re x Ap	18,5	18,61
Re x Re	50	1,603
	Freqüência Marginal	
r_{G1} = indeterminado %	r_{G2} = 18,5%	r_{med} = indeterminado

Na Tabela 19D está representada a freqüência de recombinação obtida entre pares de marcadores da população simulada que se enquadram no caso (1: 1) (1:1:1:1).

Tabela 19D. Freqüência de recombinação de três pares de marcadores que possuem cruzamentos do tipo (1:1) (1:1:1:1).

Marcadores	**Cruzamento**	**Freqüência de Recombinação**
C_{18} X C_{110}	11 12 x 12 23	$r = \frac{37}{200} = 0{,}185 = 18{,}5\%$
C_{24} X C_{311}	12 12 x 22 23	$r = \frac{18}{200} = 0{,}09 = 9\%$
C_{27} X C_{28}	11 12 x 23 13	$r = \frac{19}{200} = 0{,}095 = 9{,}5\%$
C_{211} X C_{31}	12 12 x 11 13	$r = \frac{17}{200} = 0{,}085 = 8{,}5\%$

9) Caso (1:1:1:1)(1:1): Marcas C_{17} e C_{18}

Cruzamento Tipo: $A_1A_2B_1B_1$ x $A_2A_3B_1B_2$

A_1A_2 x A_2A_3	B_1B_1 x B_1B_2
↓	↓
1 A_1 A_2	1 B_1 B_1
1 A_1 A_3	1 B_1 B_2
1 A_2 A_2	
1 A_2 A_3	

Este caso envolve a análise entre um loco informativo e o outro completamente informativo. Nota-se, de acordo com as Tabelas 20A e 20B, que somente 8 classes genotípicas são distinguíveis, de forma que as freqüências genotípicas estão confundidas entre estas classes. Comparando-se com a análise para locos completamente informativos, verifica-se que pode haver uma considerável redução no conteúdo de informação, dada o confundimento de classes.

Adicionalmente, apenas para o genitor 2 (23-12) é possível determinar, com base na distribuição marginal, os gametas parentais e recombinantes e assim estimar a porcentagem de recombinação. Para o genitor 1 (12-11), apenas o loco C_{17} é heterozigoto, de forma que os gametas segregam na proporção de 1/2:1/2 (Tabela 20A), não sendo possível estimar com base na distribuição marginal do genitor 1 a distância entre os locos.

Tabela 20A. Cruzamento Tipo: $A_1A_2B_1B_1$ x $A_2A_3B_1B_2$

Gametas		A_1B_1 P	A_1B_1 R	A_2B_1 R	A_2B_1 P	Freq. Gamética	Total
A_2B_1	P	$P^2_{12\text{-}11}$ (46)*	$PR_{12\text{-}11}$ (46)	$PR_{22\text{-}11}$ (49)	$P^2_{22\text{-}11}$ (49)	(1-r)/2	95
A_2B_2	R	$PR_{12\text{-}12}$ (3)	$R^2_{12\text{-}12}$ (3)	$R^2_{22\text{-}12}$ (4)	$PR_{22\text{-}12}$ (4)	r/2	7
A_3B_1	R	$PR_{13\text{-}11}$ (6)	$R^2_{13\text{-}11}$ (6)	$R^2_{23\text{-}11}$ (3)	$PR_{23\text{-}11}$ (3)	r/2	9
A_3B_2	P	$P^2_{13\text{-}12}$ (42)	$PR_{13\text{-}12}$ (42)	$PR_{23\text{-}12}$ (47)	$P^2_{23\text{-}12}$ (47)	(1-r)/2	89
Freq. Gamética		1/2		1/2			
Total		97		103			200

* Classes genotípicas entre parênteses indicam classes não distinguíveis.

Tabela 20B. Cruzamento Tipo: $A_1A_2B_1B_1$ x $A_2A_3B_1B_2$

Classes Genotípicas	Observado	Esperado	Esperado (r=0,5)
12 11	46	P^2+PR	25
12 12	3	R^2+PR	25
13 11	6	R^2+PR	25
13 12	42	P^2+PR	25
22 11	49	P^2+PR	25
22 12	4	R^2+PR	25
23 11	3	R^2+PR	25
23 12	47	P^2+PR	25
Total	200	1	200

De posse das classes esperadas pode-se montar a função de verossimilhança. Neste caso a expressão pra estimar a freqüência de recombinação é idêntica àquela apresentada no caso 8, envolvendo o cruzamento (1:1:1:1) (1:1). Sendo assim, aqui esta expressão foi ilustrada de forma resumida, dada a seguir:

$$L(r;n_i) = \lambda\left(P^2 + PR\right)^{n1}(R^2 + PR)^{n2}$$

$$r = \frac{n_2}{N}$$

De posse desta expressão geral, e com base nos dados relatados no exemplo acima para os marcadores C_{17} e C_{18} temos:

$$r = \frac{16}{200} = 0{,}08 = 8\%$$

Verifica-se na Tabela 20C que a estimativa da distância com base na distribuição conjunta iguala-se à estimativa com base no genitor 2. Entretanto, a fase de ligação permanece incerta – entre Ap x Ap e Re x Ap– pois, como comentado, não é possível distinguir no genitor 1 os gametas parentais e recombinantes.

Tabela 20C. Cruzamento Tipo: $A_1A_2B_1B_1$ x $A_2A_3B_1B_2$

Marcas: C_{17} e C_{18}	Tipo de Acasalamento: (1:1:1:1) (1:1)	
	Freqüência Conjunta	
Fase de Ligação	% de Recombinação	LOD escore
Ap x Ap	8	35,993
Ap x Re	50	2,151
Re x Ap	8	35,993
Re x Re	50	2,151
	Freqüência Marginal	
r_{G1} = indeterminado %	r_{G2} = 8%	r_{med} = indeterminado

Na Tabela 20D está representada a freqüência de recombinação obtida entre pares de marcadores da população simulada que se enquadram no caso (1:1:1:1) (1: 1).

Tabela 20D. Freqüência de recombinação de seis pares de marcadores que possuem cruzamentos do tipo (1:1:1:1) (1: 1).

Marcadores	**Cruzamento**	**Freqüência de Recombinação**
C_{17} X C_{18}	12 11 x 23 12	$r = \frac{16}{200} = 0{,}08 = 8\%$
C_{110}X C_{19}	12 12 x 23 22	$r = \frac{22}{200} = 0{,}11 = 11\%$
C_{23}X C_{24}	12 12 x 23 22	$r = \frac{22}{200} = 0{,}11 = 11\%$
C_{25} X C_{26}	12 11 x 13 23	$r = \frac{25}{200} = 0{,}125 = 12{,}5\%$

10) Caso (1:1)(1:1): EXTRA

Neste caso 10 o interesse é realçar a possibilidade de ocorrência de um outro tipo de curzamento, que não foi contemplado no caso da população simulada. Nesta situação as quatro classes esperadas possuem a mesma probabilidade, sendo do tipo P^2+2PR + R^2. Desta forma não é possível o cálculo da freqüência de recombinação.

Então deve sempre ter o cuidado de observar as classes esperadas oriundas de cada cruzamento antes de aplicar as expressões obtidas nos casos anteriores, de forma que a aplicação da expressão deve ser feita sempre respeitando o tipo de cruzamento envolvido e as classes de probabilidades esperadas.

De forma ilustrativa representa-se um exemplo hipotético, mostrado a seguir:

Cruzamento Tipo: $A_1A_2B_1B_1$ x $A_2A_2B_2B_3$

Este caso apresenta a análise de ligação entre dois locos informativos. Entretanto, a configuração dos parentais P_1 (12-22) e P_2 (22-23) impede completamente a determinação da distância entre os dois locos, seja com base na distribuição marginal ou conjunta. Isso se dá pela presença de um loco em homozigose ora em um parental ora em outro. Neste caso torna-se impossível detectar os gametas recombinantes e parentais para ambos os genitores, uma vez que todas as classes genotípicas observadas na progênie são confundidas, não podendo alocar de forma correta os genótipos.

De acordo com a Tabela 21B, apenas quatro classes podem ser detectadas neste cruzamento, não fornecendo, entretanto, conteúdo algum de informação. Isso pode ser confirmado analisando os resultados apresentados na Tabela 21C. Verifica-se que os valores de LOD escore são iguais a zero, para qualquer configuração de fase de ligação e o valor calculado da estimativa da distância é um valor inconsistente.

Tabela 21A. Cruzamento Tipo: $A_1A_2B_1B_1$ x $A_2A_2B_2B_3$

Gametas		A_2B_2 P	A_2B_3 R	A_2B_2 R	A_2B_3 P	Freqüência Gamética	Total de indivíduos
A_1B_1	P	$P^2_{12\text{-}12}$ (51)	$PR_{12\text{-}13}$ (50)	$PR_{12\text{-}12}$ (51)	$P^2_{12\text{-}13}$ (50)	1/2	101
A_1B_1	R	$PR_{12\text{-}12}$ (51)	$R^2_{12\text{-}13}$ (50)	$R^2_{12\text{-}12}$ (51)	$PR_{12\text{-}13}$ (50)		
A_2B_1	R	$PR_{22\text{-}12}$ (48)	$R^2_{22\text{-}13}$ (51)	$R^2_{22\text{-}12}$ (48)	$PR_{22\text{-}13}$ (51)	1/2	99
A_2B_1	P	$P^2_{22\text{-}12}$ (48)	$PR_{22\text{-}13}$ (51)	$PR_{22\text{-}12}$ (48)	$P^2_{22\text{-}13}$ (51)		
Freq. Gamética		(1-r)/2	r/2	r/2	(1-r)/2		
Total		(99)	(101)	(99)	(101)		200

Tabela 21B. Cruzamento Tipo: $A_1A_2B_1B_1$ x $A_2A_2B_2B_3$

Classes	Observado	Esperado	Esperado (r=0,5)
12-12	51	P^2+PR+PR+ R^2	50
12-13	50	PR+ P^2+ R^2+ PR	50
22-12	48	PR+ P^2+ R^2+ PR	50
22-13	51	P^2+PR+PR+ R^2	50
Total	200	1	200

Tabela 21C. Cruzamento Tipo: $A_1A_2B_1B_1$ x $A_2A_2B_2B_3$

	Tipo de Acasalamento: (1:1)(1:1)	
	Freqüência Conjunta	
Fase de Ligação	% de Recombinação	LOD escore
Ap x Ap	-	0
Ap x Re	-	0
Re x Ap	-	0
Re x Re	-	0
	Freqüência Marginal	
r_{G1} = indeterminado	r_{G2} = indeterminado	r_{med} = indeterminado

Após todas estas análises pode-se configurar um quadro resumindo todas as informações dos possíveis cruzamentos diferentes entre marcadores. Este resumo está na Tabela 22 a seguir.

Tabela 22. Esquema resumido com todas as segregações analisadas, suas classes de probabilidades observadas, função de verossimilhança e expressão geral para obtenção do valor da freqüência de recombinação.

Cruzamento/ Segregação	Classes	Função de Verossimilhança	Expressão
1) (1:1:1:1) (1:1:1:1)	P^2; PR; R^2	$L(r;n_i) = \lambda\left(P^2\right)^{n1}(PR)^{n2}\left(R^2\right)^{n3}$	$r = \frac{2n_3 + n_2}{2N}$
2) (1:1:1:1) (1:2:1)	P^2; PR; R^2; 2PR; P^2+R^2	$L(r;n_i) = \lambda\left(P^2\right)^{n1}(PR)^{n2}\left(R^2\right)^{n3}(PR+PR)^{n4}(P^2+R^2)^{n5}$	$r^3[-4N] + r^2[2(2N + n_2 + 2n_3 + n_4 + n_5)] + r[-2(N + n_2 + 2n_3 + n_4)] + (n_2 + 2n_3 + n_4)$
3) (1:2:1)(1:1:1:1)	P^2; PR; R^2; 2PR; P^2+R^2	$L(r;n_i) = \lambda\left(P^2\right)^{n1}(PR)^{n2}\left(R^2\right)^{n3}(PR+PR)^{n4}(P^2+R^2)^{n5}$	$r^3[-4N] + r^2[2(2N + n_2 + 2n_3 + n_4 + n_5)] + r[-2(N + n_2 + 2n_3 + n_4)] + (n_2 + 2n_3 + n_4)$
4) (1:2:1)(1:2:1)	P^2; R^2; 2PR; $2P^2$+$2R^2$	$L(r;n_i) = \lambda\left(P^2\right)^{n1}(2PR)^{n2}\left(R^2\right)^{n3}(2P^2+2R^2)^{n4}$	$r^3[-4N] + r^2[2(2N + n_2 + 2n_3 + n_4)] + r[-2(N + n_2 + 2n_3)] + (n_2 + 2n_3)$
5) (1:2:1)(1:1)	P^2 + PR; R^2 + PR; P^2 + R^2 + 2PR	$L(r;n_i) = \lambda\left(P^2+PR\right)^{n1}(R^2+PR)^{n2}\left(P^2+R^2+2PR\right)^{n3}$	$r = \frac{n_2}{n_1 + n_2}$
6) (1:1) (1:2:1)	P^2 + PR; R^2 + PR; P^2 + R^2 + 2PR	$L(r;n_i) = \lambda\left(P^2+PR\right)^{n1}(R^2+PR)^{n2}\left(P^2+R^2+2PR\right)^{n3}$	$r = \frac{n_2}{n_1 + n_2}$
7) (1:1) (1:1)	$2P^2$ + 2PR; $2R^2$ + 2PR	$L(r;n_i) = \lambda\left(2P^2+2PR\right)^{n1}(2R^2+2PR)^{n2}$	$r = \frac{n_2}{N}$
8) (1:1) (1:1:1:1)	P^2 + PR; R^2 + PR	$L(r;n_i) = \lambda\left(P^2+PR\right)^{n1}(R^2+PR)^{n2}$	$r = \frac{n_2}{N}$
9) (1:1:1:1) (1:1)	P^2 + PR; R^2 + PR	$L(r;n_i) = \lambda\left(P^2+PR\right)^{n1}(R^2+PR)^{n2}$	$r = \frac{n_2}{N}$
10) (1:1)(1:1) (extra)	P^2+PR+PR+ R^2	-	-

5. Considerações Gerais

No desenvolvimento de mapas de ligação, usualmente, inicia-se com um grupo aleatório de marcadores para os quais a posição no mapa não está disponível. É necessário, portanto, separar os marcadores em grupos de ligação e depois ordená-los. Um par de marcadores é considerado ligado quando as freqüências dos marcadores obtidas na progênie são significativamente diferentes para as freqüências esperadas na ausência de ligação (r=0,50). O teste de LOD escore é o mais utilizado, principalmente nas situações em que segregação distorcida sistemática não tem sido observada. Outros métodos podem ser utilizados para testar a ligação entre marcas, com o teste de qui-quadrado conjunto (χ^2_L) ou o teste de contingência (Maliepaard et. al., 1997).

Um importante aspecto a destacar é a acurácia das estimativas obtidas para a distância entre os marcadores. Embora o LOD escore significativo indique a ligação entre um par de marcas, isto não implica que a estimativa da freqüência de recombinação seja acurada. Dessa forma, no processo de mapeamento, o interesse não está apenas em detectar a ligação entre pares de marcas, mas também obter estimativas acuradas necessárias para determinar a ordem e as distâncias dos marcadores.

A acurácia das estimativas de Máxima Verossimilhança para as distâncias entre pares de marcadores pode ser obtida por meio de intervalos de confiança. Assim, é necessário determinar a variância das estimativas. Segundo Malieppard et. al. (1997), a variância dos estimadores de freqüência de recombinação compreende-se de dois componentes: (a) o número de eventos de recombinação criados pela amostragem dos gametas a partir da geração parental; (b) a habilidade ou impossibilidade de que estes eventos possam ser detectados para certas combinações de configurações para dois locos. O primeiro componente é determinado pela própria freqüência de recombinação, e o tamanho amostral na progênie. O segundo componente é determinado pelo tipo de segregação dos locos envolvidos (de acordo com o conteúdo de informação, como já mencionado) e a fase de ligação entre os parentais. Conforme tem sido abordado na literatura, a variância de estimadores de Máxima Verossimilhança pode ser aproximadamente determinada pelo inverso do índice de informação de Fisher. Assim, funções de informação podem ser estabelecidas por meio do inverso da derivada segunda da função de verossimilhança. Tais

funções de informação devem ser estabelecidas para diferentes configurações de acasalamento.

Neste estudo, foram abordadas diferentes configurações para pares de marcas na análise de ligação em Famílias de Irmãos Completos. Verificou-se, de modo empírico, que as diferentes configurações utilizadas variam significativamente em relação ao poder de detecção da ligação e na (im)possibilidade de se estimar a fase de ligação para ambos os parentais. Dessa forma, a acurácia do mapa genético depende da quantidade de informação contida no grupo de dados utilizado para a construção do mapa. Obviamente, a quantidade de informação depende do delineamento experimental utilizado e da configuração dos locos em análise.

Como observado para o cenário 2, diferentes conteúdos de informação são obtidos para as estimativas de freqüência de recombinação conforme a configuração dos locos em análise. Assim, após os marcadores terem sido alocados nos grupos de ligação, informações conflitantes sobre a ordem dos marcadores nos grupos de ligação podem surgir devido às diferentes estimativas das freqüências de recombinação inerentes à configuração dos marcadores nos parentais. Para situações onde não é possível estimar a ligação entre as marcas, uma alternativa de análise é alocar os marcadores não informativos baseando-se nos marcadores polimórficos informativos em sua vizinhança. Esta técnica não foi aqui retratada.

Para o presente estudo foi abordada a possibilidade de se estimar as freqüências de recombinação com base na distribuição marginal dos gametas parentais para cada genitor, de modo a se obter um mapa de ligação para cada parental, à semelhança de um esquema de pseudotestcross (Grattapaglia e Sederoff, 1994). Alternativamente, pode-se utilizar a distribuição conjunta observada na progênie para a construção de um mapa de ligação único. Verificou-se que as estimativas de porcentagem de recombinação podem variar entre as diferentes abordagens, conforme o conteúdo de informação dos locos analisados. Quando se dispõe de locos completamente informativos ambas as alternativas se igualam, isto é, a estimativa feita com base na média das freqüências marginais se iguala com a estimativa baseada na freqüência conjunta, como pode ser visto na análise do cenário 1. Entretanto, quando são utilizadas diferentes combinações de locos quanto ao seu conteúdo de informação, as estimativas nem sempre são iguais. Assim, a construção de mapas integrados pode variar em relação aos mapas obtidos para cada parental.

Quando se deseja construir mapas integrados, a disponibilidade de diferentes classes de marcadores deve ser enfatizada. Para classes de marcadores dominantes, como o caso de marcadores RAPD e AFLP, a integração de mapas pode se tornar mais complicada. Assim, recomenda-se utilizar classes de marcadores multialélicos, como o caso de marcadores SSR e RFLP, para a construção de mapas integrados. Uma vantagem extra da utilização de marcadores multialélicos é a possibilidade de aplicação destes na análise de diferentes configurações de informação, como demonstrado no presente estudo. Assim, se um número suficiente de marcadores completamente informativos (do tipo 12 x 34) estiver disponíveis, ambas as opções podem ser utilizadas: tanto o uso separado de mapas para cada parental, ou, se as diferenças entre as estimativas de recombinação não são significativas, a construção de um mapa integrado (Maliepaard et al., 1997).

6. Conclusões

1- A porcentagem de recombinação entre pares de marcas, obtidas de cruzamentos completamente informativos (A_iA_j x A_kA_l), obtidas com base em todas as classes genotípicas é a média aritmética da porcentagem de recombinação obtida com valores marginais associados às freqüências gaméticas dos genitores.

2- A porcentagem de recombinação entre pares de marcadores, em cruzamentos não completamente informativos é mais eficaz quando calculada a partir de todas as informações genotípicas. O uso das informações marginais não é apropriado pela grande quantidade de perda de informação.

3- A metodologia descrita é eficaz em gerar um mapa integrado dos genitores avaliados. O uso das informações marginais apresenta o inconveniente de requerer, adicionalmente, aplicativos para integração de mapas.

7. Referências Bibliográficas

ALBERTS, B., JOHNSON, A., LEWIS, J., RAFF, M., ROBERTS, K., WALTER, P. Molecular biology of the cell. 4th ed. **Garland Science**, 1.463p. 2002.

ALLARD, R. W. Formulas and tables to facilitate the calculation of recombination values in heredity. Hilgardia, 24: 235-278.1956.

ARUS, P., OLART, C., ROMERO, M. e VARGAS, F. Linkage analysis of tem isozyme genes in F1 segregating almond progenies. **Journal of the American Society for Horticultural Science**, 119: 339-344. 1994.

CRUZ, C.D. **Programa para análise de dados moleculares e quantitativos** – GQMOL. Viçosa: UFV, 2005.

DOLIGEZ, A., ADAM-BLONDON, A.F., CIPRIANI, G., DI GASPERO, G. LAUCOU, V., MERDINOGLU, D., MEREDITH, C.P., RIAZ, S., ROUX,C., THIS, P. An integrated SSR map of grapevine based on five mapping populations. **Theor Appl Genet,** 113:369–382. 2006.

GRATTAPAGLIA, D. e SEDEROFF R.R. Genetic linkage maps of Eucalyptus grandis and E. urophylla using a pseudo-testcross mapping strategy and RAPD markers, **Genetics** 137: 1121-1137.1994

GEORGE, A.W., MENGERSEN ,K.L. e DAVIS, G.P., A Bayesian approach to ordering gene markers. **Biometrics** 55: 419–429. 1999.

GUILHERME, J.M.R., YANDELL, B.S. e GIANOLA D., A Bayesian approach for constructing genetic maps when markers are miscoded. **Genet. Sel. Evol**. 34: 353–369. 2002.

HALDANE, J.B.S. The combination of linkage values and the calculation of distances between the loci of linked factors. **J. Genet**. 8: 299-309. 1919.

HASEMAN, J.K., ELSTON, R.C., The investigation of linkage between a quantitative trait and a marker locus. **Behav. Genet**., 2: 3-19. 1972.

HE, P., Li, J.Z., ZHENG, X.W., SHEN, L.S., LU, C.F., CHEN, Y., ZHU, L.H. Comparison of molecular linkage maps and agronomic trait loci between DH and RIL populations derived from the same rice cross. **Crop Science**, 41: 1240– 1246, 2001.

JANSEN, J. Construction of linkage maps in full-sib families of diploid outbreeding species by minimizing the number of recombinations in hidden inheritance vectors. **Genetics**, 170: 2013-2025. 2005.

KOSAMBI, D.D. The estimation of map distances from recombination values. **Ann Eugen**. 12: 172-175. 1944.

LANDER, E. S., e GREEN P., Construction of multilocus genetic maps in humans. **Proc. Natl. Acad. Sci**. USA 84: 2363–2367. 1987.

LIU, B.H. Statistical genomics: linkage, mapping, and QTL analysis. Boca Raton, Flórida, USA: **CRC Press**. 568p. 1998.

LU, Q., CUI, Y., WU, R., A multilocus likelihood approach to joint modeling of linkage, parental diplotype and gene order in a full sib family. **BMC genetics**, 5: 20, 2004.

LYNCH, M., WALSH, B., **Genetics and analysis of quantitative traits**. 1th ed. Sunderland, MA: Sinuauer Associets, Inc. 980p. 1998.

MA, C.-X., LIN, M., LITTELL, R.C., YIN, T., WU, R., A likelihood approach for mapping growth trajectories using dominats markers in a phase unknown full-sib family. **Theor. Applied Genet**. 108: 699-705. 2004.

MALIEPAARD, C.; JANSEN, J.; VAN OOIJEN, J.W. Linkage análisis in a full-sib family of an outbreeding plant species: overview and consequences. **Genet. Res. Camb**. 70; 237-250. 1997.

MATTER, k.. **The Measurement of Linkage in Heredity**. London :Methuen. 1951.

NOVAES, E. Detecção de QTls para qualidade de madeira em Eucalyptus grandis x Eucalyptus urophylla e ancoragem de clones BAC no mapa genético. **Dissertação de mestrado** (Apresentada a Universidade Federal de Viçosa (UFV) no programa de Genética e Melhoramento). 171p. 2006.

RITTER, E. e SALAMINI, F. The calculation of recombination frequencies in crosses of allogamous plant species with applications to linkage mapping. **Genetical Research** 67, 55-65.1996.

RITTER, E., GEBHARDT, C. e SALAMINI, F. Estimation os recombination frequencies and construction of RFLP linkage maps in plants from crosses between heterozygous parents. **Genetics** 125: 645-654.1990.

SCHUSTER, I., CRUZ, C.D. **Estatística genômica aplicada a populações derivadas de cruzamentos controlados**. 1. ed. Viçosa, MG: Imprensa Universitária. 568p. 2004.

TARTAGLIA., Método para obtenção de raízes de equação de terceiro grau. Obtido em: http://pessoal.sercomtel.com.br/matematica/medio/polinom/tartaglia.htm, acessado em 10/02/2008).

UKRAINETZ, N.K., RITLAND, K., MANSFIELD, S.D. An AFLP linkage map for Douglas-fir based upon multiple full-sib families. **Tree Genetics & Genomes** 4: 181–191. 2008.

VERHAEGEN, D.; PLOMION, C. Genetic mapping in Eucalyptus urophylla and Eucalyptus grandis using RAPD markers. **Genome**, 39: 1051-1061, 1996.

WEBER, W.E. e WRICKE, G. Genetic Markers in Plant Breeding. Advances in plant breeding 16. Berlin: **Parey Scientific**. 1994.

WU, R. e MA, C. Simultaneous Maximum Likelihood Estimation of Linkage and Linkage Phases in outcrossing species. **Theoretical Population Biology**, 61: 349-363, 2002.

ZAMIR, D., TADMOR, Y., Unequal segregation of nuclear genes in plants. **Botanical Gazette**, 147: 355-358, 1986.

CAPÍTULO 2

TAMANHO DE POPULAÇÃO IDEAL PARA MAPEAMENTO GENÉTICO EM FAMÍLIAS DE IRMÃOS COMPLETOS

VIÇOSA
MINAS GERAIS – BRASIL
2008

RESUMO

BHERING, Leonardo Lopes, D.Sc., Universidade Federal de Viçosa, fevereiro de 2008. **Mapeamento genético em famílias simuladas de irmãos completos: Tamanho de poulação ideal para mapeamento.** Orientador: Cosme Damião Cruz. Co-orientadores: José Marcelo Soriano Viana e Pedro Crescêncio Souza Carneiro.

O mapeamento genético facilita o trabalho de melhoramento uma vez que uma ou mais marcas do genótipo podem ser associadas a genes controladores de características qualitativas e quantitativas (QTL). Um dos fatores de fundamental importância para se obter dados consistentes em um trabalho de mapeamento é o tamanho da amostra ou população a ser trabalhada. Deste modo, o objetivo deste trabalho foi gerar e analisar dados a partir da simulação de genoma e de populações, e com base nestes dados simulados avaliar o tamanho ótimo de populações para estudo de mapeamento genético de famílias de irmãos completos. Foram simulados genomas parentais e amostras de populações de família de irmãos completos do tipo completamente informativa, e também que não fossem completamente informativas. As amostras geradas foram de tamanho 100, 200, 400 e 600 indivíduos com três grupos de ligação cada e 11 marcas moleculares codominantes e multialélicas espaçadas a 10 centimorgans por grupo de ligação. Foram realizadas 100 repetições por amostra. Concluiu-se que para populações completamente informativas um tamanho populacional de 200 indivíduos seria o suficiente para resgatar as informações originais. Contudo, para a população não completamente informativa seria necessária a utilização de uma população maior, constituída de 600 indivíduos.

Termos de indexação**:** populações exogâmicas, estatística genômica, tamanho amostral.

ABSTRACT

BHERING, Leonardo Lopes, D.Sc., Universidade Federal de Viçosa. February, 2008. **Genetic mapping in simulated full siblings Family (2): Optimum population size for genetic mapping in full siblings' families.**Adviser: Cosme Damião Cruz. Co-Advisers: José Marcelo Soriano Viana and Pedro Crescêncio Souza Carneiro.

The genetic mapping facilitates the breeding work once one or more marks of the genotype can be associated to controlling genes of qualitative and quantitative traits (QTL). One important factor to obtain solid data in a mapping work is the size of the sample or population to be worked. This way, the objective of this work was to generate and to analyze data starting from simulation of the genome and of populations, and based in these simulated data to evaluate the optimum size of populations for study of genetic mapping of full siblings' families. There were simulated parental genomes and samples of populations of full siblings' families of both types, completely informative and not completely informative. The generated samples had 100, 200, 400 and 600 individuals with three linkage groups each and 11 marks molecular codominantes and multi-allelic spaced 10 centimorgans in each linkage group. 100 repetitions were accomplished by sample. In completely informative populations, an optimum size of 200 individuals would be enough to rescue the original information, however, for the population not completely informative it would be necessary a larger population, constituted of 600 individuals.

Indexation terms: exogamic populations, statistical genomic, size sample

1. Introdução

O mapeamento genético facilita o trabalho de melhoramento uma vez que marcas no genoma podem ser associadas a um ou mais genes controladores de características qualitativas e quantitativas (QTL). Desse modo, com o genoma da espécie mapeado, o trabalho de melhoramento poderá ser otimizado, tanto na eficiência do programa, quanto na velocidade de obtenção de ganhos. Entretanto a disponibilidade de um mapa genético fidedigno depende de uma série de fatores, como, o tipo de marcador utilizado, o tipo de população analisada e o tamanho da população. Além disto, alguns aspectos de natureza metodológica também devem ser considerados no processo de obtenção do mapa genético tais como análise de segregação de locos individuais, análise de segregação conjunta, níveis máximos de recombinação e mínimos de LOD escores para estabelecimento de ligação, rotinas de ordenamento de locos, dentre outras. Com isso, nota-se que uma das ferramentas disponíveis, e que vem se tornando cada vez mais necessária, ao melhorista é a Genômica.

Genômica é a denominação dada a ciência que estuda o genoma de forma completa, pela integração de várias áreas tradicionais da genética tais como a Genética Mendeliana, a Citogenética, a Genética Molecular, a Genética de Populações e a Genética Quantitativa, incluindo também a ciência da computação e os sistemas automatizados (Liu, 1998).

Um dos fatores de fundamental importância para se obter dados consistentes em um mapeamento é o tamanho da amostra ou população a ser trabalhada. A resolução do mapa e a capacidade de se determinar a seqüência de marcadores no mapa estão diretamente relacionadas ao tamanho da amostra ou população. É certo que as amostras pequenas, por exemplo, com menos de 50 indivíduos, provavelmente terão baixa resolução de mapeamento, principalmente na detecção de QTLs de pequeno efeito (Young, 1994).

O mapeamento genético baseado na análise da freqüência dos genótipos foi idealizado por Stutervant (1913), que publicou o primeiro mapa genético. Alfred H. Stutervant trabalhou com seis genes ligados ao sexo em *Drosophila melanogaster* e, além de produzir o primeiro mapa genético com todos os genes na sua ordem correta, também propôs o princípio básico do mapeamento genético, da utilização da freqüência de recombinantes para estimar a distância entre dois genes.

É fundamental que se calculem adequadamente as distâncias entre genes e marcas, estabelecendo um ordenamento correto e formando grupos de ligações que reflitam o

número básico de cromossomos da espécie. Como os eventos de permutação ocorrem ao acaso ao longo do cromossomo, a probabilidade de recombinação é maior para locos que se encontram a uma maior distância entre si do que para aqueles mais próximos. Isto pode ser considerado como sendo a idéia básica do mapeamento genético, ou seja, a taxa de recombinação entre os locos é usada como referência para cálculo de distância e ordenamento dos genes (ou marcadores) nos cromossomos (Schuster e Cruz, 2004).

Confirmada a existência de ligação entre duas marcas é indispensável adotar métodos quantitativos para estudar o grau de associação entre elas. A metodologia de Máxima Verossimilhança é utilizada no mapeamento genético para a obtenção de várias estimativas, inclusive as da freqüência de recombinação (Liu, 1998). Este método permite a obtenção de estimadores consistentes, com eficiência assintótica e variância mínima. A confiabilidade do posicionamento das marcas ao longo do grupo de ligação pode ser avaliada considerando as variâncias associadas às estimativas de recombinação (Liu, 1998).

Ainda hoje, tanto o tamanho de população quanto o número de marcas para representação de cromossomos em grupos de ligação não são bem definidos, existindo falta de padrão pra análise de dados de trabalhos de mapeamento (Cruz, 2006).

O estabelecimento de mapas genéticos em populações exogâmicas (FIC e FMI) apresenta determinadas complicações que não são encontradas quando utilizados delineamentos a partir de linhagens endogâmicas, como populações F_2, RILs (Recombinant Inbred Lines), duplo-haplóides, retrocruzamentos, dentre outros. Em populações segregantes derivadas de linhagens endogâmicas todos os locos estarão segregando para apenas dois alelos. Em adição, a fase de ligação do duplo heterozigoto pode ser claramente determinada com base na análise da segregação dos gametas recombinantes da população (Lynch e Walsh, 1998).

Em determinadas espécies de plantas não é possível obter populações segregantes derivadas de linhagens endogâmicas, devido à auto-incompatibilidade, à depressão endogâmica ou ao longo período juvenil. Em tais espécies é preciso empregar delineamentos experimentais de populações exogâmicas, como famílias de meio irmãos e famílias de irmãos completos.

Diferentes configurações de marcadores podem estar segregando em famílias de irmãos completos originadas do cruzamento entre genitores derivados de uma população exogâmica. De acordo com Haseman e Elston (1972) para o caso geral de sistemas multialélicos com quatro ou mais alelos, haverá basicamente três categorias e sete distintos

tipos de acasalamentos ou diferentes tipos de pares de irmãos que caracterizam a herança de marcas individuais, sendo : (1) cruzamento entre genitores homozigotos; (2) cruzamento entre um genitor homozigoto e um genitor heterozigoto e (3) cruzamento entre dois genitores heterozigotos.

De acordo com Lynch e Walsh (1998), existem três categorias de acasalamentos quanto ao grau de informação da progênie: (1) famílias derivadas de cruzamentos completamente informativos; (2) famílias derivadas de 'retrocruzamentos' e (3) famílias derivadas de 'intercruzamentos'. Considerando-se dois locos segregando e marcas codominantes, 81 configurações podem surgir da combinação dos diferentes tipos de acasalamentos citados anteriormente. Destas configurações, 17 proporcionam informação sobre ligação e o restante não contém informação de ligação. Para marcadores dominantes, 7 de 9 configurações proporcionam informação de ligação (Liu, 1998).

O objetivo deste trabalho foi o de gerar e analisar dados a partir de um programa de simulação de genoma e de populações, e com base nestes dados simulados avaliar o tamanho ótimo de populações para estudo de mapeamento genético com famílias de irmãos completos.

2. Revisão de Literatura

O tamanho da amostra assume fundamental importância em trabalhos relacionados a mapeamento, uma vez que a resolução do mapa e a capacidade de se determinar a seqüência de marcadores em grupos de ligações estão diretamente relacionadas com o número de indivíduos genotipados. Alguns trabalhos constataram que amostras de tamanho reduzido, com menos de 50 indivíduos, proporcionado baixa resolução no mapeamento e torna difícil a detecção de QTLs de pequeno efeito (Young, 1994).

Populações com poucos genótipos podem não permitir a observação da quebra de ligação entre os marcadores e a conseqüente determinação da distância entre estes. O tamanho adequado da população também depende do tipo de população, sendo que populações de retrocruzamento apresentam, aproximadamente, metade do conteúdo informativo de populações F_2 (Cruz, 2006).

Lannes et al., (2004), em estudo com arroz, empregaram duas populações de retrocruzamento, uma com 53 e a outra com 74 indivíduos, utilizando 138 primers, sendo 131 RAPD obtidos da University of British Columbia e sete que haviam amplificado bandas polimórficas mais consistentes entre os pais. Eles comentaram que os mapas genéticos obtidos em seus estudos poderiam ser apenas considerados como básicos, uma vez que o número de marcadores mapeados e os tamanhos das populações de mapeamento foram relativamente pequenos. Uma outra observação feita pelos autores foi a presença de sete intervalos que apresentaram distância de 0 cM entre marcadores adjacentes. Segundo os autores, isto pode ter acontecido devido ao pequeno tamanho da população, 53 indivíduos, utilizada para a formação deste mapa. Nesta situação a falta de observação de recombinantes, pelo reduzido tamanho da amostra, faz com que as estatísticas assumam valor de distância igual a zero que, muitas vezes, não é verdade.

Weller (1986) e Patterson et al., (1988), grandes variedades de efeitos quantitativos podem ser associadas a marcadores genéticos específicos e sugerem que uma porção significativa dos efeitos de características quantitativas é de uma magnitude que poderia ser prontamente detectada em experimentos com aproximadamente 1000 indivíduos de população F_2 ou de retrocruzamento. Porém Cruz (2006) sugere que não há necessidade de um tamanho tão grande de indivíduos para que sejam detectados efeitos de características quantitativas em populações de retrocruzamento, apesar deste tipo de população, em certos casos, necessitar de um maior número de indivíduos que populações F_2.

Sabe-se que o desenvolvimento da metodologia de construção de mapas genéticos remonta do início do século passado. Após a redescoberta do trabalho de Mendel em 1900, trinta e quatro anos depois de sua publicação, várias pesquisas foram realizadas com o objetivo de ampliar e validar suas conclusões em relação ao mecanismo de herança de características quantitativas. Bateson e Punnet (1905), citados por Stutervant (1965), trabalhando com características cor da flor e formato do grão de pólen, em ervilha, publicaram um dos primeiros relatos de ligação gênica (Rocha et al., 2003). Entretanto, a primeira evidência de que os genes estão localizados em posições definidas foram dados por Morgan (1910), em seu trabalho com análise do padrão de herança de um gene mutante ligado ao sexo, em *Drosophila melanogaster*.

O mapeamento genético baseado na análise da freqüência dos genótipos recombinantes foi idealizado por Stutervant (1913), que publicou o primeiro mapa genético. Alfred H. Stutervant trabalhou com seis genes ligados ao sexo em *Drosophila melanogaster*. Ele não apenas produziu o primeiro mapa genético com todos os genes na sua ordem correta, como também propôs o princípio básico do mapeamento genético, da utilização da freqüência de recombinantes para estimar a distância entre dois genes.

Para obter mapas de ligações é necessário o cálculo das distâncias entre pares de genes que possibilita o estabelecimento de um agrupamento e posterior ordenamento desses genes, resultando na formação dos grupos de ligações. A probabilidade de recombinação é maior para locos que se encontram a uma maior distância entre si do que para aqueles mais próximos, sendo este fato considerado a idéia básica do mapeamento genético. Assim, a taxa de recombinação entre os locos é usada como referência para ordenamento dos genes (ou marcadores) nos cromossomos (Schuster e Cruz, 2004).

De modo prático, para populações como retrocruzamentos, a freqüência de recombinação entre dois locos é obtida pelo número de recombinantes dividido pelo total de indivíduos analisados. Sendo assim, conhecendo-se as freqüências de recombinação entre diversos locos do mesmo grupo de ligação, é possível estabelecer um agrupamento e, posteriormente, estimar a ordem dos locos no grupo de ligação. Define-se um grupo de ligação como um conjunto de marcadores genéticos que possuam menos de 50% de recombinação entre dois marcadores consecutivos (Schuster e Cruz, 2004).

A existência de falta de aditividade das freqüências de recombinação levou ao desenvolvimento de funções de mapeamento, que são utilizadas para converter freqüências de recombinação em medidas de distâncias, com propriedades mais interessantes para o

ordenamento de locos. Das funções de mapeamento, as mais conhecidas são as de Haldane (1919), que admite a independência das permutas nos intervalos adjacentes, e a de Kosambi (1944), que considera a interferência.

A função de Haldane é expressa por:

$$m = \frac{-\ln(1-2r)}{2}$$

em que m é a distância em Morgan. Para obter a distância em centiMorgan (cM), basta multiplicar por 100.

Para obter a freqüência de recombinação a partir da distância em Morgan utiliza-se a expressão:

$$r = \frac{1-e^{-2m}}{2}$$

O mapa de ligação de uma espécie pode ser definido conjunto de marcadores ligados e não ligados. Dois marcadores são ditos ligados sempre que menos de 50% dos gametas produzidos apresentam genótipos recombinantes para esses dois genes (Stutervant, 1913).

Para Ooijen (1992) o desenvolvimento de mapas de ligação, com grande número de marcadores moleculares, tem estimulado a busca por métodos para mapeamento de genes envolvidos no controle de características quantitativas (QTL). Um método promissor proposto por Lander e Botstein (1989), já citado, emprega pares de marcadores vizinhos para obter o máximo de informação da ligação de QTLs dentro do segmento de cromossomo analisado. Os autores investigaram a acurácia deste método foi investigada por simulação computacional. Os resultados obtidos mostraram que existe uma probabilidade razoável de detecção de QTL que explique pelo menos 5 % da variância. Tanto o número de indivíduos quanto o tamanho relativo do efeito genotípico do QTL são fatores importantes na determinação da precisão do mapa em detectar possíveis genes de interesse. Na média, um QTL com capacidade para explicar 5 ou 10% da variância do

caráter é possível de ser mapeado a distâncias de 40 ou 20 cM, respectivamente. É claro que QTLs com maiores efeitos genotípicos serão localizados mais precisamente. Contudo, deve ser notado que o comprimento do intervalo é variável.

É comentado por Ooijen (1992) que o procedimento de mapeamento de QTL de Lander e Botstein (1989) permite determinar a posição de QTLs em um mapa, com limitações. QTLs com pequeno efeito aditivo ($\sigma^2_{exp} = 1\%$, sendo σ^2_{exp} a variância explicada pelo QTL) são muito difíceis de serem detectadas. Uma população de pelo menos 200 indivíduos é necessária, a menos que se esteja interessado apenas em genes de efeito muito grande ($\sigma^2_{exp} > 10\%$). O tamanho de população de 400 indivíduos parece ser o maior número possível praticamente em relação a trabalhos com RFLP. Contudo pode ser esperado que, com este procedimento de mapeamento, QTLs que explicam uma variância de pelo menos 5% terão boa chance de ser detectados.

Para a maioria dos experimentos destinados ao mapeamento, o mapa de ligação é sempre estabelecido pelos mesmos indivíduos nos quais o mapeamento de QTL já foi, ou será, feito. Tal mapa pode ser menos preciso, mas irá relacionar melhor os eventos de recombinação nos cruzamentos disponível para o pesquisador.

Outra questão que surge é o fato de que, na prática, sempre haverá dados perdidos nos genótipos marcadores. Uma conseqüência imediata é a redução do LOD escore, uma vez que a quantidade de dados perdidos vai variar dependendo do marcador, isto tem um efeito na comparabilidade dos LOD escores de diferentes partes do cromossomo, a não ser que a quantidade não seja excessiva.

O tipo de marcador utilizado na análise também merece atenção. Marcadores dominantes levarão a um menor LOD escore e conseqüentemente comparações de LOD escores baseados em dominantes com LOD escores baseados em marcadores co-dominantes não serão apropriadas.

Existem algumas etapas a serem seguidas para construção de um mapa de ligação. Em geral, o primeiro passo na construção de um mapa de ligação está relacionado com a escolha dos genitores a serem cruzados, de forma que maximize o polimorfismo genético. Uma vez selecionados os genitores é necessário o desenvolvimento de uma população segregante, composta de pouco mais de uma centena de indivíduos. É importante lembrar que o número de marcadores polimórficos depende do polimorfismo genético entre os

genitores, e que a precisão das estimativas de recombinação depende fundamentalmente da escolha da população (Gratapaglia e Sederoff, 1994).

Após a escolha dos genitores e o desenvolvimento da população segregante, a etapa seguinte envolve a obtenção de marcas contrastantes entre os genitores que apresentam segregação mendeliana na população de mapeamento. A estratégia de busca pelas marcas polimórficas depende, principalmente, do tipo de marcadores utilizados e da diversidade genética da espécie estudada (Rocha et al., 2003).

Após a escolha dos marcadores polimórficos é necessário analisar o padrão de amplificação dos indivíduos avaliados no mapeamento e obter as estimativas de recombinação. Para construir um mapa genético todos os marcadores devem ser analisados dois a dois, verificando a independência ou a existência de ligação entre eles (Liu, 1998). Com base no princípio de que os genótipos recombinantes alterados por permuta simples ou por permuta dupla são gerados em freqüências diferentes, utiliza-se o teste de aderência χ^2 (qui-quadrado) para confirmar a ligação entre os marcadores. O teste de χ^2 é qualitativo, pois apenas comprova a existência ou não de ligação gênica. Fundamenta-se na comparação dos desvios entre os resultados esperados, sem a ocorrência de permutas, com os resultados observados, sendo sensível à magnitude destes desvios e ao número de genótipos amostrados, Falconer (1987).

Confirmada a existência de ligação entre duas marcas é indispensável adotar métodos quantitativos para estudar o grau de associação entre essas marcas. A metodologia de Máxima Verossimilhança é utilizada no mapeamento genético para a obtenção de várias estimativas, inclusive as da freqüência de recombinação (Liu, 1998). Este método permite a obtenção de estimadores consistentes, de distribuição normal, eficiência assintótica, e variância mínima. A confiabilidade do posicionamento das marcas ao longo do grupo de ligação pode ser avaliada considerando as variâncias associadas às estimativas de recombinação (Liu, 1998).

Tanto o tamanho de população quanto o número de marcas para representação de cromossomos em grupos de ligação não são bem definidos nos estudos atuais, existindo falta de padrão pra análise de dados de trabalhos de mapeamento. Com as informações do Quadro 1, pode-se ter idéia da falta de padrão em relação ao número de indivíduos e de marcas a serem usadas em trabalhos com mapeamento genético.

Quadro 1. Exemplos de trabalhos de diversos autores com espécies, tipo de população trabalhada, tamanho de população e número de marcadores diferentes.

Espécie	Autor	Tipo de População	Tamanho	Nº de Marcadores
Bovinos	Telles (2001)	3 Rebanhos	66	133
Erva Mate	Vidor (2002)	Acessos	2000	168
Eucalipto	Falcão (2004)	Parentais	8	1261
Feijão	Corrêa (2000)	F_2	302	440
Feijão	Faleiro (2003)	RILs	154	70
Maracujá	Ganga (2004)	Acessos	36	123
Milho	Brunelli (2002)	linhagens, hibrido e F_2 das mesmas	165	142
Milho	Silva (2002)	F_2	250	140
Morango	Conti (2002)	Cultivares	26	63
Simulado	Carbonell (1993)	Retrocruzamento e Duplo haplóides	25000	48
Simulado	Knott (1992)	F_2	1000	22
Soja	Corrêa (1999)	F1	50 e 120	238

Tamanho: é número total de indivíduos analisados; Nº de Marcadores: é o número de marcadores utilizados no genoma de cada indivíduo.
Fonte: Cruz(2006)

2.1. Simulação

A simulação consiste em construir um sistema que imite o funcionamento de uma realidade, com a finalidade de averiguar o que aconteceria no sistema real se alterações de interesse fossem efetuadas em seu funcionamento (Dachs, 1988). Informações valiosas podem ser extraídas desse sistema simulado, com menor custo e maior rapidez. No sistema real, muitas das opções de alteração são inviáveis de serem avaliadas, sejam pelos custos que podem ser elevados, pelos longos períodos de resposta ou pela incerteza de direção e sentido das respostas, fatores estes que podem conduzir a um dano irreparável a este sistema.

No melhoramento genético, o tempo se torna fator limitante. Além do tempo necessário, a necessidade de laboratórios bem equipados aumenta os custos, dificultando muitas vezes a realização de determinados trabalhos de pesquisa. Para contornar este problema, pesquisadores utilizam técnicas de simulação, que permitem a obtenção de um

grande volume de dados em um curto período de tempo, sem os custos de implantação e condução de experimentos com animais ou plantas e com laboratórios (Corrêa, 2001).

Ao se fazer estudo baseado em simulação, deve-se considerar que nada é tão simples que possa ser compreendido e controlado sem abstração. Essa abstração consiste em substituir o objeto de interesse por um modelo semelhante, porém com estrutura mais simples. Outro aspecto importante na simulação é a modelagem. O modelo deve ser suficientemente simples para ser operacionalizado e interpretado adequadamente, mas seu desempenho deve ser comparável com o modelo real e, se a defasagem for grande, ele deve ser eliminado ou refinado (Cruz, 2001). Ao utilizar-se de uma técnica de simulação o pesquisador deve precaver-se contra erros, seja estes devidos a problemas como levantamentos amostrais, escolha inadequada das distribuições de probabilidades nos eventos de natureza aleatória, simplificação inadequada da realidade e erros de implantação do sistema simulado. Para a garantia de sua eficiência pode-se lançar mão de processos de validação. Essa validação consiste em fazer o sistema simulado operar nas condições do sistema real e verificar através de testes de hipóteses e outras análises estatísticas ou através de comparação com situações reais já analisadas, se os resultados observados na simulação condizem com os observados no sistema real (Cruz, 2001).

Dempster et al., (1977) apresentam abordagem sobre um método computacional iterativo para estimações de máxima verossimilhança, quando as observações são classificadas como dados incompletos. A partir do momento em que cada iteração do algoritmo consiste em uma etapa de estimativa e outra de maximização este algoritmo é chamado de EM. O algoritmo tem uma ampla aplicabilidade, incluindo dados perdidos, dados truncados, modelos de “misturas finitas” e estimação de componentes de variância, análise de fatores e outros, além de ser relativamente simples.

Martinez et al., (1999) desenvolveram procedimento de mapeamento baseado em modelo randômico utilizado para se investigar sua robustez e adequação para mapeamento de QTL em populações onde prevalecem estruturas de famílias de meios irmãos. Sob o modelo randômico, a localização do QTL e componentes de variância foram estimadas usando técnicas de máxima verossimilhança. A estimação de parâmetros é feita baseando-se na abordagem do modelo de pares de irmãos. A proporção de genes idênticos por descendência (IBD) no QTL foi estimada através de dois marcadores flanqueadores. Já as estimativas para os parâmetros e poder de QTL foram obtidas usando dados simulados, variando o número de famílias analisadas, a herdabilidade da característica, a variância, o

número de marcadores e o número de alelos no QTL. Os fatores mais importantes que influenciaram o poder e os parâmetros do QTL foram a herdabilidade e a variância genética da característica. Segundo os autores, o número de alelos do QTL não influenciou as estimativas dos parâmetros avaliados, nem o poder de detecção do QTL. Com uma herdabilidade mais alta, foi observado um confundimento entre QTL e componentes poligênicos.

Em relação ao trabalho de Martinez et al., (1999), a técnica de simulação Monte Carlo foi utilizada para gerar dados genotípicos e fenotípicos. O mapeamento de QTL foi considerado para um segmento de cromossomo de 100 cM de comprimento, coberto por seis marcadores distribuídos igualmente ao longo do cromossomo a uma distância de 20 cM entre eles. Todos os marcadores tinham igual número de alelos e de mesma freqüência. Um único QTL com vários alelos codominantes de mesma freqüência e de efeito aditivo foi simulado no meio do segmento cromossômico, 50cM, os pais foram gerados através da alocação aleatória dos genótipos em cada loco, assumindo equilíbrio de Hardy-Weinberg. A fase de ligação parental foi assumida como sendo desconhecida. Descendentes foram gerados assumindo-se não haver interferência, desse modo, o evento de recombinação em um intervalo não irá interferir num evento de recombinação em um intervalo adjacente. As frações de recombinação, para cada loco, foram calculadas usando a função de mapeamento de Haldane. É importante mencionar que em cada simulação dois tamanhos diferentes de amostras foram considerados, 50 e 100 famílias com 25 descendentes cada.

2.2. Contribuição dos estudos de simulação na análise genômica

Direta ou indiretamente a simulação tem contribuído grandemente para o avanço tanto na genômica quanto para as demais áreas do melhoramento.

Uma dessas contribuições é a possibilidade de serem simulados diferentes tipos de populações, permitindo menor gasto em material e mão de obra, além do ganho de tempo. As simulações podem também ser feitas com base em dados reais, Frish, et al., 1999, para verificar a confiabilidade do software utilizado, compararam o mapa de ligação original utilizado em seu trabalho, no qual foi baseado em dados experimentais de F_2, com um mapa de ligação construído a partir de dados simulado de indivíduos F_2, pelo programa MAPMAKER (Lander et al., 1987). Pela comparação foi visto que os mapas estavam em

excelente concordância, confirmando que os modelos usados nos dois programas eram similares.

Mais uma vantagem do uso da simulação pode ser encontrada no trabalho de Visscher et al. (1996) onde a simulação foi utilizada para se avaliar eficiência do uso da seleção assistida na introgressão de genes em programas de retrocruzamento. No trabalho de Martínez e Curnan (1992) também foram feitas simulações correspondentes a um retrocruzamento, para se verificar, dentre outras, o efeito de QTL.

Uma outra vantagem da simulação é que esta pode ser empregada em qualquer etapa de um programa de melhoramento, seja este clássico ou não, para se comprovar ou refutar novos procedimentos e técnicas. Uma das formas de simulação vista ao longo dos textos anteriores foi a utilização da simulação para obtenção de dados sobre determinadas populações. Desse modo, após se chegar a um modelo para detecção de QTL, por exemplo, podem-se testar sua eficiência em populações de retrocruzamento, F_1, F_2, duplo-haploides, linhas endogâmicas e outras, sem o custo operacional e laboratorial que poderia ser gerado. Para serem simuladas populações de diversos tipos de famílias podem-se ser usadas características obtidas de dados reais, como tipo de distribuição, média, variância e demais características que se mostrarem necessárias.

Uma clara vantagem do uso da simulação é o número de amostras que podem ser geradas. Como exemplo pode-se citar o trabalho de Martinez e Curnan (1992), que teve o objetivo de ilustrar como é possível aparecer um “QTL fantasma” quando é usada a análise de mapeamento por intervalo de Lander e Botstein (1989) ou o mapeamento por regressão usando apenas marcadores flanqueadores. Neste estudo foi utilizada uma amostra com 2000 observações. Este número elevado de observações foi necessário para que fosse observado melhor o viés sistemático, ao invés da variação de amostra proporcionada pelo tamanho reduzido de observações.

Em seu trabalho sobre técnicas de máxima verossimilhança para mapeamento e análise de QTL com auxílio de marcadores genéticos, Weller (1986) apresentou dados simulados que indicaram boa concordância entre os efeitos preditos e os efeitos obtidos em dez repetições de 2000 indivíduos F_2. Pelo comentado, a técnica se mostrou mais eficiente para genes codominantes que para dominantes.

No trabalho de Knott e Haley (1992), para se investigar as propriedades dos métodos analíticos, dados simulados foram usados. Foram gerados descendentes de parentais completamente heterozigotos, assumindo que não exista interferência e a

aditividade em Morgan das distâncias de mapa. Cada conjunto de dados continha 1000 indivíduos F_2, o genótipo de cada indivíduo foi composto por um par de cromossomos de 100 cM em comprimento. Foram simulados onze locos marcadores a um intervalo de 10 cM entre eles. Também foram simuladas diferentes situações de ligação de QTLs.

Uma das dificuldades atualmente enfrentadas pelos pesquisadores é a identificação da posição de QTLs e do tamanho do seu efeito. Através da simulação pode-se inferir e testar, sob determinadas pressuposições, o que aconteceria sob determinadas condições, como um maior distanciamento entre marcas moleculares, qual seria o efeito da presença de um ou mais QTLs próximos a um QTL flanqueado por marcadores, o que aconteceria caso existissem QTLs externamente a um intervalo limitado por marcadores, porém, sem a existência de QTL entre eles, e outras situações que se mostrassem de interesse a serem analisadas. Martínez (1992), no trabalho em que procurou estimar a localização e o tamanho dos efeitos de QTLs, usando marcadores flanqueadores, gerou um conjunto de dados simulados que exemplificaram o problema a ser pesquisado. Os dados foram analisados pelo método de mapeamento por intervalo e pelo modelo de regressão.

2.3. Aplicativos utilizados para simulação

Diversos softwares são encontrados com a finalidade de realizar simulação de dados. Aqui será feito um breve relato sobre os dois que foram utilizados neste estudo.

O programa GENES, amplamente utilizado em análises de modelos aplicados ao melhoramento de plantas e animais, é um software destinado à análise e processamento de dados por meio de diferentes modelos biométricos, contando com procedimentos uni e multivariados, enfatizando estimação de parâmetros genéticos. Também estão disponíveis procedimentos para análise de dados binários, geralmente obtidos de estudos moleculares, permitindo a análise e interpretação de fenômenos particulares desta área (Cruz, 2004). Possui ainda o modo “simulação”, onde o usuário poderá avaliar tamanhos de amostras, número de famílias, plantas, repetições, em diferentes estudos.

O programa GENES está disponível para download, gratuitamente, sem nenhuma restrição de uso ou de divulgação, no endereço http://www.ufv.br/dbg/genes/genes.htm. Conta com manual comercializado pela Editora UFV. O E-mail: editora@ufv.br ou site: http://www.livraria.ufv.br/ .

O programa GQMOL (Cruz, 2005), foi desenvolvimento com o propósito de analisar dados obtidos de estudos moleculares. O programa pode ser usado análise de segregação de locos individuais, estimação de porcentagem de recombinação, agrupamento de marcas moleculares e mapeamento, incluindo estudos de QTL com populações controladas, populações exogâmicas, simulação e análise de imagem. Além disso, conta com um módulo de ensino, apresentando vários procedimentos para entendimento de princípios estatísticos e genéticos envolvidos na análise genômica. Possui também o módulo "simulação", onde diferentes genomas podem ser simulados, levando em conta diferentes tamanhos amostrais, tipos de populações, variáveis quantitativas entre outras variáveis.

O programa GQMOL está disponível para download, gratuitamente, sem nenhuma restrição de uso ou de divulgação, no endereço http://www.ufv.br/dbg/gqmol/gqmol.htm.

Os dois programas citados tem sido de grande utilidade na área de melhoramento genético uma vez que conta com recursos para análise e processamento de dados fundamentados em diferentes metodologias biométricas. Também tem sido indispensável em muitos estudos por contar com módulo de simulação em que é possível estabelecer diferentes genomas, níveis de saturação, tipo de marcadores, grau de polimordismo, tipo de população segregante e características quantitativas.

O uso destes aplicativos em estudos genômicos tem sido rotineiro, proporcionando valiosas contribuições na experimentação científica.

2.4. Mapeamento em Família de Irmãos Completos

A utilização de família de irmãos completos para avaliações genéticas, pode contribuir muito para o avanço nos estudos de mapeamento genético e detecção de QTLs em diferentes culturas. Culturas em que o cálculo da freqüência de recombinação entre pares de locos é mais complexo de ser obtidas, seja por possuírem ciclo longo, sejam por possuir uma elevada depressão por endogamia que dificulta ou até mesmo impossibilita a obtenção de linhagens homozigóticas e, portanto, o emprego de delineamentos clássicos de mapeamento, tais como populações F_2, retrocruzamentos e linhagens recombinantes endogâmicas (RILs). Dessa maneira, as famílias disponíveis para mapeamento são derivadas do cruzamento de parentais heterozigóticos, podendo ter na população de mapeamento até quatro alelos segregando para cada loco, tornando complicadas as análises de ligação por causa da existência de muitos tipos de razões de segregação entre pares de

marcas (Maliepaard et al., 1997). Com a estratégia de pseudocruzamentoteste (Grattapaglia e Sederoff, 1994), onde os marcadores segregantes são analisados separadamente para cada parental, a utilização de genitores heterozigóticos para a construção de mapas de ligação tornou-se possível.

Outro fato importante é que família de irmãos completos é uma família de melhoramento e de mapeamento. Isto é uma crítica realizada, por exemplo, quando se utiliza RILs, que é um tipo de delineamento que é usado apenas para mapeamento, e não tem fins práticos de melhoramento.

Sendo assim se torna necessário a utilização das informações dos irmãos completos, tanto no mapeamento e também na detecção de QTL, para que consiga utilizar a maior quantidade possível de informações geradas no mapeamento.

Novaes (2006), utilizando uma família de irmãos completos de Eucalyptus constituída de 188 indivíduos obtidas através do cruzamento entre *E. grandis* x *E. urophylla*, realizou o mapeamento genético e detecção de QTL para qualidade de madeira. Em seu estudo foi utilizada a estratégia de pseudocruzamento e posterior integração de mapas. Este autor encontrou QTL associados a todas as características fenotípicas avaliadas.

É importante além de poder usar estas informações deste tipo de população, ter uma consistência em relação ao tamanho de populações ideais a serem utilizadas no mapeamento de uma família de Irmãos Completos. Na literatura, diferentes tamanhos de populações são utilizadas sem uma maior preocupação de avaliar se o tamanho utilizado é realmente suficiente para a obtenção de qualidade confiável. Ukrainetz et al. (2008), usou 8 famílias de irmãos completos com 40 indivíduos cada, de Douglas-fir (*Pseudotsuga menziesii)* para fazer mapa de ligação utilizando marcadores AFLP, a partir da técnica descrita por Hu et al.,(2004) para a construção do mapa. Os autores ainda comenta que o mínimo de quatro famílias deve ser utilizado, para que pelo menos uma delas seja segregante, e o tamanho destas varia de acordo com a freqüência de recombinação encontrada. Para freqüência de recombinação inferior a 0,1 ($r<0,1$) um tamanho de 10 indivíduos seria adequado. Para $r < 0,25$ de 20 a 30 indivíduos seriam necessários por família, e para $r < 0,3$ seriam necessários 40 indivíduos.

Outro trabalho relevante ao que diz respeito a simulação para inferir sobre o tamanho de populações ideal de irmãos completos foi descrito por Silva et al. (2004), o qual avaliou diferentes de tamanhos de populações necessários para que fosse possível a

identificação de QTL. Este autor sugere a utilização de pelo menos 50 FIC com 10 indivíduos cada para que aumente as chances de identificação de QTL, confrontando com o sugerido por Silva (2002), que sugere a utilização de 50 famílias com 40 indivíduos cada. Entretanto, segundo Goddard et al. (1999), aumento na acurácia para a detecção de QTLs em pequenas famílias de irmãos completos pode ser obtido se forem analisadas amostras selecionadas de irmãos, excluindo-se indivíduos não-informativos, concordando com as afirmações de Guo e Elston (2000). Tal afirmação é confirmada por Chatziplis e Haley (2000), os quais afirmaram que pares de irmãos completos, oriundos de famílias com baixa variância, não estariam segregando para o QTL e, dessa forma, são não-informativos para a análise.

3. Material e Métodos

3.1. Simulação de dados

Para gerar os dados foi utilizado o módulo de simulação do aplicativo computacional GQMOL (Cruz, 2005), o qual permite gerar informações sobre genomas, genótipos, genitores, indivíduos de diferentes tipos de populações e dados de características quantitativas. Foram simulados genomas parentais e amostras de populações de família de irmãos completos do tipo completamente informativas, e também de famílias de irmãos completos que não fossem completamente informativas. As amostras geradas foram de tamanho 100, 200, 400 e 600 indivíduos com 3 grupos de ligação cada. Também foram geradas matrizes de distâncias entre pares de locos. Depois de feita simulação e gerada a matriz de distância foi avaliada a eficácia de reconstituição do número de grupos de ligação a partir dos diferentes tamanhos e tipos de amostras.

3.1.1. Simulação do genoma

Foi tomada como referência uma espécie diplóide fictícia com 2n = 2x =6 cromossomos, cujo comprimento total do genoma, por grupo de ligação, foi estipulado em 100 cM. Foi gerado o genoma com nível de saturação de 11 marcas moleculares (ou 10 cM de intervalo) por grupo de ligação. Cada genoma foi composto por 3 grupos de ligação, 100 cM em cada grupo, com comprimento total de 300 cM(Figura 1).

Foram usadas nas simulações marcas codominantes tanto para a população completamente informativa, quanto para a população não-completamente informativa.

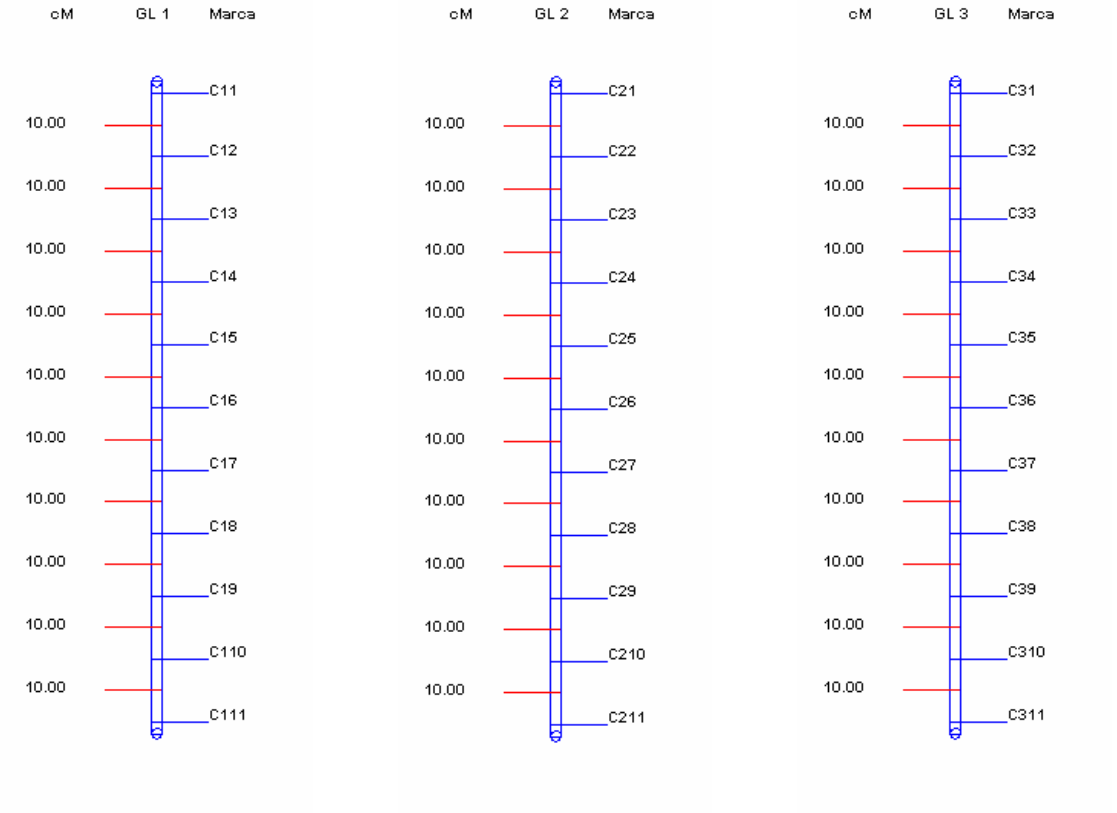

Figura 1. Grupos de ligação 1, 2 e 3 com 11 marcadores cada, evidenciando o grau de saturação do genoma simulado.

3.1.2. Simulação dos genitores

Para a simulação dos genitores foram utilizadas duas situações, uma para os marcadores completamente informativos, em que foi considerado genitores do tipo A_iA_j x A_kA_l a fim de originar as famílias de irmãos-completos FIC (Figura 2), e outra para a simulação dos genitores não completamente informativos para formar a segunda população de irmão completos (Figura 3). Para esta segunda situação considerou-se a presença de quatro alelos na população tomados ao acaso, e com freqüências iguais de 0,25 para cada alelo, para que tivesse a mesma probabilidade de ocorrências das combinações genotípicas, fazendo com que todas combinações possíveis aparecessem na população simulada. Nesta condição o PIC (conteúdo de informação de polimorfismo) é igual a 0,7031.

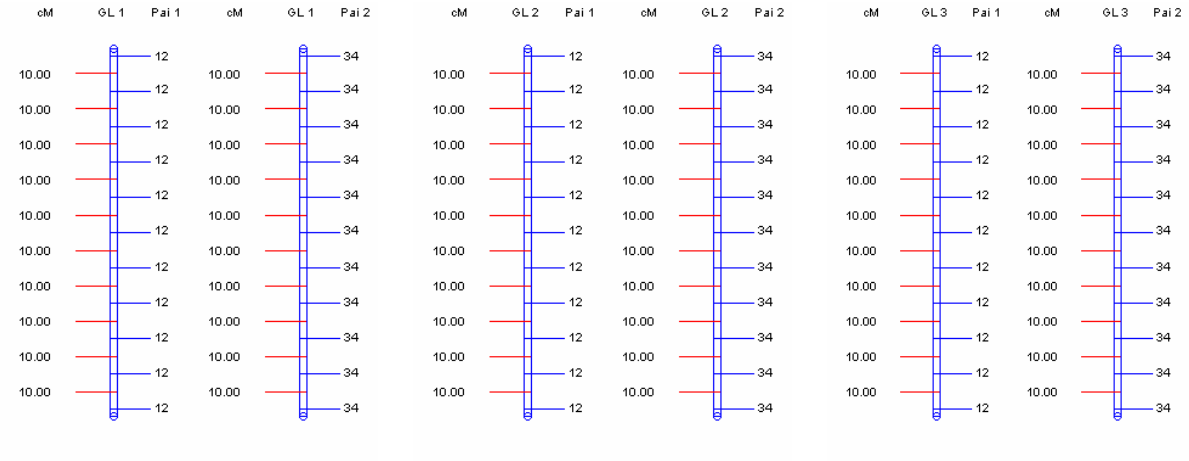

Figura 2. Genoma dos pais 1 e 2, nível de saturação de 10cM, para os genitores da população completamente informativa (12 x 34).

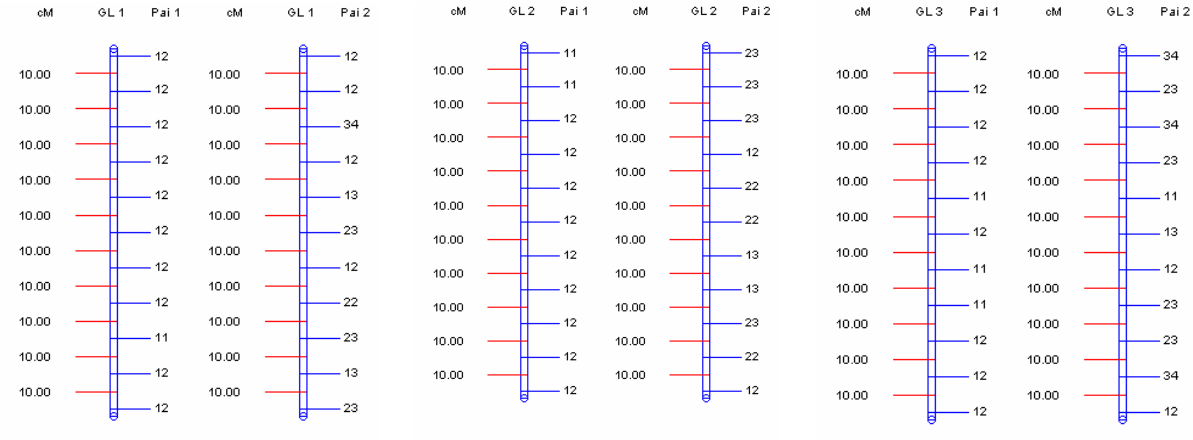

Figura 3. Genoma dos pais 1 e 2, nível de saturação de 10cM, para os genitores da população não completamente informativa.

3.1.3. Tamanho da população

Tanto para a população de irmãos completos completamente informativa quanto para a não completamente informativa, foram geradas amostras com 100, 200, 400 e 600 indivíduos, com saturações de 11 marcas por grupo de ligação e 100 repetições, chegando a um total de 800 simulações, sendo 400 famílias completamente informativas e 400 não completamente informativas.

3.1.4. Procedimento de simulação dos indivíduos das FIC

A estratégia básica de simulação é caminhar ao longo dos cromossomos, realizando permutas em cada intervalo entre marcas adjacentes, de acordo com as distâncias dos marcadores, conforme descrito por Silva (2005).

O processo de simulação das famílias de irmãos completos completamente informativas bem como as não completamente informativas seguiu os seguintes passos:

i) a partir do genoma simulado foram construídos os genótipos parentais, de forma que a população completamente informativa apresentaria genitores que seriam sempre de padrão genotípico 12 x 34; já para a população não completamente informativa cada pai poderia possuir qualquer uma das 16 possíveis combinações alélicas (11, 12, 13,..., 34, 44).

ii) a partir dos genótipos parentais foram gerados os gametas para a formação dos indivíduos das populações de irmãos completos. A produção de gametas foi feita simulando-se o pareamento dos homólogos e realizando-se permutas ao longo dos cromossomos, considerando a não existência de interferência, nas regiões delimitadas por dois marcadores adjacentes. A probabilidade de ocorrência de recombinação numa região entre marcadores adjacentes foi dada de acordo com a distância destes marcadores no genoma simulado. Sendo que uma maior distancia implica numa maior possibilidade de ocorrência de recombinação.

O programa GQMOL considera o encontro aleatório de gametas para a simulação dos indivíduos. Sendo assim, um novo processo acontece para cada indivíduo simulado dentro de cada repetição.

3.2. Análise genômica – Mapeamento

Após a geração dos dados, seguiram-se as etapas do processo de mapeamento, como descrito a seguir.

3.2.1. Análise de segregação de locos individuais

Foram aplicados testes de qui-quadrado (χ^2) para verificação da razão de segregação em cada marca de todas as populações geradas. No processo de mapeamento foram utilizadas todas as marcas, mesmo as que não segregaram de acordo com a

proporção esperada de 1:1:1:1 (A_iA_k:A_iA_l:A_jA_k:A_jA_l) para completamente informativos e 1:1 ou 1:2:1 ou 1:1:1:1 na população não completamente informativa.

A estatística de qui-quadrado é dada por:

$$\chi^2 = \sum_{i=1}^{n}\left[\frac{(Obs_i - Esp_i)^2}{Esp_i}\right]$$

em que

χ^2 é o valor de qui-quadrado calculado;

Obs_i e Esp_i, são os valores observados e esperados, para a i-ésima classe fenotípica (i = 1, 2, ..., n), respectivamente.

A hipótese (H_0) de segregação dos locos foi testada a 5 % de probabilidade. Nas situações em que o valor da probabilidade calculado foi inferior ao pré-estabelecido, a hipótese H_0 foi rejeitada significando que a segregação não ocorreu de acordo com o esperado.

3.2.2. Estimação da percentagem de recombinação

Após a aplicação dos testes de segregação, seguiu-se a etapa da estimação da porcentagem de recombinação entre pares de marcas, utilizando o método da máxima verossimilhança.

As expressões utilizadas para a estimação considerando os diferentes tipos de cruzamentos, foram aquelas detalhadamente explicadas no capítulo 1.

3.2.2.1. População Completamente informativa

A descendência da população completamente informativa foi estabelecida a partir do cruzamento de genitor A_1A_2 com o genitor A_3A_4.

A partir deste cruzamento, quatro possibilidades de fase de ligação acontecerão para os dois locos envolvidos, sendo elas: aproximação-aproximação, aproximação-repulsão, repulsão-aproximação e repulsão-repulsão. Desta forma é possível estimar quatro diferentes

medidas de distância, sendo adotada como a verdadeira fase de ligação entre os locos que proporcionar a maior estimativa de LOD.

A estimação da percentagem de recombinação para marcadores completamente informativos pode ser feita pelo método da máxima verossimilhança levando-se em consideração a função de verossimilhança:

$$L(r;n_i) = \lambda\left(P^2\right)^{n1}(PR)^{n2}\left(R^2\right)^{n3}$$

esta estimação foi toda detalhada no capítulo 1, de forma que aqui só será mostrado a expressão para a obtenção do valor da freqüência de recombinação, dado por :

$$r = \frac{2n_3 + n_2}{2N}$$

3.2.2.2. População não completamente informativa

A descendência da população não completamente informativa foi estabelecida a partir do cruzamento de genitores onde para cada loco era possível a ocorrência de um dos quatro alelos; de forma que haveria, pra cada genitor, a possibilidade de formação de 16 tipos de combinações alélicas para formar sua constituição genotípica. Considerou-se a presença de quatro alelos na população com freqüências iguais de 0,25 para cada alelo, de forma que este loco seja classificado de altamente polimórfico, uma vez que, seu alelo mais freqüente possui freqüência alélica inferior a 0,55 Ott (1992b). Nesta situação o PIC (conteúdo de informação de polimorfismo) é igual a 0,7031. O PIC é usado para quantificar o polimorfismo da população. Acredita-se que quanto maior o PIC, maior o conteúdo de informação de ligação. Para esta situação onde os quatro alelos possuem a mesma freqüência a expressão usada pra obter o PIC é a seguinte:

$$PIC = 1 - \frac{1}{l} - \frac{1}{l^2} + \frac{1}{l^3}$$

em que, *l* é o número de alelos, com mesma freqüência, presente na população.

De posse da constituição genotípica de cada genitor observa-se o tipo de segregação para cada loco afim da obtenção do tipo de cruzamento envolvido. Mais uma vez é possível a ocorrência teórica de quatro fases de ligação, de modo que se adota como a real fase de ligação entre os marcadores aquela onde for obtido o maior LOD.

Para o caso de marcadores que não sejam completamente informativos existem diferentes expressões para a obtenção para a freqüência de recombinação. No Quadro 2 encontram-se, de forma resumida, as expressões obtidas por meio da função de verossimilhança para cada situação de cruzamento. Mais uma vez deve-se enfatizar que a obtenção destas expressões foi mostrada de forma detalhada no capítulo 1.

É importante salientar que para a estimação da freqüência de recombinação entre locos não informativos é necessário fazer uso da integração de mapas. Isto é feito para que seja possível estimar a porcentagem de recombinação nos locos não informativos. Esta porcentagem de recombinação é obtida usando informações das distâncias entre locos adjacentes às marcas cuja distância não pode ser determinada por impossibilidade de detecção de classes recombinantes. Estas marcas adjacentes funcionam como âncoras e a estimativa da freqüência de recombinação é obtida de forma indireta; depois de obtidas esta estimativa, é possível alocar tais marcas não informativas no mapa genético.

Para exemplificar a utilização da integração de mapas considera-se o cruzamento $A_1A_2B_1B_2C_1C_1$ x $A_1A_1B_3B_4C_1C_2$. Observe que não foi possível estimar a distância A e C de forma direta.

		½ A_1C_1	½ A_1C_2
½	A_1C_1	r *= ?	
½	A_2C_1		

(*) Apesar de ambos os locos segregarem, o valor da porcentagem de recombinação (r) é indeterminado.

Neste tipo de situação, é possível obter valores de distância entre A e B e entre B e C. Desta forma, a âncora B possibilitará a integração das informações e posterior estimação da distância entre A e C. Sendo assim, no caso em que nem toda progênie do cruzamento é completamente informativa, pode ser necessário o uso da estratégia de integração de mapas a fim de obter a estimativa da freqüência de recombinação de forma indireta. Um outro ponto importante a ser lembrado é que para populações não completamente informativas a estimativa da freqüência de recombinação deve ser feita através da Função de Máxima Verossimilhança, pois esta, leva em consideração todos os indivíduos da população.

Quadro 2. Esquema resumido com todas as segregações analisadas, suas classes de probabilidades observadas, função de verossimilhança e expressão geral para obtenção da freqüência de recombinação.

Cruzamento/ Segregação	Função de Verossimilhança	Expressão
(1:1:1:1) (1:1:1:1)	$L(r;n_i) = \lambda\left(P^2\right)^{n1}(PR)^{n2}\left(R^2\right)^{n3}$	$r = \frac{2n_3 + n_2}{2N}$
(1:1:1:1) (1:2:1)	$L(r;n_i) = \lambda\left(P^2\right)^{n1}(PR)^{n2}\left(R^2\right)^{n3}(2PR)^{n4}(P^2+R^2)^{n}$	$r^3[-4N] + r^2[2(2N + n_2 + 2n_3 + n_4 + n_5)$ $r[-2(N + n_2 + 2n_3 + n_4)] + (n_2 + 2n_3 + n_4$
(1:2:1)(1:1:1:1)	$L(r;n_i) = \lambda\left(P^2\right)^{n1}(PR)^{n2}\left(R^2\right)^{n3}(2PR)^{n4}(P^2+R^2)^{n}$	$r^3[-4N] + r^2[2(2N + n_2 + 2n_3 + n_4 + n_5)$ $r[-2(N + n_2 + 2n_3 + n_4)] + (n_2 + 2n_3 + n_4$
(1:2:1)(1:2:1)	$L(r;n_i) = \lambda\left(P^2\right)^{n1}(2PR)^{n2}\left(R^2\right)^{n3}(2P^2+2R^2)^{n4}$	$r^3[-4N] + r^2[2(2N + n_2 + 2n_3 + n_4)] +$ $r[-2(N + n_2 + 2n_3)] + (n_2 + 2n_3)$
(1:2:1)(1:1)	$L(r;n_i) = \lambda\left(P^2 + PR\right)^{n1}(R^2 + PR)^{n2}\left(P^2 + R^2 + 2PR\right)^{n3}$	$r = \frac{n_2}{n_1 + n_2}$
(1:1) (1:2:1)	$L(r;n_i) = \lambda\left(P^2 + PR\right)^{n1}(R^2 + PR)^{n2}\left(P^2 + R^2 + 2PR\right)^{n3}$	$r = \frac{n_2}{n_1 + n_2}$
(1:1) (1:1)	$L(r;n_i) = \lambda\left(2P^2 + 2PR\right)^{n1}(2R^2 + 2PR)^{n2}$	$r = \frac{n_2}{N}$
(1:1) (1:1:1:1)	$L(r;n_i) = \lambda\left(P^2 + PR\right)^{n1}(R^2 + PR)^{n2}$	$r = \frac{n_2}{N}$
(1:1:1:1) (1:1)	$L(r;n_i) = \lambda\left(P^2 + PR\right)^{n1}(R^2 + PR)^{n2}$	$r = \frac{n_2}{N}$

3.3. Comparação de genomas

O termo genoma simulado refere-se, genericamente, àquele obtido conforme descrito na seção 3.1.1. do item material e métodos. Já o termo genoma analisado refere-se, genericamente aos genomas construídos a partir das populações simuladas.

Foi comparado, para os genomas analisados o número de grupos de ligação obtidos, o tamanho dos grupos de ligação, as distâncias médias entre dois marcadores adjacentes nos grupos de ligação, as variâncias das distâncias entre marcas adjacentes nos grupos de ligação, o estresse, e a inversão da ordem dos marcadores, verificada pela correlação de Spearman. Todas estas comparações foram realizadas com o módulo “Comparação de Genomas” do aplicativo computacional GQMOL (Cruz, 2005).

Nas análises apresentadas foram utilizadas apenas as repetições em que houve recuperação dos 3 grupos de ligação no mapeamento genético.

3.3.1. Número de grupos de ligação e marcas por grupo

Para todos os genomas analisados fez-se uma contagem do número de grupos de ligação e do número de marcas por grupo, obtidos do mapeamento das populações simuladas.

3.3.2. Tamanho do grupo de ligação

O tamanho do grupo de ligação foi obtido somando-se as distâncias entre marcas adjacentes no grupo de ligação do genoma analisado, como segue:

$$L = \sum_{k=1}^{m-1} d_k$$

em que: L é o tamanho do grupo de ligação e d_k é à distância entre marcas adjacentes m_k e m_{k+1} no grupo de ligação do genoma analisado (k = 1,..., m-1). Sendo que m é o número de marcadores no grupo de ligação do genoma analisado.

3.3.3. Média das distâncias entre marcadores adjacentes no grupo de ligação

É a razão do tamanho do grupo de ligação pelo número de intervalos entre marcas adjacentes no grupo de ligação, como segue:

$$\overline{d} = \frac{L}{I}$$

em que: $\overline{d}$ é a distância média de dois marcadores adjacentes no grupo de ligação do genoma analisado, L é o tamanho do grupo de ligação do genoma analisado e I é o número de intervalos entre marcas adjacentes, dado por m – 1, onde m é o número de marcas no grupo de ligação.

3.3.4. Variância das distâncias entre marcas adjacentes

É a razão do somatório do quadrado dos desvios entre as distâncias de marcas adjacentes e a distância média de dois marcadores adjacentes no grupo de ligação pelo número de intervalos (I) no grupo de ligação menos 1, como segue:

$$\hat{\sigma}^2 = \frac{\sum_{k=1}^{m-1}\left(d_k - \overline{d}\right)^2}{I-1}$$

em que: $\hat{\sigma}^2$ é a variância das distâncias entre marcas adjacentes, d_k é a distância entre marcas adjacentes m_k e m_{k+1} no grupo de ligação do genoma analisado (k=1, ..., m-1), $\overline{d}$ é a média da distância de dois marcadores adjacentes no grupo de ligação do genoma analisado e I é o número de intervalos entre marcas adjacentes, dado por m-1, onde m é o número de marcadores no grupo de ligação.

3.3.5. Correlação de Spearman

A correlação de spearman, também conhecida como correlação de rank, é utilizada quando não é possível mensurar variações contínuas, como variáveis x e y nos n membros de uma população. Porém, é possível mensurar um x e um y, em forma de nota (rank), onde cada nota pode ser colocada em ordem para os n membros. Esta correlação expressa o grau de concordância nas notas das duas variáveis.

Adaptando-se a correlação de Spearman, dada por Clarke (1994), para análise de genomas, tem-se:

$$r_s = 1 - \frac{6\sum_{k=1}^{m} \Delta_k^2}{m\left(m^2 - 1\right)}$$

em que: r_s é o valor estimado da correlação de Spearman, para um grupo de ligação do genoma analisado ($-1 \leq r_s \leq 1$), Δ_k é a diferença da nota do marcador m_k (k=1, ..., m) na posição k do grupo de ligação do genoma simulado e a posição do marcador m_k na posição k do grupo de ligação do genoma analisado. Neste trabalho espera-se correlações entre 0 e 1, uma vez que teve-se a precaução de comparar genomas sem estarem invertidos.

Onde, m é o número de marcas no grupo de ligação do genoma simulado e a nota do marcador m_k, tanto no grupo de ligação do genoma simulado quanto no grupo de ligação do analisado, é o valor do índice k do referido marcador. O valor da nota do marcador não é alterado quando sua posição em relação ao grupo de ligação do genoma simulado é alterada. Deve ser lembrado que a nota refere-se à posição que determinado marcador ocupa no grupo de ligação de referência e não ao valor ou importância do marcador. Por exemplo, Figura 4, se no grupo de ligação do genoma simulado a ordem dos marcadores for: m_1-m_2-m_3-...-m_9, então a nota do marcador m_1 será 1, a do m_2 será 2, a do m_3 3.... E, se no grupo de ligação do genoma analisado a ordem dos respectivos marcadores for m_2-m_1-m_3-...-m_9, a nota do marcador m_2 continuará sendo 2, do m_1 será 1, do m_3 será 3 e até o ultimo marcador.

Portanto, os valores de Δ_k serão:

$$\Delta_1 = (1-2) = -1,\ \Delta_2 = (2-1) = 1,\ \Delta_3 = (3-3) = 0\ \ldots\ \Delta_3 = (9-9) = 0$$

Com isso o valor estimado da correlação de spearman é:

$$r_s = 1 - \frac{6\left[(-1)^2 + (1)^2 + (0)^2 + (0)^2 + (0)^2 + (0)^2 + (0)^2 + (0)^2 + (0)^2\right]}{9\left(9^2 - 1\right)} = 0{,}9833$$

Esta correlação não é afetada pelas distâncias entre marcadores.

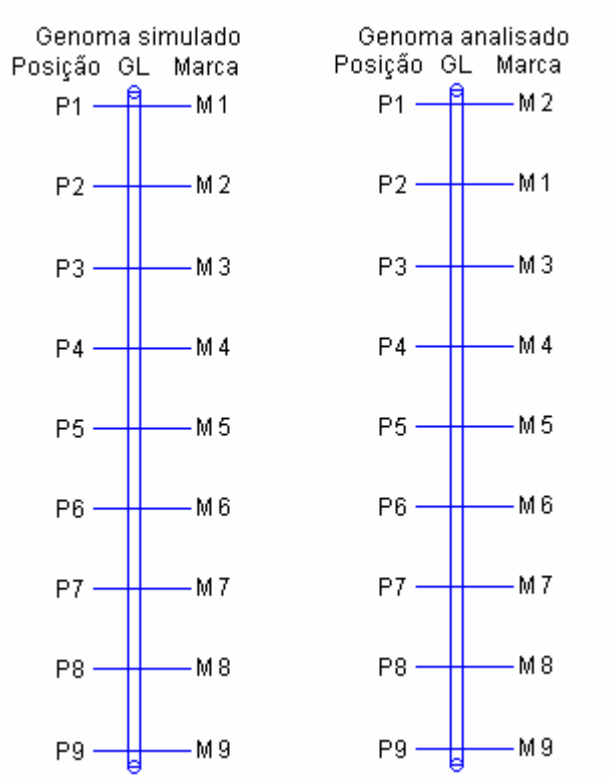

Figura 4. Grupos de ligação simulado e analisado com nove marcas e inversão das marcas M2 e M1, no grupo de ligação analisado.

3.3.6. Estresse

Para estimar os efeitos das mudanças nas distâncias entre os marcadores no genoma analisado em relação ao genoma originalmente simulado, foi utilizado o coeficiente de estresse.

O coeficiente de estresse (S) é utilizado como medida de adequação da representação gráfica de medidas de dissimilaridade convertidas em escores relativos às variáveis x e y em estudos de divergência genética (Cruz e Carneiro, 2003). A sua descrição para análise de genomas é demonstrada a seguir:

$$S = 100.\sqrt{\frac{\sum_{k=1}^{m-1}\left(d_{ok} - d_k\right)^2}{\sum_{k=1}^{m-1} d_{ok}^2}}$$

em que: S é o valor estimado do estresse, em percentagem, para o grupo de ligação do genoma analisado; d_{ok} é a distância entre marcas adjacentes m_k e m_{k+1} no grupo de ligação do genoma simulado; d_k é a distância entre marcas adjacentes m_k e m_{k+1} no grupo de ligação do genoma analisado (k=1, ..., m-1). Sendo que m é o número de marcadores no grupo de ligação do genoma simulado e no grupo de ligação analisado.

Se os marcadores no genoma analisado mantiverem as mesmas distâncias do genoma simulado, o valor estimado do estresse será zero.

3.4. Testes de comparação múltipla

As médias das variáveis como tamanho do grupo de ligação, distância média de marcas adjacentes, variância e estresse para cada grupo de ligação obtido para vários tamanhos de população foram comparados pelo teste de Tukey em nível de probabilidade de 5% (erro tipo I) com auxílio do aplicativo computacional GENES (Cruz, 2004). Também foram comparadas as médias gerais (médias de todos os grupos de ligação) para cada tamanho de população.

3.5. Fluxograma da simulação utilizada

Para facilitar o entendimento da logística utilizada no trabalho de simulação é apresentado o fluxograma abaixo (Figura 5). Os passos seguidos no processo de simulação foram:

1º) simulação dos genomas com nível de saturação de 10cM e 11 marcas moleculares;

2º) a partir dos genomas simulados foram construídos os genótipos dos genitores, sendo estes de dois tipos: completamente informativo e não informativos;

3º) Simulação das populações com diferentes números de indivíduos (100, 200, 400 e 600);

4º) as populações segregantes foram mapeadas;

5º) os mapas obtidos foram comparados com o genoma simulado. Para a comparação foram utilizados os critérios apresentados no quadro mostrado no interior do fluxograma.

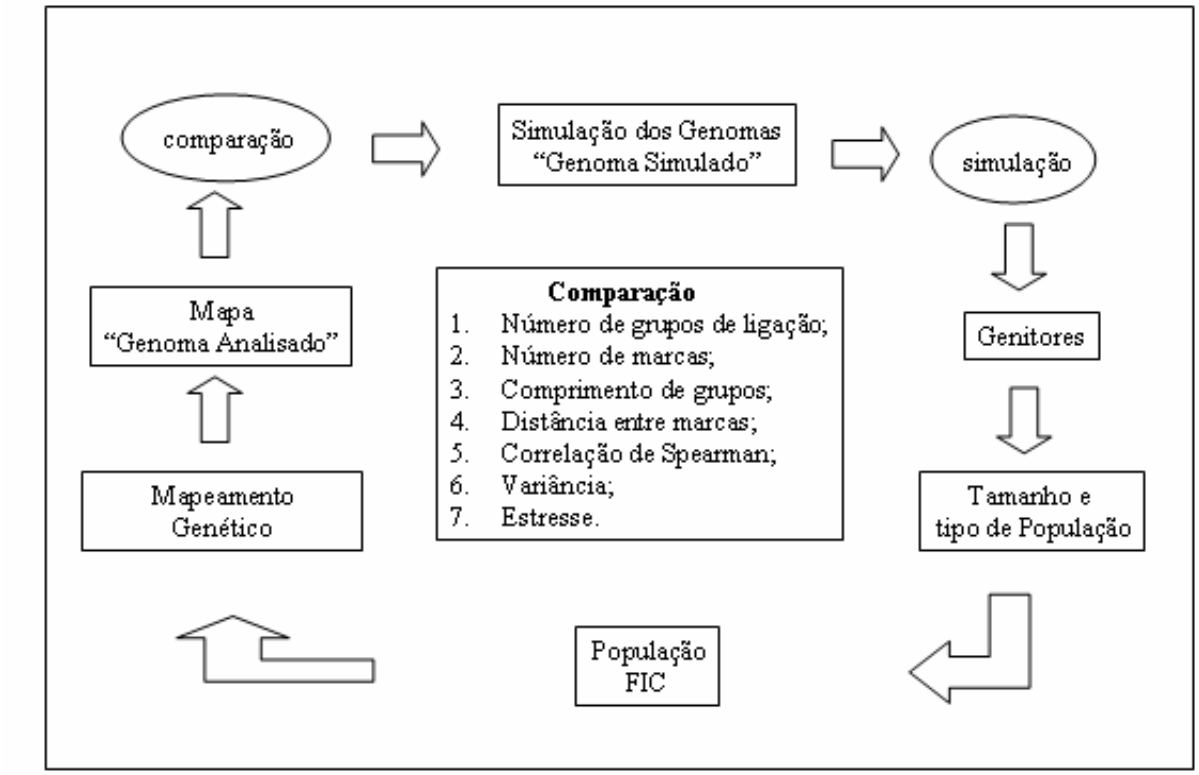

Figura 5. Fluxograma ilustrativo do processo de simulação utilizado neste trabalho.

4. Resultados e Discussão

4.1. População Completamente Informativa

Na população oriunda do cruzamento entre genitores completamente informativos não houve a recuperação dos três grupos de ligação (GL) originalmente estabelecidos considerando a saturação com 11 marcadores codominantes em uma repetição da população de 100 indivíduos. Com isso, esta repetição foi descartada do conjunto de dados, de forma que as análises apresentadas serão originadas das 99 repetições em que foi possível fazer a reconstituição dos 3 GL. Nesta repetição houve formação de 3 grupos de ligação, porém estes não foram formados conforme o esperado, ou seja, com 11 marcas por grupo de ligação. Assim, pode-se ter uma idéia preliminar do que acontece quando não se têm números de indivíduos e de marcas adequados, como ocorreu com apenas 100 indivíduos.

Um fator a desqualificar um determinado tamanho da população para análise é a junção de grupos de ligação. Essa junção pode ser total, em que um grupo inteiro se liga a outro grupo inteiro ou parcial, quando um grupo de ligação se liga à parte de outro grupo de ligação.

Na Figura 6 é apresentada uma situação em que um grupo de ligação se ligou a uma parte de outro grupo de ligação, não ocorrendo diminuição do número de grupos de ligação. Como pode ser visto, esse fenômeno resultou na formação de dois grupos de ligação, sendo um formado por um grupo de ligação inteiro e parte de outro grupo de ligação e o outro, por apenas fragmento de grupo de ligação, com quatro marcas.

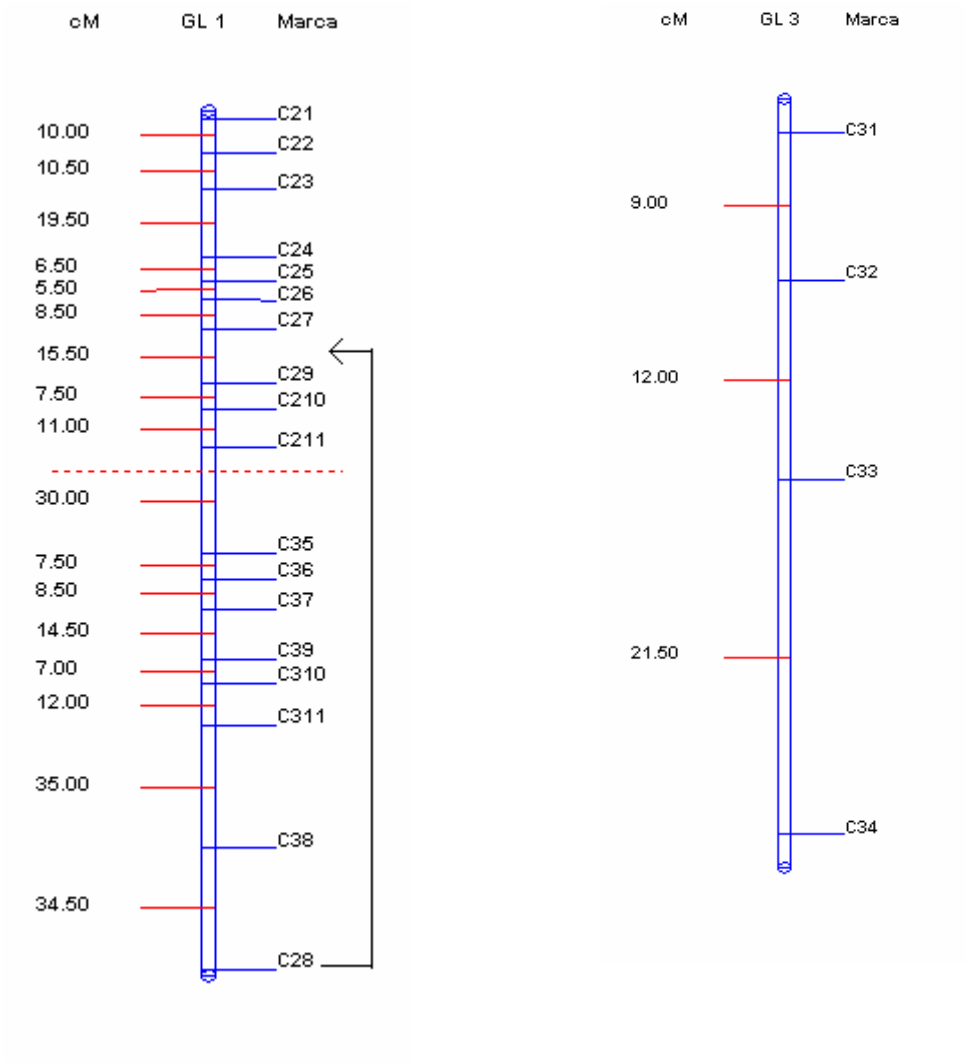

Figura 6. Representação de um grupo de ligação oriundo da junção de dois outros grupos de ligação. O trecho tracejado separa visualmente um grupo do outro. A seta indica onde a marca deveria estar se não tivesse ocorrido à inversão.

A inversão também é um fator importante a ser verificado para se ter uma idéia e confiabilidade dos dados obtidos. Na Figura 7, é apresentada um exemplo de inversão que aconteceu numa repetição desta população de 100 indivíduos.

As inversões podem se dar de várias formas, havendo casos em que o grupo de ligação é formado, mas com alterações da ordem de uma ou mais marcas. Além disso, as inversões podem acontecer, também, dentro de grupos de ligação já invertidos, dificultando a decisão de considerar para fins de análise este grupo de ligação.

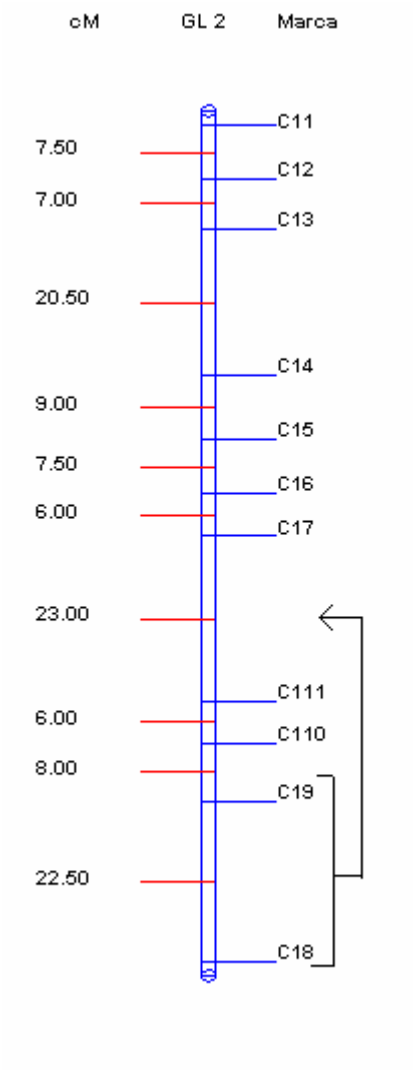

Figura 7. Representação da inversão de marcadores em um grupo de ligação já invertido. As setas indicam os pares de marcadores invertidos nos grupos de ligação.

Estes problemas, comumente encontrados na população de 100 indivíduos, são também verificados em escala muito menor em amostras com maior número de indivíduos e tende a diminuir à medida que o número de indivíduos de amostra aumentam. Nesta simulação considerando amostra de 200 indivíduos não ocorreu nenhuma repetição em que não conseguisse recuperar os 3 grupos de ligação esperados, e ocorreu apenas uma inversão de marcas nestas 100 repetições. Para as populações simuladas constituídas de 400 e 600 indivíduos nenhum destes problemas foi detectado.

4.1.1. Recuperação de grupos de ligação

O número de grupos de ligação esperado no processo de mapeamento das populações completamente informativas simuladas neste estudo seria de 3 grupos de ligação, que é o número de grupos de ligação que se tem no genoma original usado para a simulação das populações.

No Quadro 3 está apresentado o número de repetições que recuperou os 3 grupos de ligação (GL) para tamanhos de populações de FIC com 100, 200, 400 e 600 indivíduos, bem como o número de inversões para cada um destes tamanhos. Observa-se que apenas uma repetição da população com 100 indivíduos não recuperoiu os 3 grupos originalmente estabelecido no genoma simulado. Cabe ressaltar que nesta mesma repetição ocorreu inversão.

Verificou-se que a medida que o número de indivíduos aumenta o número de grupos de ligação recuperado que tende a ser igual ao número de grupos de ligação no genoma original utilizado para a simulação.

4.1.2. Correlação de Spearman entre medidas de distância

A correlação de Spearman foi utilizada para a identificação de inversão de posição de marcas dentro de cada um dos três grupos de ligação formados no mapeamento. A avaliação da inversão só foi feita em repetições em que foram recuperados os três grupos de ligação.

Por se tratarem de populações hipotéticas, foi considerado não ser necessário especificação das repetições e em quais grupos de ligação ocorreu inversão sendo relatados apenas o número de repetições a recuperar três grupos de ligação e quantos destes grupos apresentaram inversões, independentemente do número de inversões ocorridas dentro de cada repetição.

A obtenção de valores de correlação de Spearman iguais a 1 indicam que as ordens das marcas nos grupos de ligação obtidos no mapeamento das populações não foram alteradas em relação à ordem previamente conhecida do genoma original utilizado para a geração das simulações. Porém, caso os valores de correlação de Spearman fossem menores que 1 a indicação seria que a ordem das marcas nos grupos de ligação obtidos no

mapeamento da população segregante foi alterada em relação à ordem do genoma de referência.

Repetições com presença de valores de correlação de Spearman menor que 1, em um ou mais dos três grupos de ligação, foi obtido apenas na população de 200 indivíduos. Na população de 100 indivíduos a única repetição que teve a presença de inversões constatada foi a mesma repetição onde não conseguiu recuperar os três grupos de ligação, sendo portanto descartada da análise (Quadro 3).

Quadro 3. Tamanho da população, número de repetições com três grupos de ligação e número de inversões na população completamente informativa

Tamanho da População	Repetições	Número de Inversões
100	99	0
200	100	1
400	100	0
600	100	0

É importante observar que, com o aumento do tamanho das populações, os problemas foram só diminuindo, deixando inicialmente de não existir repetições que não conseguissem resgatar os três grupos de ligações iniciais, e posteriormente não existindo nem repetições em que verificasse a presença de inversões.

4.1.3. Comprimento dos grupos de ligação

Com relação ao comprimento médio de cada grupo de ligação, foi obtida a média aritmética dos comprimentos dos grupos de ligação apenas nas repetições que recuperaram 3 grupos de ligação. Os valores dos comprimentos médios são apresentados no Quadro 4.

O comprimento esperado nos grupos de ligação após o mapeamento das populações completamente informativas era de 100 cM, uma vez que esse é o comprimento de cada grupo de ligação no nível de saturação do genoma utilizado para a simulação das populações.

Na avaliação dos comprimentos médios dos grupos de ligação foi feita a análise de variância seguida do teste comparativo de médias (teste de Tukey), em que as diferenças entre médias foram avaliadas com nível de significância de 0,05. O teste de Tukey permitiu a avaliação estatística do efeito do tamanho da população no comprimento médio dos grupos de ligação formados no mapeamento das populações simuladas.

Quadro 4. Média aritmética do comprimento, em cM, de três grupos de ligação em quatro tamanhos de população completamente informativas.

Tamanho da População	Grupos de Ligação			Média Geral	Números de dados
	1	2	3		
100	102,722[(a)1]	102,152[(a)]	104,172[(a)]	103,015[(a)]	99
200	101,891[(a)]	103,244[(a)]	103,458[(a)]	102,864[(a)]	100
400	102,825[(a)]	103,025[(a)]	103,829[(a)]	103,226[(a)]	100
600	102,967[(a)]	102,586[(a)]	102,788[(a)]	102,780[(a)]	100

[1] indica que dentro dos parêntesis estão as letras correspondentes ao resultado da análise de médias feita pelo teste de Tukey a 5%. Nas colunas, médias seguidas pela mesma letra não diferem estatisticamente entre si, pelo teste de Tukey.

No Quadro 4 são apresentados os resultados do teste de Tukey em nível de significância de 0,05. Verificou-se que não houve diferenças entre as médias dos diferentes tamanhos de população. Como pode ser visto na penúltima coluna do Quadro 4, não houve diferença entre as médias gerais das quatro populações avaliadas. Desde modo pode-se concluir que não houve variação significativa no comprimento médio dos grupos de ligação com o aumento no tamanho da população.

Uma das maneiras adicionais de observar o comportamento dos comprimentos médios dos grupos de ligação se dá através da análise do desvio-padrão. Com o aumento no número de indivíduos em um mesmo nível de saturação do genoma espera-se que o desvio-padrão diminua, fato este, que pode ser observado no Quadro 5, observando-se os valores de desvio-padrão na penúltima coluna. Desta forma, torna-se evidente a tendência de redução na amplitude de variação observada nas médias com o aumento do tamanho populacional.

Quadro 5. Desvio-padrão do comprimento de três grupos de ligação em quatro tamanhos de populações completamente informativas.

Tamanho da População	Grupos de Ligação			Média Geral	Números de dados
	1	2	3		
100	6,895	7,193	7,356	7,148	99
200	5,007	6,097	5,741	5,615	100
400	3,663	3,741	3,683	3,696	100
600	3,026	3,060	3,229	3,105	100

Assim, com o aumento no número de indivíduos das populações não existe diferença significativa entre as médias gerais do comprimento dos grupos de ligação, porém há uma diminuição na amplitude do desvio-padrão com o aumento no número de indivíduos nas populações.

Apesar das médias do comprimento dos grupos de ligação entre os diferentes tamanhos de população não terem se mostrado significativas, observou-se que a população constituída de 600 indivíduos obteve o menor valor de média (102,780), aproximando-se do comprimento originalmente estabelecido de 100 cM. Fato semelhante pode ser observado para o desvio padrão desta população (3,105), que é o menor entre os quatro, e ainda, é menos da metade do valor obtido na população de 100 indivíduos (7,148).

4.1.4. Média das distâncias entre marcas adjacentes

A média das distâncias entre marcas adjacentes ao longo de cada grupo de ligação foi obtida fazendo-se duas médias aritméticas sucessivas. A primeira média foi feita apenas entre os valores de distância encontrados dentro de cada grupo de ligação. A segunda foi obtida pela média aritmética das médias anteriormente obtidas. Neste processo só foram utilizadas repetições que tinham os três grupos de ligação reconstituídos, as demais foram desconsideradas para efeito de análise.

Os valores de média das distâncias entre marcas adjacentes e seu desvio padrão obtido para os vários tamanhos de população segregantes utilizados no mapeamento genético estão apresentados nos Quadros 6 e 7, respectivamente.

Quadro 6. Média das distâncias, em cM, entre marcas adjacentes nas repetições que recuperaram três grupos de ligação, em quatro tamanhos de população completamente informativas.

Tamanho da População	Grupos de Ligação			Média Geral	Números de dados
	1	2	3		
100	10,272$^{(a)1}$	10,215$^{(a)}$	10,417$^{(a)}$	10,301$^{(a)}$	99
200	10,189$^{(a)}$	10,324$^{(a)}$	10,345$^{(a)}$	10,286$^{(a)}$	100
400	10,282$^{(a)}$	10,302$^{(a)}$	10,382$^{(a)}$	10,322$^{(a)}$	100
600	10,296$^{(a)}$	10,258$^{(a)}$	10,278$^{(a)}$	10,278$^{(a)}$	100

[1] indica que dentro dos parêntesis estão as letras correspondentes ao resultado da análise de médias feita pelo teste de Tukey a 5%. Nas colunas, médias seguidas pela mesma letra não diferem estatisticamente entre si, pelo teste de Tukey

Quadro 7. Desvio-padrão das médias das distâncias entre marcas adjacentes nas repetições que recuperaram três grupos de ligação em quatro tamanhos de populações completamente informativas.

Tamanho da População	Grupos de Ligação			Média Geral	Números de dados
	1	2	3		
100	0,689	0,719	0,735	0,714	99
200	0,500	0,609	0,574	0,561	100
400	0,366	0,374	0,368	0,369	100
600	0,302	0,306	0,322	0,310	100

Como as populações foram simuladas a partir de um genoma pré-determinado, a média das distâncias entre marcas adjacentes esperada já é conhecida. Sendo assim espera-se que as médias se aproximem ao máximo da distância do genoma simulado que era de 10 cM. Como pode ser observado no Quadro 6 , independente do tamanho da população todos os valores ficaram acima dos inicialmente esperados, porém em relação ao teste de Tukey, realizado com $P<0,05$, não foi encontrada diferença entre as médias para os tamanhos populacionais avaliados.

A falta de tendência à diminuição no tamanho da distância entre marcas adjacentes pode ser claramente vista, por exemplo, no grupo de ligação 2 em que na população com

100 indivíduos a distância entre marcas é de 10,215 cM, enquanto na população com 600 indivíduos a distância entre médias é de 10,258 cM; como visto, ocorre pequeno aumento nos valores, mesmo que esta variação não seja significativa esperava-se uma variação em sentido contrário, ou seja, quanto maior o tamanho da população espera-se que as distâncias médias entre marcas adjacentes diminuíssem. Como ainda pode ser observado no Quadro 6 apenas no grupo de ligação três a distância média entre marcas adjacentes foi menor na população de 600 indivíduos, mostrando claramente esta falta de tendência nas estimativas desta distância.

Embora a redução nos valores das médias entre marcas adjacentes não tenha ocorrido de forma evidente, o desvio-padrão apresenta comportamento diferente. No Quadro 7 estão os desvios das médias das distâncias entre marcas adjacentes nas repetições que recuperaram três grupos de ligação, avaliados em quatro tamanhos populacionais diferentes. É interessante observar que, com o aumento do tamanho da população, existe a tendência clara à redução da amplitude de variação nos valores das médias das distâncias entre marcas adjacentes. Esta diminuição pode ser observada tanto nos valores pertencentes ao mesmo grupo de ligação, como também, através da média geral de cada tamanho populacional. Veja que para o tamanho de população de 100cM o desvio era de 0,714 enquanto para o tamanho de 600 indivíduos o valor do desvio diminui para 0,310.

Por não apresentarem nenhuma diferença significativa, as médias das distâncias entre marcas deveriam exibir valores de desvio-padrão com comportamento semelhante, mas pelo que foi demonstrado, apesar de o tamanho da população não influenciar no tamanho das médias entre marcas adjacentes, existe uma influência na diminuição na variação da amplitude do tamanho das médias com o aumento das populações avaliadas.

Segundo Hospital et al. (1992), marcadores são mais úteis quando sua posição no mapa é conhecida. Com isso, tem-se que, quanto menor a variância e, conseqüentemente, o desvio das distâncias entre marcas, melhor será a capacidade de localização e, por conseqüência, mais eficiente o trabalho de mapeamento.

4.1.5. Variância das distâncias entre marcas adjacentes

A partir das distâncias entre marcas adjacentes obtidas nos grupos de ligação foi estimada a variância amostral, como apresentado no Quadro 8. Como em quadros anteriores, para cada grupo de ligação são dadas médias aritméticas, nesse caso, das

variâncias obtidas em cada repetição, em que houve a formação de três grupos de ligação no mapeamento das populações completamente informativas simuladas.

Os valores de variância encontrados na análise dos mapas obtidos das populações simuladas são referentes aos erros para qualquer tamanho de população dos genomas avaliados, pois os genomas utilizados para geração das populações segregantes tinham seus marcadores distribuídos de forma eqüidistante dentro dos três grupos de ligação, sendo que o genoma possuía nível de saturação de 11 marcas por grupo de ligação os marcadores encontravam-se distribuídos a distâncias eqüidistante de 10 cM ao longo dos grupos de ligação.

Os valores de variância podem ser interpretados de forma que quanto menores os valores de variância mais eqüidistantes estarão distribuídas às marcas dentro dos grupos de ligação e conseqüentemente menor o erro. Portanto quanto menores os valores de variância mais próximos estarão os valores do esperado, indicando uma boa recuperação do genoma com o mapeamento das populações segregantes.

Quadro 8. Variância das distâncias entre marcas adjacentes nas repetições que recuperaram três grupos de ligação em quatro tamanhos de populações completamente informativas.

Tamanho da População	Grupos de Ligação			Média Geral	Números de dados
	1	2	3		
100	5,327[(a)]	5,707[(a)]	5,462[(a)]	5,498[(a)]	99
200	2,797[(b)]	3,017[(b)]	2,736[(b)]	2,85[(b)]	100
400	1,245[(c)]	1,236[(c)]	1,259[(c)]	1,246[(c)]	100
600	0,838[(c)]	0,875[(c)]	0,867[(c)]	0,86[(d)]	100

[1] indica que dentro dos parêntesis estão letras correspondentes ao resultado da análise de médias, pelo teste de Tukey ao nível de significância de 5%. Nas colunas, médias seguidas de uma mesma letra não diferem entre si, pelo teste de Tukey.

Analisando o efeito do tamanho da população, verificou-se redução da média da variância em relação às distâncias entre marcas adjacentes. Na análise da média geral houve diferença significativa entre todas as médias dos diferentes tamanhos de população.

A redução nos valores de variância com o aumento do tamanho da população é evidenciada pelas diferenças significativas dadas pelo teste de médias, em que as menores médias estão associadas às maiores populações, portanto, levando à maior precisão no mapeamento genético.

Em relação à confiabilidade, quanto menores os valores de variância dos valores entre marcas adjacentes, maior a confiabilidade dos dados, sendo assim, com o aumento do tamanho populacional temos maior confiabilidade dos dados.

Além da redução na média das variâncias, observa-se redução na amplitude de variação dos valores de variância entre repetições, à medida que aumentou o tamanho da população. Essa redução pode ser demonstrada pela redução do desvio-padrão. Como exemplo, pode-se observar os desvios nas populações de 100 e 600 indivíduos, onde na primeira tem-se um desvio padrão de 4,324 e na segunda uma estimativa de 0,459 (Quadro 9).

Quadro 9. Desvio-padrão das médias das variâncias das distâncias entre marcas adjacentes nas repetições que recuperaram três grupos de ligação em quatro tamanhos de populações completamente informativas.

Tamanho da População	Grupos de Ligação			Média Geral	Números de dados
	1	2	3		
100	4,774	4,645	3,553	4,324	99
200	2,111	1,899	1,636	1,882	100
400	0,653	0,723	0,886	0,754	100
600	0,442	0,436	0,499	0,459	100

Observa-se portanto, que para população completamente informativa, o aumento do tamanho da população acarreta uma redução da variância das médias entre marcas adjacentes, e também redução no desvio-padrão das médias das variâncias entre as marcas adjacentes, como foi verificado anteriormente nos Quadros 8 e 9.

4.1.6. Estresse

O estresse foi utilizado para expressar o grau de concordância dos valores de distância entre cada par de marcas adjacentes nos grupos de ligação simulados em relação às distâncias nos respectivos pares de marcas no genoma de referência.

Para obtenção do estresse médio foi feita uma média aritmética dos valores de estresse obtidos nas repetições em que houve a formação de três grupos de ligação no

mapeamento genético. Os valores obtidos são apresentados a seguir, no Quadro 10, juntamente com o resultado de comparações (Tukey a 5%) dos valores médios do estress.

Quadro 10. Valores do estresse médio (%) em função do tamanho da população em populações completamente informativas.

Tamanho da População	Grupos de Ligação			Média Geral	Números de dados
	1	2	3		
100	21,942[(a)]	22,852[(a)]	22,712[(a)]	22,502[(a)]	99
200	15,931[(b)]	17,256[(b)]	16,450[(b)]	16,546[(b)]	100
400	11,187[(c)]	11,152[(c)]	11,431[(c)]	11,257[(c)]	100
600	9,377[(d)]	9,412[(d)]	9,511[(d)]	9,434[(d)]	100

[1] indica que dentro dos parêntesis estão letras correspondentes ao resultado da análise de médias, pelo teste de Tukey ao nível de significância de 5%. Nas colunas, médias seguidas de uma mesma letra não diferem entre si, pelo teste de Tukey.

Observa-se claramente uma tendência de redução dos valores de estresse médio à medida que o tamanho da população avaliada aumenta. Esta tendência ocorreu tanto para cada grupo de ligação avaliado individualmente quanto para as médias gerais de estresse, apresentada na penúltima coluna do Quadro 10, em que valores de estresse médio passaram de 22,502 para 9,434 %, considerando-se as populações de 100 e 600 indivíduos, respectivamente.

Pelo teste de Tukey verificou-se que os valores de estresse médio diferiram estatisticamente para os diferentes tamanhos de população utilizadas.

Em relação à amplitude da variação dos valores médios de estresse, percebe-se que ocorreu uma diminuição na amplitude com o aumento do tamanho das populações avaliadas (Quadro 11).

Quadro 11. Desvio-padrão dos valores de estresse médio nas repetições em que recuperaram três grupos de ligação em quatro tamanhos de populações completamente informativas.

Tamanho da População	Grupos de Ligação			Média Geral	Números de dados
	1	2	3		
100	7,279	6,914	6,873	7,022	99
200	4,893	4,619	4,525	4,679	100
400	2,874	3,160	3,303	3,112	100
600	2,341	2,491	2,393	2,408	100

Conclui-se com base nos dados apresentados nos Quadros 10 e 11, que com o aumento do tamanho das populações houve uma tendência de diminuição no estresse médio e também na amplitude de variação dos valores médios de estresse, considerando populações completamente informativas.

4.2. População Não - Completamente Informativa

Nas populações oriundas do cruzamento de genitores não completamente informativos não houve a recuperação dos 3 grupos de ligação esperados considerando a saturação com 11 marcadores co-dominantes nas populações de 100 e 200 indivíduos. Com isso as repetições onde não foi capaz de recuperar os três grupos de ligação foram descartadas dos dados de forma que as análises apresentadas posteriormente serão originadas apenas das repetições onde conseguiu fazer a reconstituição dos 3 GL.

Um fato que chama atenção nos dados originados, é que o marcador 5 do grupo de ligação 3 teve genótipo A_1A_1 nos dois pais. Desta forma este não segrega, portanto, não foi utilizado para análise de mapeamento (Figura 8). Assim, o grupo de ligação 3 para a população não completamente informativa terá apenas 10 marcas, sendo portanto descartado das análises realizadas posteriormente. Porém a fim de mostrar as informações deste grupo de ligação, sempre que possível os dados serão apresentados de forma resumida.

Como já citado anteriormente, um fator a desqualificar as populações para análise é a junção de grupos de ligação, podendo esta junção ser total ou parcial. Na Figura 8 é

apresentada uma situação em que um grupo de ligação se ligou a outro grupo de ligação, ocorrendo assim a redução do número de grupos de ligação. Como pode ser visto, esse fenômeno resultou na formação de um único grupo de ligação, constituídos das marcas dos dois grupos de ligação.

Outro fator importante a ser observado na verificação da qualidade dos dados é a ocorrência de inversões. Na Figura 8 também é apresentada um exemplo de inversão, onde as setas pretas indicam os pares de marcadores invertidos nos grupos de ligação.

Diferentes formas de inversões podem ser observadas, sendo uma delas quando ocorre a translocação de uma ou mais marcas dentro do grupo de ligação formado.

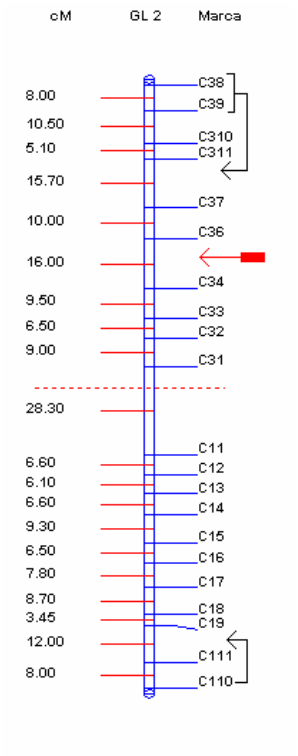

Figura 8: Representação da ligação entre um grupo de ligação a outro grupo de ligação. O trecho tracejado separa visualmente um grupo do outro. As setas pretas indicam onde as marcas deveriam estar se não tivesse ocorrido à inversão. A seta (⟵) chama a atenção para a inexistência do marcador C35.

Na Figura 9 pode ser verificado um outro caso de ligação entre grupos de ligação. Neste caso houve a ligação total dos grupos de ligação 2 e 3 e ainda a ligação parcial do grupo de ligação 1. Desta forma que no final apareceu apenas dois grupos de ligação da população como se observa na figura

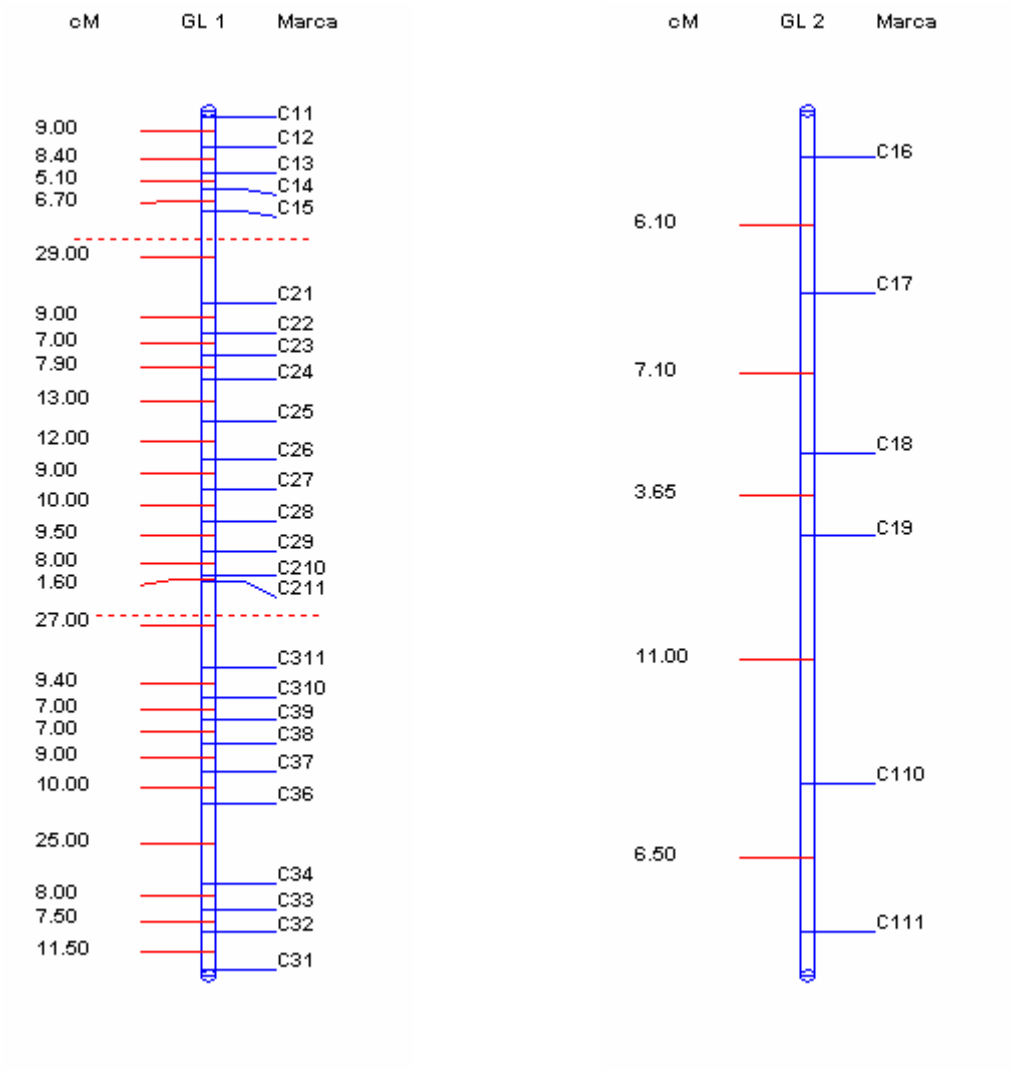

Figura 9. Representação da ligação total entre dois grupos de ligação e parte de um terceiro grupo de ligação, e os 6 marcadores restantes formando um segundo grupo de ligação.

Na Figura 10 pode-se verificar a formação de quatro grupos de ligação. Desta forma não foi possível recuperar os três grupos de ligação originalmente informados no genoma simulado. Nota-se também que o último mapa que corresponderia ao grupo de ligação 3 também não possui a marca C_{35}, como já explicado anteriormente. Veja ainda, que o grupo de ligação dois se desmembrou em dois novos grupos de ligação (segundo e terceiros mapas de ligação na Figura 10), o primeiro constituído de nove marcadores e o segundo constituído de dois marcadores. Nesta figura ainda chama-se a atenção para inversões acontecidas entre os marcadores, que estão assinaladas por setas pretas.

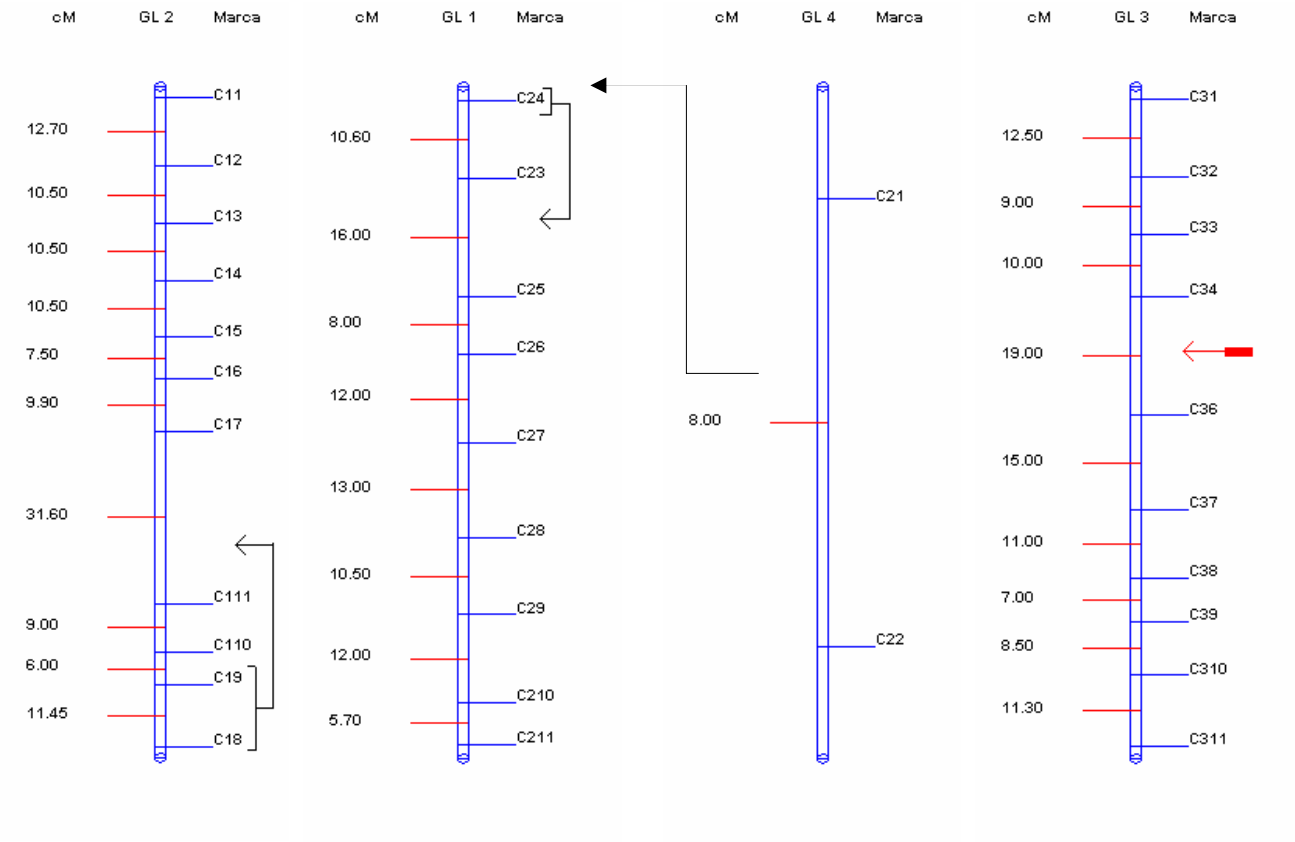

Figura 10. Representação da formação de quatro grupos de ligação após ocorrer a quebra de um dos grupos de ligação, ocorrência de inversões e a falta de marca C35.

Estes problemas encontrados tendem a aparecerem em freqüência muito menor à medida que o número de indivíduos da amostra aumenta. Fatos estes que serão evidenciados posteriormente neste trabalho.

4.2.1. Recuperação de grupos de ligação

O número de grupos de ligação esperado no processo de mapeamento das populações não completamente informativas simuladas seria de 3 grupos de ligação, que é o número de grupos de ligação que se tem no genoma original usado para a simulação das populações. Apesar disso, o número de grupos de ligação não foi recuperado na população de 100 e 200 indivíduos. Sendo assim, a medida que ocorre o aumento do número de indivíduos o número de grupos de ligação recuperado tende a ser igual ao número de grupos de ligação no genoma original utilizado para a simulação.

4.2.2. Correlação de Spearman entre medidas de distância

Para que fosse possível identificar a ocorrência de inversões das marcas dentro de cada grupo de ligação formado utilizou-se a correlação de Spearman. Apenas nas repetições em que foram recuperados os três grupos de ligação foi realizada a avaliação da inversão.

Nesta avaliação foi relatado apenas o número de repetições a recuperar três grupos de ligação e quantos destes grupos apresentaram inversões, independentemente do número de inversões ocorridas dentro de cada repetição.

Valores de correlação de Spearman iguais a 1 significa que as ordens das marcas nos grupos de ligação obtidos no mapeamento das populações não foram alteradas em relação à ordem previamente conhecida, do genoma original, utilizado para a geração das simulações. Valores de correlação de Spearman menores que 1 indicam que a ordem das marcas nos grupos de ligação obtidos no mapeamento da população segregante foi alterada em relação à ordem do genoma de referência.

Observa-se que em todos os tamanhos de população foram observados a presença de valores de correlação de Spearman menores que 1 para as repetições avaliadas, uma vez que em todos tamanhos populacionais foi observado a ocorrência de inversões (Quadro 12).

Quadro 12. Tamanho da população, número de repetições com três grupos de ligação e número de inversões detectadas em populações não completamente informativas.

Tamanho da População	Repetições	Número de Inversões	% de Inversões
100	95	32	11,22
200	96	23	7,98
400	100	23	7,66
600	100	7	2,33

A porcentagem de inversões ocorridas tenderam a diminuir com o aumento do tamanho das populações. Na população composta de 100 indivíduos 95 repetições das 100 realizadas conseguiram resgatar os três grupos de ligação originalmente definidos como padrão. Nesta mesma população 32 grupos de ligação apresentaram algum tipo de inversão, fazendo com que a correlação de spearman fosse inferior a 1, ou seja, dos 285 grupos de ligação possíveis de serem analisados, 11,22% apresentaram algum tipo de inversão. A população de maior tamanho (600 indivíduos) não apresentou nenhuma repetição com problema em resgatar os três grupos de ligação. Além disso, apenas 7 grupos de ligação,

dos 300 analisados apresentaram algum tipo de inversão, o que corresponde a 2,33% dos grupos de ligação analisados, valor este aproximadamente 5 vezes menor do que o apresentado pela população de 100 indivíduos.

Desta forma pode-se concluir que ao analisar a porcentagem de inversões observadas que existe uma tendência de diminuição da quantidade de inversões com o aumento do tamanho da população.

4.2.3. Comprimento dos grupos de ligação

Para avaliação do comprimento dos grupos de ligação, foi calculada a média aritmética dos comprimentos dos grupos de ligação obtidos nas repetições com três grupos de ligação. Os valores dos comprimentos médios de cada grupo de ligação são apresentados no Quadro 13.

O comprimento esperado nos grupos de ligação após o mapeamento era de 100 cM, uma vez que esse é o comprimento de cada grupo de ligação no nível de saturação do genoma utilizado para a simulação das populações. Sendo assim, quanto mais próximos de 100 cM melhores serão os dados obtidos.

Após a obtenção dos dados, realizou-se a análise de variância seguida por teste comparativo de médias (teste de Tukey) em nível de significância de 5%. O teste de Tukey permitiu a avaliação estatística do efeito do tamanho populacional nos grupos de ligação formados no mapeamento das populações simuladas.

Quadro 13. Média aritmética do comprimento, em cM, de três grupos de ligação em quatro tamanhos de população não completamente informativas.

Tamanho da População	Grupos de Ligação			Média Geral	Números de dados
	1	2	3		
100	105,081$^{(ab)}$	102,276$^{(a)}$	103,165$^{(a)}$	103,507$^{(a)\,1}$	95
200	105,192$^{(ab)}$	103,640$^{(a)}$	101,169$^{(ab)}$	103,333$^{(a)}$	96
400	105,455$^{(a)}$	103,546$^{(a)}$	101,548$^{(ab)}$	103,516$^{(a)}$	100
600	102,099$^{(b)}$	103,056$^{(a)}$	100,664$^{(b)}$	101,939$^{(a)}$	100

[1] indica que dentro dos parêntesis estão as letras correspondentes ao resultado da análise de médias feita pelo teste de Tukey a 5%. Nas colunas, médias seguidas pela mesma letra não diferem estatisticamente entre si, pelo teste de Tukey

Com base no teste de Tukey pode-se observar que apenas para o grupo de ligação 2 não apresentou diferenças entre as médias dos diferentes tamanhos de população. Para os grupos de ligação 1 e 3 houve diferença significativa entre os quatros tamanhos de população. Como pode ser visto na penúltima coluna do Quadro 13, não houve diferença entre as médias gerais das quatro populações avaliadas. Desde modo podemos concluir que não houve variação significativa no comprimento médio dos grupos de ligação com o aumento no tamanho da população.

Apesar de se esperar uma tendência de diminuição do comprimento médio do grupo de ligação com o aumento do tamanho populacional, este fato não ficou evidenciado nos dados apresentados no Quadro 13, uma vez que, não aconteceu sempre esta diminuição da estimativa do comprimento dos grupos de ligação ao aumentar o tamanho populacional.

Uma análise do desvio padrão dos comprimentos médios dos grupos de ligação é uma alternativa auxiliar para observar o comportamento e tendência destes comprimentos, uma vez que, com o aumento no número de indivíduos em um mesmo nível de saturação do genoma espera-se que o desvio-padrão diminua, fato este, que não pôde ser bem evidenciado com os dados do Quadro 14, uma vez que observando os valores de desvio-padrão médio na penúltima coluna, as populações de 200 e 400 indivíduos discordaram desta tendência, uma vez que se esperava para a população de 200 indivíduos (6,792) uma estimativa de desvio-padrão médio superior à estimativa apresentada para a população de 400 indivíduos (6,978). Sendo assim, não se tornou evidente a tendência de redução na amplitude de variação observada nas médias com o aumento do tamanho populacional.

Quadro 14. Desvio-padrão do comprimento de três grupos de ligação em quatro tamanhos de populações não completamente informativas.

Tamanho da População	Grupos de Ligação			Média Geral	Números de dados
	1	2	3		
100	11,007	9,574	7,943	9,508	95
200	8,927	6,289	5,162	6,792	96
400	9,831	5,870	5,233	6,978	100
600	5,781	4,160	3,238	4,393	100

Após a interpretação dos dados apresentados anteriormente, não ficou evidente e não foi possível concluir que, com o aumento no número de indivíduos das populações

haverá uma diminuição no valor das estimativas do comprimento médio dos grupos de ligação, e também não conseguiu evidenciar que com este aumento populacional tem-se uma diminuição na amplitude do desvio-padrão.

Mesmo que não tenha sido comprovada estatisticamente a diferença entre as médias e a tendência de diminuição do desvio padrão, pode-se observar que para a população constituída de 600 indivíduos foi onde teve o menor valor de média (101,939), aproximando-se do comprimento originalmente estabelecido de 100 cM; o desvio padrão da população de 600 indivíduos também foi o menor dos valores observados (4,393), que é menos da metade do valor obtido na população de 100 indivíduos (9,508).

4.2.4. Média das distâncias entre marcas adjacentes

Para a obtenção da média das distâncias entre marcas adjacentes ao longo de cada grupo de ligação foram realizadas duas médias aritméticas sucessivas. A primeira entre os valores de distância encontrados dentro de cada grupo de ligação e a segunda obtida através da média aritmética das médias anteriormente obtidas. Deve-se ressaltar que para a obtenção destes valores médios utilizaram-se apenas repetições que tiveram os três grupos de ligações reconstituídos, as demais foram desconsideradas para efeito de análise. Quanto mais estes valores médios de distâncias entre marcas adjacentes aproximarem de 10 cM, melhor será os dados, uma vez que no genoma paramétrico usado, tinha-se marcas espaçadas 10 cM.

Deve-se neste momento chamar a atenção para o fato que o grupo de ligação 3 só possui 10 marcas no seu genoma, e não 11 como ocorrem nos outros dois grupos de ligação. Este fato já havia sido realçado anteriormente, e se deve ao fato de que a marca C35 é de constituição genotípica A_1A_1 x A_1A_1. Desta forma ela não foi utilizada para prover as estimativas de distância no mapa. Sendo assim, nas análises mostradas a seguir as discussões serão sempre feitas chamando a atenção para o fato de usar ou não os valores deste terceiro grupo de ligação.

Nos Quadros 15 e 16 a seguir têm-se, respectivamente, os valores de média das distâncias entre marcas adjacentes e seu desvio padrão obtido para os vários tamanhos de população analisadas.

Quadro 15. Média das distâncias, em cM, entre marcas adjacentes nas repetições que recuperaram três grupos de ligação, em quatro tamanhos de população não completamente informativas.

Tamanho da População	Grupos de Ligação (GL)			Média Geral	Média GL_1 e GL_2	Números de dados
	1	2	3			
100	$10,508^{(ab)}$	$10,227^{(a)}$	$11,462^{(a)}$	$10,732^{(a)}$	$10,367^{(a)}$	95
200	$10,519^{(ab)}$	$10,363^{(a)}$	$11,240^{(ab)}$	$10,707^{(a)}$	$10,441^{(a)}$	96
400	$10,545^{(a)}$	$10,354^{(a)}$	$11,283^{(ab)}$	$10,727^{(a)}$	$10,449^{(a)}$	100
600	$10,209^{(b)}$	$10,305^{(a)}$	$11,184^{(b)}$	$10,566^{(a)}$	$10,257^{(a)}$	100

[1] indica que dentro dos parêntesis estão as letras correspondentes ao resultado da análise de médias feita pelo teste de Tukey a 5%. Nas colunas, médias seguidas pela mesma letra não diferem estatisticamente entre si, pelo teste de Tukey

Quadro 16. Desvio-padrão das médias das distâncias entre marcas adjacentes nas repetições que recuperaram três grupos de ligação em quatro tamanhos de populações não completamente informativas.

Tamanho da População	Grupos de Ligação (GL)			Média Geral	Média GL_1 e GL_2	Números de dados
	1	2	3			
100	1,100	0,957	0,882	0,979	1,028	95
200	0,892	0,628	0,573	0,697	0,76	96
400	0,983	0,587	0,581	0,717	0,785	100
600	0,578	0,416	0,359	0,451	0,497	100

Ao utilizar um genoma pré-determinado, para que de posse dele fosse realizado as simulações, já se tem a priori o conhecimento do valor esperado para a média das distâncias entre marcas adjacentes. Espera-se então que as médias se aproximem ao máximo da distância do genoma simulado que era de 10cM, válido para o grupo de ligação 1 e 2, constituídos de 11 marcas. Já para o grupo de ligação 3 considerando a presença de apenas 10 marcas, espera-se uma distância média de 11,11 cM ao longo do genoma. Independente do tamanho da população todos os valores ficaram acima dos inicialmente esperados (Quadro 15).

Com base no teste de Tukey, não foi encontrada diferença significativa ($P< 0,05$) entre as médias para os tamanhos populacionais avaliados no grupo de ligação 2. Para os

grupos de ligação 1 e 3 o resultado foi semelhante, onde a população de 600 indivíduos apresentou a menor estimativa de distância, apesar de que estatisticamente este valor não diferiu dos valores obtidos por outros tamanhos populacionais.

Com a avaliação de cada grupo de ligação separadamente e também para os valores médios obtidos, não se conseguiu observar claramente uma tendência de diminuição no tamanho da distância entre marcas adjacentes. Para o grupo de ligação 1 este fato se torna mais evidente, uma vez que se espera que com o aumento do tamanho da população ocorra uma diminuição do tamanho médio das distâncias adjacentes, ao contrário disto, observa-se que da população de 100 indivíduos (10,508) para a população de 400 indivíduos (10,545) ocorre um aumento do valor da distância média, porém estes valores não são estatisticamente diferentes.

O comportamento da média geral, tanto considerando a avaliação dos 3 grupos de ligação, ou apenas considerando os valores do grupo de ligação 1 e 2, se apresenta de modo semelhante, uma vez que as estimativas obtidas para os diferentes tamanhos populacionais não foram estatisticamente diferentes entre si pelo teste Tukey a 5% de probabilidade.

De forma semelhante são os resultados apresentados com a análise baseada no desvio-padrão, uma vez que com o aumento do tamanho populacional não ocorreu em todos os grupos de ligação a tendência de diminuição das estimativas do desvio-padrão. Esta diminuição só pode ser observada nos valores pertencentes grupo de ligação 2. Para os outros dois grupos de ligação, como também considerando a média dos três grupos de ligação, ou apenas a média dos grupos de ligação 1 e 2, a tendência esperada de diminuição da amplitude de variação não foi claramente observada. Apesar de que, no tamanho populacional de 600 indivíduos a amplitude de variação se mostrou muito inferior à amplitude de variação na população de 100 indivíduos. Porém, mais uma vez deve-se enfatizar que as populações de 200 e 400 indivíduos não tiveram esta tendência de diminuição da variação (Quadro 16).

Desta forma, podemos concluir que para populações não completamente informativas e com base nos dados apresentados anteriormente, com o aumento do tamanho populacional não se pode concluir que haverá uma tendência clara de diminuição da estimativa da distância média entre marcas adjacentes, como também não se evidencia uma diminuição da amplitude de variação destas estimativas de distâncias médias.

4.2.5. Variância das distâncias entre marcas adjacentes

A estimativa de variância amostral (Quadro 17) foi obtida com base nas distâncias entre marcas adjacentes dos diferentes grupos de ligação. Como já apresentado anteriormente, para cada grupo de ligação são obtidas as médias aritméticas das variâncias obtidas em cada repetição em que se resgataram os três grupos de ligação.

Os valores de variância são referentes aos erros observados nos diferentes tamanhos de populações avaliados, uma vez que, os genomas utilizados para geração das populações segregantes tinham seus marcadores distribuídos de forma eqüidistante dentro dos três grupos de ligação, sendo o genoma constituído de 11 marcas eqüidistantes. Vale ressaltar novamente, que o grupo de ligação 3 apresenta apenas 10 marcas, portanto haverá uma pequena alteração no valor esperado da distância entre marcas.

Quanto menores os valores de variância observados mais eqüidistantes estarão às marcas dentro dos grupos de ligação e, conseqüentemente, menores serão os erros, de forma que, os valores simulados estarão mais próximos dos valores esperados, sendo isto, um indicativo de boa recuperação do genoma com o mapeamento das populações segregantes.

Quadro 17. Variância das distâncias entre marcas adjacentes nas repetições que recuperaram três grupos de ligação em quatro tamanhos de populações não completamente informativas.

Tamanho da População	Grupos de Ligação (GL)			Média Geral	Média GL_1 e GL_2	Números de dados
	1	2	3			
100	14,972[(a)]	10,106[(a)]	14,233[(a)]	13,103[(a)]	12,539[(a)]	95
200	12,810[(a)]	5,324[(b)]	11,441[(b)]	9,858[(a)]	9,067[(a)]	96
400	11,351[(a)]	3,209[(c)]	9,873[(bc)]	8,144[(a)]	7,280[(a)]	100
600	4,953[(b)]	1,928[(d)]	8,871[(c)]	5,250[(a)]	3,440[(a)]	100

[1] indica que dentro dos parêntesis estão letras correspondentes ao resultado da análise de médias, pelo teste de Tukey ao nível de significância de 5%. Nas colunas, médias seguidas pela mesma letra não diferem estatisticamente entre si, pelo teste de Tukey de uma mesma letra não diferem entre si, pelo teste de Tukey.

No Quadro 17 pode-se verificar a redução da média da variância em relação às distâncias entre marcas adjacentes considerando os diferentes tamanhos populacionais. Porém, na análise da média geral estas diferenças entre os diferentes tamanhos populacionais não foram significativas estatisticamente. Fato semelhante foi observado quando considerou apenas os dois primeiros grupos de ligação e descartou das análises o terceiro grupo de ligação que era constituído de apenas 10 marcas moleculares. Sendo assim, neste estudo, a presença do terceiro grupo de ligação não modificou o resultado das análises.

Menores valores de variância entre marcas adjacentes implicam em maior confiabilidade dos dados, o que enfatiza que com o aumento do tamanho populacional temos maior confiabilidade dos dados.

A utilização do desvio padrão (Quadro 18) tem como finalidade observar a ocorrência na redução na amplitude de variação dos valores de variância entre repetições, à medida que aumenta o tamanho da população. Esta tendência de diminuição não ficou evidenciada, uma vez que, apenas o grupo de ligação 3 apresentou a tendência clara de declínio do desvio-padrão com o aumento da população. Ao observar o valor da média geral, ocorreu o declínio desejado, porém se desconsiderar o grupo de ligação 3 e analisar a média entre os outros grupos de ligação, mais uma vez a tendência de declínio esperado não pode ser observada , uma vez que a estimativa média da população de 400 indivíduos foi superior a de 200 indivíduos.

Com base nas informações contidas nos Quadros 17 e 18, temos que ao fazer uso de populações não completamente informativas, o aumento do tamanho da população até 600 indivíduos, como foi avaliado neste estudo, não necessariamente acarreta em uma diminuição da variância das médias entre marcas adjacentes, e também não acarreta em uma redução no desvio-padrão das médias das variâncias entre as marcas adjacentes.

Quadro 18. Desvio-padrão das médias das variâncias das distâncias entre marcas adjacentes nas repetições que recuperaram três grupos de ligação em quatro tamanhos de populações não completamente informativas.

Tamanho da População	Grupos de Ligação (GL)			Média Geral	Média GL_1 e GL_2	Números de dados
	1	2	3			
100	15,401	5,409	7,147	9,319	10,405	95
200	16,931	2,687	4,970	8,196	9,809	96
400	17,213	2,877	2,964	7,684	10,045	100
600	10,654	1,039	2,857	4,85	5,846	100

4.2.6. Estresse

De posse da distância pré-estabelecida entre os pares de marcas do genoma de referência realizou-se uma comparação com as distâncias encontradas no genoma simulado. A concordância dos valores de distâncias é expressa através do estresse. O desejado é que se tenha os menores valores possíveis de stress após as análises. Quanto menor são estes valores de stress melhor é a qualidade da população, sendo que espera-se que com o aumento do tamanho populacional ocorra um declínio nos valores médios de stress.

Foi feita uma média aritmética dos valores de estresse obtidos nas repetições em que houve a formação de três grupos de ligação no mapeamento genético, a fim de obter o estresse médio (Quadro 19).

Deve-se ter atenção ao analisar valores de estresse por estes apresentarem significados diferentes em níveis de saturação diferentes, ou seja, devem ser comparados níveis de estresse entre genomas com o mesmo nível de saturação, uma vez que valores iguais de porcentagem em níveis de saturação diferentes têm significados diferentes. Desta forma os valores de estresse do grupo de ligação 3 serão apenas apresentados a seguir e não discutidos, uma vez que o nível de saturação deste grupo de ligação é diferente dos demais.

Quadro 19. Valores do estresse médio (%) em razão do tamanho da população em populações não completamente informativas.

Tamanho da População	Grupos de Ligação (GL)			Média Geral	Média GL_1 e GL_2	Números de dados
	1	2	3			
100	25,706[(a)]	30,492[(a)]	18,179[(a)]	24,792[(a)]	28,099[(a)]	95
200	18,050[(b)]	22,474[(b)]	14,713[(b)]	18,412[(ab)]	20,262[(ab)]	96
400	15,117[(c)]	17,281[(c)]	11,524[(c)]	14,640[(ab)]	16,199[(b)]	100
600	13,386[(c)]	13,734[(d)]	8,459[(d)]	11,859[(b)]	13,560[(b)]	100

[1] indica que dentro dos parêntesis estão letras correspondentes ao resultado da análise de médias, pelo teste de Tukey ao nível de significância de 5%. Nas colunas, médias seguidas de uma mesma letra não diferem entre si, pelo teste de Tukey.

Com base no quadro anterior é fácil notar a tendência de redução dos valores de estresse médio à medida que aumenta o tamanho da população, seja com base na análise dos grupos de ligação separadamente, seja ao utilizar os dados das médias gerais do estresse, em que valores de estresse médio passam de 28,099 para 13,560 % considerando-se as populações de 100 e 600 indivíduos, respectivamente.

Com base no teste de Tukey, verifica-se que os valores de estresse médio diferem estatisticamente com o aumento do tamanho da população, exceção feita para as populações de 400 e 600 indivíduos do grupo de ligação 1 que não apresentaram diferença significativa entre suas médias, mostrando mais uma vez o efeito do tamanho populacional sobre a variável estresse. Para a média dos grupos de ligação 1 e 2 não se pode observar com clareza o efeito do aumento populacional sobre o estresse, uma vez que não houve diferença significativa entre todos os tratamentos avaliados.

Com base na análise utilizando o desvio-padrão, que representa à amplitude da variação dos valores médios de estresse, pode afirmar que ocorre redução na amplitude com o aumento do tamanho das populações avaliadas, uma vez que era esperado que esta redução ocorresse em todos os grupos de ligação analisados, porém como se pode constatar apenas para o grupo de ligação 1 ocorreu à diminuição gradual dos valores de desvio-padrão ao passo que ia aumentando o tamanho populacional. O mesmo ocorreu com a média entre os grupos de ligação 1 e 2, onde os valores de desvio padrão considerando a população de 100 e 600 indivíduos diminuem expressivamente, saindo de 7,915 na primeira população para 3,8025 na última população; Porém nas populações intermediárias

houve uma inversão do que era esperando, uma vez que a média da população de 400 indivíduos se mostra superior a média da população de 200 indivíduos (Quadro 20).

Sendo assim, apenas com base nas informações obtidas pela análise do estresse médio e do desvio-padrão relativo ao estresse não se pode concluir que com o aumento do tamanho das populações haverá uma tendência de diminuição no estresse médio e também uma diminuição na amplitude de variação dos valores médios de estresse considerando populações não completamente informativas.

Quadro 20. Desvio-padrão dos valores de estresse médio nas repetições em que recuperaram três grupos de ligação em quatro tamanhos de populações não completamente informativas.

Tamanho da População	Grupos de Ligação (GL)			Média Geral	Média GL_1 e GL_2	Números de dados
	1	2	3			
100	7,651	8,180	8,239	8,023	7,915	95
200	6,032	5,173	5,376	5,527	5,602	96
400	5,433	6,110	6,501	6,014	5,771	100
600	4,214	3,391	3,872	3,825	3,8025	100

4.3. Comparação entre a População Completamente Informativa e a População não Completamente Informativa

Após análise de cada um dos tipos de populações realizadas nos itens 4.1 e 4.2, anteriormente discutidos, fez-se necessário avaliar os resultados obtidos nos dois tipos de populações ao mesmo tempo, a fim de comparar os efeitos do tamanho populacional permitindo estabelecer uma correspondência entre estes dois tipos de populações. Na prática espera-se que o pesquisador escolha marcadores com alto grau de polimorfismo de forma que suas análises fossem feitas a partir de locos, prioritariamente, completamente informativos. Entretanto, será pouco provável que todos (ou a maioria) sejam desta natureza. Portanto, locos não-completamente informativos certamente serão também utilizados em muitos estudos de mapeamento. Assim os dois tipos de populações certamente serão utilizadas ao mesmo tempo, fazendo necessário que tenhamos a noção de como diferentes tamanhos de populações se comportam em diferentes tipos de populações.

Como já visto anteriormente a população não completamente informativa apresenta maiores complexidades, que vão desde a estimação das distâncias entre as marcas, até mesmo com o resgate das informações do genoma após simulação.

4.3.1. Recuperação de grupos de ligação e Correlação de Spearman entre medidas de distância

A princípio, o número de grupos de ligação esperado no processo de mapeamento das duas populações seria de 3 grupos de ligação, que é o número de grupos de ligação que se tem no genoma original usado para a simulação das populações.

A correlação de Spearman foi utilizada para a identificação de inversão de posição de marcas dentro de cada um dos três grupos de ligação formados no mapeamento. A avaliação da inversão só foi feita em repetições em que foram recuperados os três grupos de ligação.

Como pode ser verificado pelo Gráfico 1 a população não completamente informativa apresenta maiores dificuldades durante o mapeamento, havendo maior freqüência de inversões, além de que a teve um maior número de repetições descartadas do estudo por não conseguirem resgatar os 3 grupos de ligação previamente declarado.

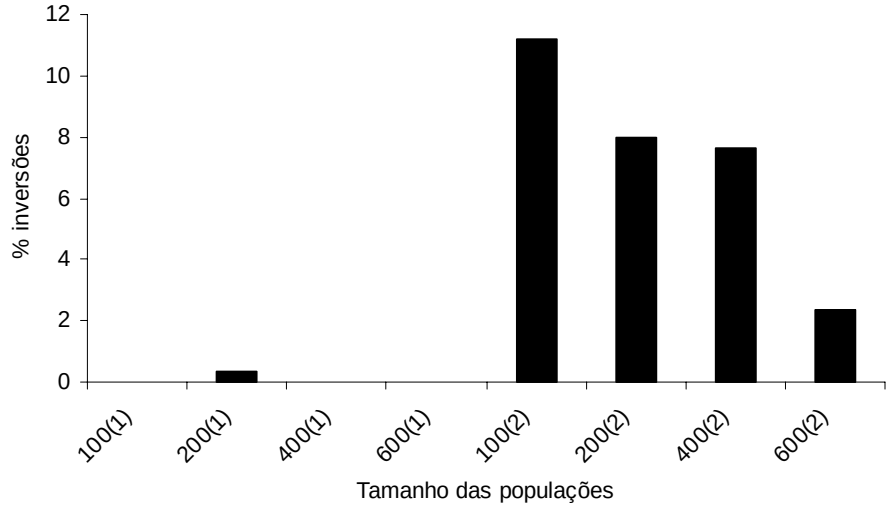

Gráfico 1. Comparação entre as populações completamente informativas (CI) (1) e as populações não completamente informativas (NI) (2), referentes aos tamanhos de população e % de inversões observadas.

Com base nos dados anteriormente apresentados verifica-se que é melhor trabalhar com uma população de 200 indivíduos completamente informativa do que com uma

população de 600 indivíduos não completamente informativas, pois ambas conseguem resgatar os três grupos de ligação porém o percentual de inversões na população completamente informativa (0,33%) é expressivamente inferior do que o apresentado pela população não informativa (2,33%).

4.3.2. Comprimento dos grupos de ligação

Como já mencionado, o comprimento médio esperado para cada grupo de ligação após o mapeamento das populações era de 100 cM, uma vez que esse é o comprimento de cada grupo de ligação no nível de saturação do genoma utilizado para a simulação das populações.

A média dos comprimentos de grupos de ligação está plotada no Gráfico 2. Apesar da variação encontrada entre as medidas do comprimento, ao realizar o Teste de Tukey (5%), não foi observada diferença estatística significativa entre nenhuma das oito situações analisadas.

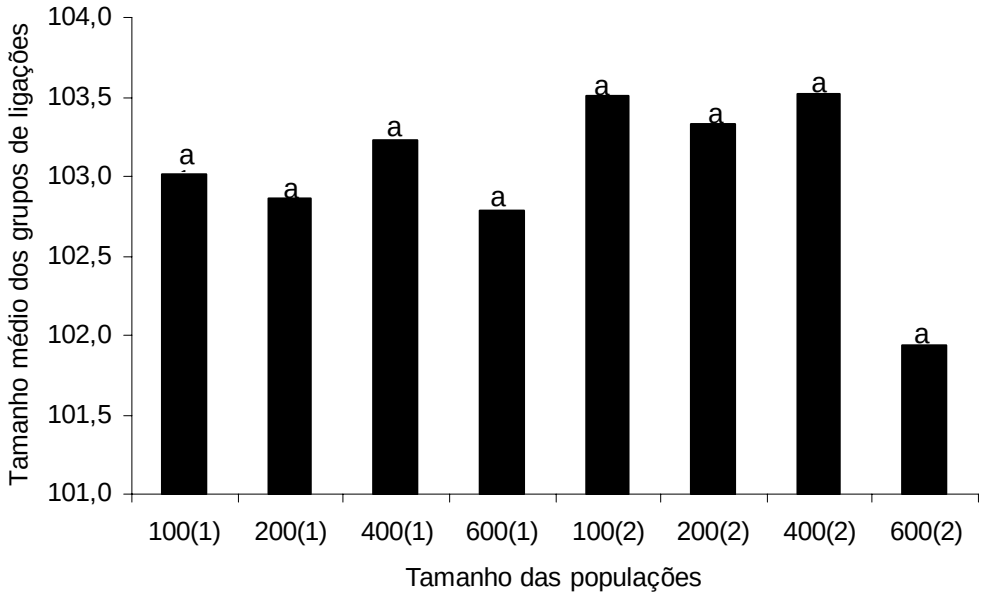

Gráfico 2. Comparação entre as populações completamente informativas (1) e as populações não completamente informativas (2), referentes aos tamanhos de população e comprimento médio dos grupos de ligação, e Teste Tukey a 5% de significância.

[1] indica as letras correspondentes ao resultado da análise de médias feita pelo teste de Tukey a 5%, em que, médias seguidas pela mesma letra não diferem estatisticamente entre si.

Desde modo pode-se concluir que não houve variação significativa no comprimento médio dos grupos de ligação com o aumento no tamanho da população e com os diferentes tipos de populações. Não se pode, portanto fazer uma referência de determinado tamanho da população completamente informativa que se aproximasse da população não completamente informativa.

A análise do desvio-padrão referente ao comprimento médio do grupo de ligação é apresentado no Gráfico 3. Ao observar os dados médios do desvio-padrão nota-se que os valores obtidos para populações completamente informativas são inferiores daqueles obtidos para população não completamente informativa em todos os tamanhos de populações avaliados. Observa-se ainda que o valor obtido para a população de 100 indivíduos completamente informativa, se aproxima muito daquele obtido na população de 400 indivíduos não completamente informativa. Porém ao realizar o Teste de Tukey a 5% verificou-se que o valor obtido para a população de 100 indivíduos C.I. não foi estatisticamente diferente aos valores de 100, 200 e 400 indivíduos da população N.I.

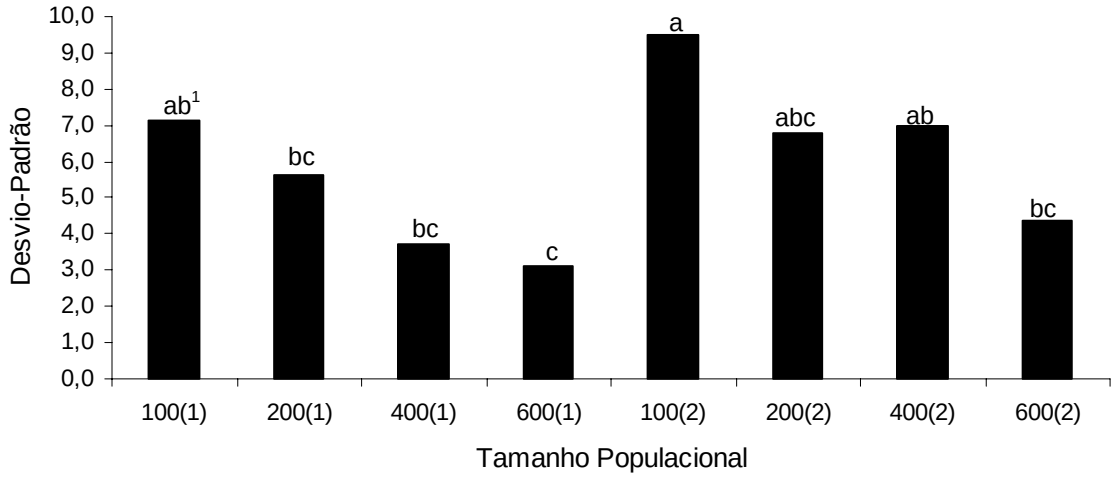

Gráfico 3. Comparação entre as populações completamente informativas (1) e as populações não completamente informativas (2) para o desvio-padrão do comprimento do grupo de ligação e Teste Tukey a 5% de significância.

[1] indica as letras correspondentes ao resultado da análise de médias feita pelo teste de Tukey a 5%, em que, médias seguidas pela mesma letra não diferem estatisticamente entre si.

4.3.3. Média das distâncias entre marcas adjacentes

A comparação feita utilizando os valores de média das distâncias entre marcas adjacentes e seu desvio padrão obtido para os vários tamanhos de população segregantes

estão representados pelos Gráficos 4 e 5, respectivamente. É importante salientar que para a obtenção destes valores médios representados foi considerado os dados oriundos do grupo de ligação 3 da população não completamente informativa, uma vez que, este grupo de ligação contou com a presença de apenas 10 marcas, apresentando um grau de saturação diferente dos demais grupos de ligação. Então, evitando problemas na comparação das informações aqui relatadas, estes valores serão excluídos.

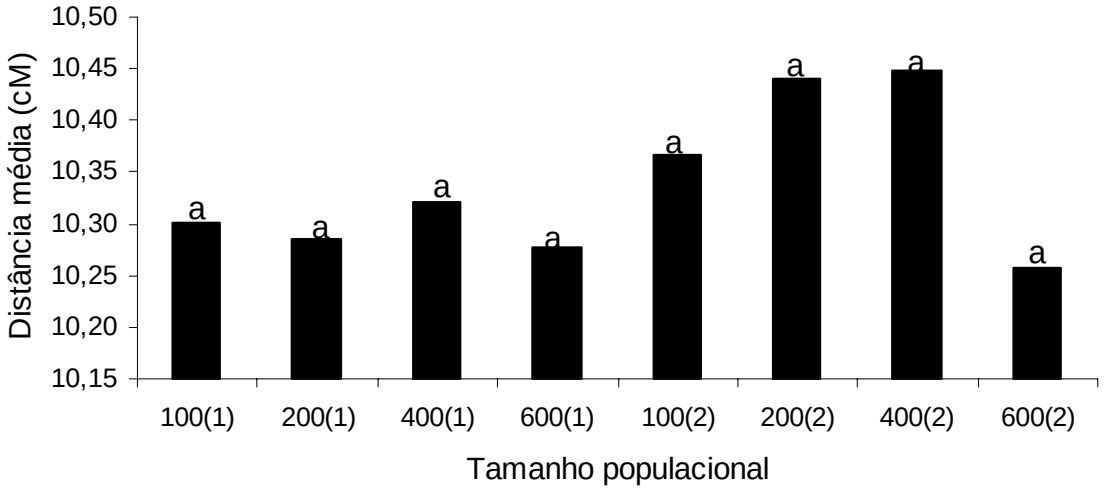

Gráfico 4. Média das distâncias, em cM, entre marcas adjacentes nas populações completamente informativas (1) e as populações não completamente informativas (2), e teste Tukey a 5% de significância.

[1] indica as letras correspondentes ao resultado da análise de médias feita pelo teste de Tukey a 5%, em que, médias seguidas pela mesma letra não diferem estatisticamente entre si.

Em relação ao teste de Tukey, realizado com $P<0,05$, não foi encontrada diferença entre as médias para os tamanhos populacionais avaliados, isto mostra que com a variação no tamanho populacional, e com os dois tipos de populações avaliados não foi possível proporcionar uma vantagem significativa no momento de tentar resgatar os valores do genoma original (Gráfico 4).

Para o desvio-padrão das médias das distâncias entre marcas adjacentes o teste Tukey realizado a 5% de significância mostra que existe diferença entre os tratamentos avaliados (Gráfico 5). Observa-se que a população completamente informativa de 200 indivíduos mostrou comportamento semelhante à população não completamente

informativa de 600 indivíduos. Isto nos dá uma idéia do tamanho necessário de determinado tipo de população pra ser mais bem representada. Assim uma idéia preliminar é que ao trabalharmos com populações completamente informativas poderemos fazer uso de apenas um terço do tamanho populacional daquele usado pra populações não informativas.

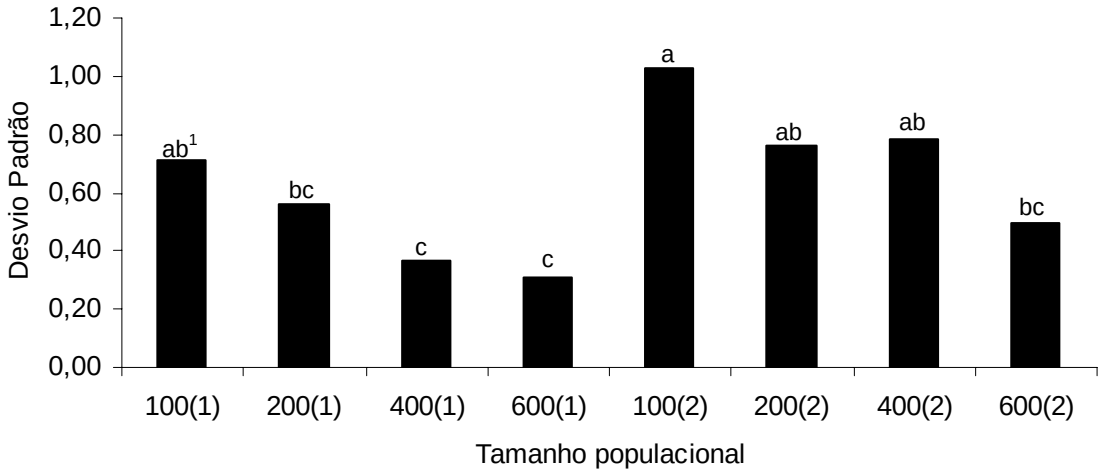

Gráfico 5. Desvio-padrão das médias das distâncias entre marcas adjacentes nas populações completamente informativas (1) e as populações não completamente informativas (2), e teste Tukey a 5% de significância.

[1] indica as letras correspondentes ao resultado da análise de médias feita pelo teste de Tukey a 5%, em que, médias seguidas pela mesma letra não diferem estatisticamente entre si.

4.3.4. Variância das distâncias entre marcas adjacentes

A partir das distâncias entre marcas adjacentes obtidas nos grupos de ligação foi estimada a variância amostral média, que pode ser interpretada de forma que quanto menores os valores de variância, mais eqüidistantes estarão distribuídos às marcas dentro dos grupos de ligação e conseqüentemente menor o erro. Portanto quanto menores os valores de variância mais próximos estarão os valores do esperado, indicando uma boa recuperação do genoma com o mapeamento das populações segregantes.

Analisando o efeito do tamanho e dos tipos de população, observa-se que os valores obtidos para a população completamente informativa são menores do que aqueles obtidos para a população não completamente informativa considerando os mesmos tamanhos populacionais. Verifica ainda que o valor para a população de 200 indivíduos

completamente informativa é inferior a estimativa apresentada na população de 600 indivíduos não completamente informativa. Porém, ao realizar o teste de médias, estes valores destas populações não foram significativamente diferentes. Sendo assim, mais uma vez percebe-se uma tendência de correlação entre os dados apresentados pela população de 200 indivíduos completamente informativa e a de 600 indivíduos não completamente informativa (Gráfico 6).

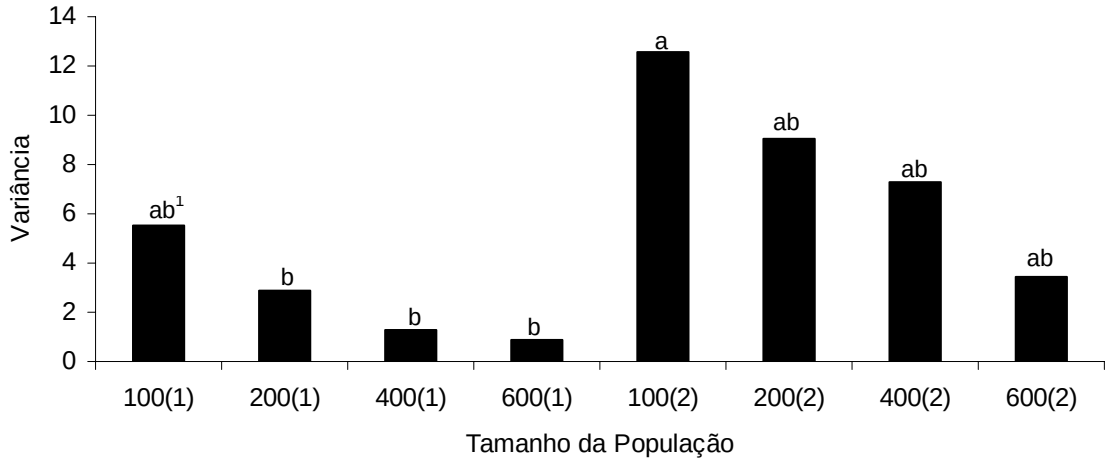

Gráfico 6. Variância das distâncias entre marcas adjacentes nas populações completamente informativas (1) e as populações não completamente informativas (2), e teste Tukey a 5% de significância.

[1] indica as letras correspondentes ao resultado da análise de médias feita pelo teste de Tukey a 5%, em que, médias seguidas pela mesma letra não diferem estatisticamente entre si.

No Gráfico 7, tem-se o desvio padrão das médias das variâncias das distâncias entre marcas, e o teste Tukey realizado a 5% de significância. Observa-se que os valores obtidos pela população completamente informativa são bem inferiores àqueles obtidos pela população não completamente informativa, chegando a ponto de que o maior valor da população C.I.(4,324) é menor do que o menor valor obtido na população N.I.(5,846). Porém, apesar desta diferença entre os valores das duas populações o teste de média não apontou nenhuma diferença estatística significativa entre os valores, não sendo possível, tirar maiores conclusões a respeito destes dados.

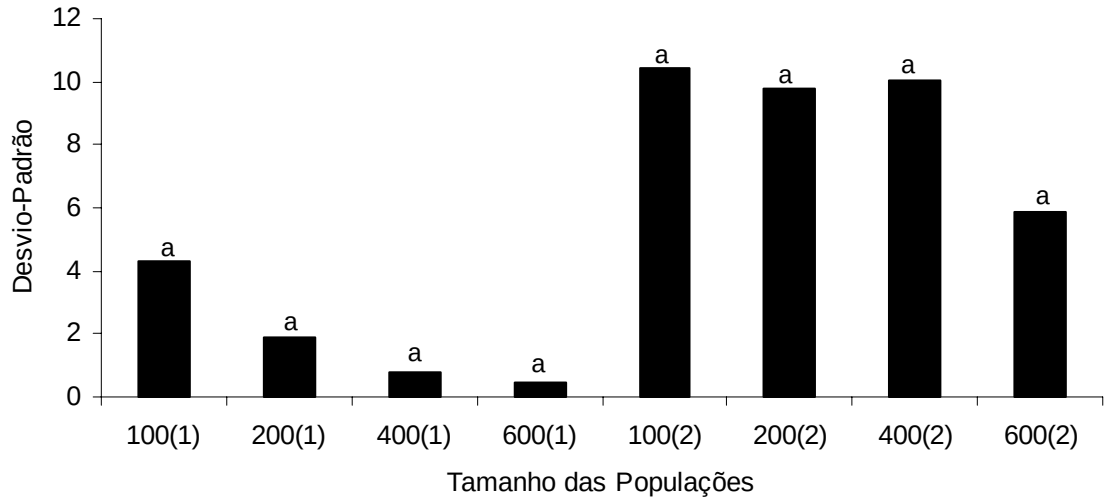

Gráfico 7. Desvio-padrão da média das variâncias das distâncias entre marcas adjacentes nas populações completamente informativas (1) e as populações não completamente informativas (2), e teste Tukey a 5% de significância.

[1] indica as letras correspondentes ao resultado da análise de médias feita pelo teste de Tukey a 5%, em que, médias seguidas pela mesma letra não diferem estatisticamente entre si.

4.3.5. Estresse

Os valores de estresse obtidos são apresentados a seguir, no Gráfico 8, juntamente com o resultado de comparações feitas com teste de médias. Analisando os efeitos do tamanho da população observa-se claramente a tendência de redução dos valores de estresse médio à medida que aumenta o tamanho da população nas duas populações avaliadas. Através do teste de Tukey, realizado com significância de 5%, verifica-se que os valores de estresse médio diferem estatisticamente, e que os valores obtidos na população C.I. de 200 indivíduos se assemelham muito aqueles obtidos na população N.I. de 400 indivíduos, porém em termos estatísticos os valores da população C.I. de 200 indivíduos não se diferem dos valores das populações de 200, 400 e 600 indivíduos na população N.I., se tornando assim, difícil de observar uma tendência clara de relação entre os tipos de populações e seus tamanhos.

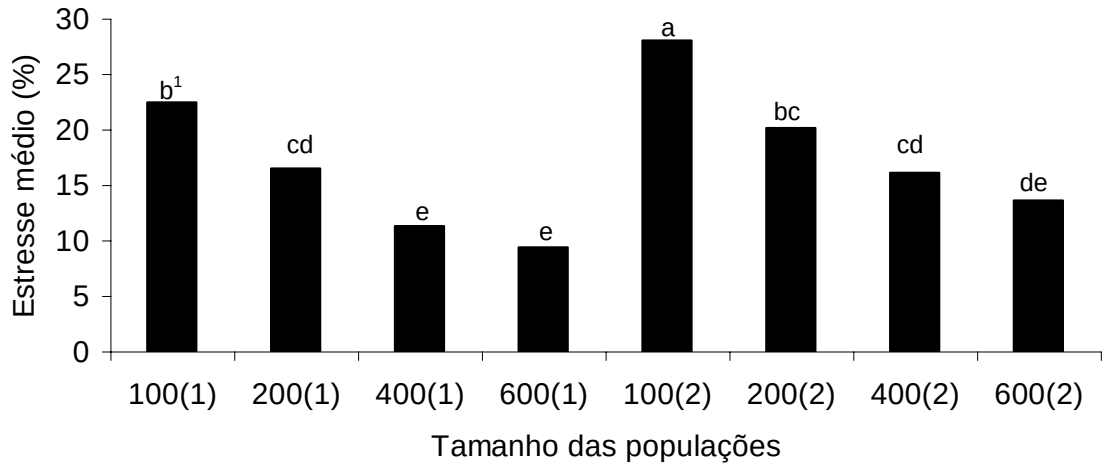

Gráfico 8. Valores do estresse médio (%) em razão do tamanho da população em populações completamente informativas (1) e populações não completamente informativas (2), e teste Tukey a 5% de significância.

[1] indica as letras correspondentes ao resultado da análise de médias feita pelo teste de Tukey a 5%, em que, médias seguidas pela mesma letra não diferem estatisticamente entre si.

Em relação à amplitude da variação dos valores médios de estresse (Gráfico 9), percebe-se através do teste de Tukey apresentado a 5% de probabilidade que para os dois tipos de populações o tamanho de 100 indivíduos comportou-se de forma semelhante. A população de 200 indivíduos C.I. teve comportamento semelhante as populações de 200, 400 e 600 indivíduos na população N.I., uma vez que o teste de média não permitiu diferenciar estas populações. A população de 400 indivíduos C.I., apresentou comportamento semelhante ao da população de 600 indivíduos N.I.. Sendo assim, mais uma vez não se pode concluir de forma evidente uma relação entre o tipo de população completamente informativa e não completamente informativa.

Porém, apesar de não ter ficado evidente a relação entre estes tipos de populações, ficou claro que em todas as análises necessitou de menor tamanho de população completamente informativa do que não completamente informativa. Este resultado é muito importante em vista a aplicação dos resultados deste trabalho, uma vez que, como já citado anteriormente, na prática provavelmente sempre trabalha-se com marcadores pertencentes a ambos os tipos de populações, sendo raras às vezes em que tem-se apenas um tipo de população, onde os marcadores são completamente informativos, sendo analisada em determinado estudo.

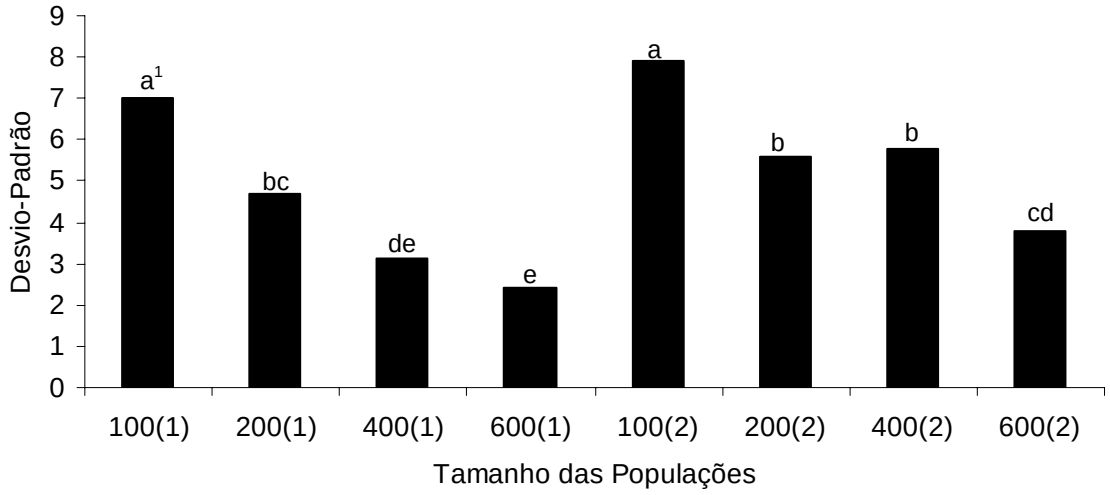

Gráfico 9. Desvio-padrão dos valores de estresse médio nas populações completamente informativas (1) e nas populações não completamente informativas (2), e teste Tukey a 5% de significância.

[1] indica as letras correspondentes ao resultado da análise de médias feita pelo teste de Tukey a 5%, em que, médias seguidas pela mesma letra não diferem estatisticamente entre si.

5. Considerações Finais

A disponibilidade de mapas genéticos fidedignos depende do número de marcas moleculares e do número de indivíduos analisados. Porém, o tamanho de famílias de irmãos completos ainda não havia sido alvo específico de estudos, como já aconteceu com populações F_2, retrocruzamento (Cruz, 2006) e RILs (Silva, 2005). Sendo assim, o tamanho utilizado de família de irmãos completos era determinado, primordialmente pela disponibilidade de material e recursos. Como resultado, os mapas disponíveis na literatura foram construídos, cada um, com um número diferente de indivíduos. Portanto, sem atenção necessária aos efeitos prejudiciais causados pelo número insuficiente de indivíduos utilizados no mapeamento, conforme demonstrado neste trabalho de simulação.

Sendo assim, diferentes tamanhos de populações e saturação de mapa são encontrados na literatura. Por exemplo, Ukrainetz et al., (2008) trabalhou com 8 famílias de irmãos completos com 40 indívudos cada, usando marcadores AFLP para realização do mapa genético em Douglas-fir (*Pseudotsuga menziesii)*. O mapa realizado era constituído de 19 grupos de ligações, com comprimento total de 938cM e com saturação média de 9,3 cM entre marcas. Doligez et al., 2006, em trabalho semelhante de mapeamento em videira

(*Vitis vinifera*) utilizou 5 populações de irmão completos de tamanhos variando de 46 a 153 para fazer um mapa integrado, utilizando 439 pares de primers (426 destes SSR) Este autor comenta ainda que os mapas de cada população separada foram bem conservados, e que o mapa integrado possuía 1,647 cM distribuídos em 19 grupos de ligação e uma distância média entre marcas de 3,3 cM.

É importante salientar que na prática ao utilizar famílias de irmãos completos para mapeamento as duas situações de marcadores serão encontradas, tanto completamente informativos quanto não informativos. Além disso, deve-se ressaltar novamente a importância deste tipo de população para estudos genômicos, uma vez que estas são utilizadas como populações de melhoramento, e devem cada vez mais ser utilizadas também como populações de mapeamento, o que faz com que possam ter uma grande importância na pesquisa de diversas culturas e espécies animais.

Todavia, espera-se, que este trabalho contribua para um melhor entendimento dos efeitos do número de indivíduos no mapeamento de famílias de irmãos completos, assim como para fornecer subsídios para a escolha adequada desta variável quando da condução de futuros trabalhos.

6. Conclusões

Para obtenção de informações suficientes de modo que sejam gerados mapas confiáveis, análises devem estar associadas à utilização de tamanho de amostra e número de marcas adequadas. Mapas com sérias distorções serão obtidos com utilizações de uma quantidade inadequada de indivíduos.

A maior confiabilidade dos mapas não aumenta indefinidamente com aumento do tamanho da população, sendo que, com base no apresentado, pode-se afirmar que, para populações com marcadores completamente informativos, um tamanho populacional de 200 indivíduos seria o suficiente para resgatar as informações originais de forma satisfatória, podendo desta forma economizar esforços de avaliar e genotipar grande quantidade de indivíduos. Porém, conforme verificado nos dados quanto maior o tamanho da população avaliada melhor seriam as estimativas obtidas.

Para a população com locos não completamente informativa, usando marcadores com PIC = 0,7, seria mais confiável utilizar um tamanho populacional de 600 indivíduos, uma vez que, apesar da população de 400 indivíduos ter conseguido resgatar os 3 grupos de

ligação, os valores das estimativas anteriormente discutidas para a população de 400 indivíduos estão muito próximas dos valores obtidos nas estimativas da população de 200 indivíduos, e muitas vezes estas populações tiveram comportamento invertido nas análises.

7. Referências Bibliográficas

BRUNELLI, K.R., SILVA, H.P., CAMARGO, L.E.A. Mapeamento de genes de resistência a *Puccinia polysora* em milho. **Fitopatologia Brasileira** 27: 134-140. 2002.

CARBONELL, E.A., ASINS, M.J., BASELGA, M., BALANSARD, E., GERIG, T.M. Power studies in the estimation of genetic parameters and the localization of quantitative trait loci for backcross and doubled haploid populations. **Theor. Appl. Genet**. 86: 411-416. 1993.

CHATZIPLIS, D., HALEY, C.S. Selective genotyping for QTL detection using sib pair analysis in outbred populations with hierarchical structures. **Genet. Sel. Evol**., 32: 547-560, 2000.

CLARKE, G.M. **Statistics and experimental design: an introduction for biologists and biochemists**. Third Edition, New York-Toronto, 1994, 280p.

CONTI, J.H., MINAMI, K., GOMES, L.H.,TAVARES, F.C.A. Estimativa da similaridade genética e identificação de cultivares de morangueiro por análise de RAPD. **Horticultura Brasileira**, Brasília, 20: n. 2, p. 145-152, junho 2002.

CORRÊA, F.J.C. **Avaliação de métodos de seleção tradicionais, assistida por marcadores moleculares e por genes candidatos, com dados simulados**. Viçosa, MG: UFV, 2001, 54p. Tese (Mestrado em Zootecnia) Universidade Federal de Viçosa, 2001.

CORRÊA, R.X., ABDELNOOR, R.V., FALEIRO, F.G., CRUZ, C.D., MOREIRA, A.M., BARROS, E.G. Genetic distances in soybean base don RAPD markers. **Bragantina**, Campinas, 58 (1): 15-22, 1999.

CORRÊA, R.X., GOOD-GOD, P.I.V., OLIVEIRA, M.L.P., NIETSCHE, S., MOREIRA, M.A. e BARROS, E.G. Herança da resistência à mancha-angular do feijoeiro e identificação de marcadores moleculares flanqueando o loco de resistência. **Fitopatologia Brasileira** 26: 27-32. 2000.

CRUZ, C.D. **GENES: Programa de análise e processamento de dados baseados em modelos de genética e estatística experimental**. Versão 2004.2.1 Viçosa: UFV.

CRUZ, C.D. **Programa para análise de dados moleculares e quantitativos – GQMOL**. Viçosa: UFV, 2005.

CRUZ, C.D., CARNEIRO, P.C.S. **Modelos biométricos aplicados ao melhoramento genético. Vol. 2**, Viçosa: Imprensa Universitária, 2003. 585p.

CRUZ, C.D. In: NASS, L.L., VALOIS, A.C.C., MELO, I.S., VALADARES-INGLIS, M.C. **Recursos Genéticos & Melhoramento.** Rondonópolis, MT: Fundação MT, 2001. 1183 p.

CRUZ, E.M.. **Efeito da saturação e do tamanho de populações f2 e de retrocruzamento sobre a acurácia do mapeamento genético** (2006). Tese (Doutorado em Genética e Melhoramento) -Universidade Federal de Viçosa, Viçosa.

DACHS, J.N.W. **Estatística Computacional.** Rio de Janeiro: LTC Editora,1988, 236p

DEMPSTER, A.B., LAIRD, N.M., RUBIN, D.B. Maximum Likelihood from Incomplete Data via the EM Algorithm. **Journal Royal Statistical Society (B)**. 39: 71361-71398, 1977.

DOLIGEZ, A., ADAM-BLONDON, A. F., CIPRIANI, G., DI GASPERO, G. LAUCOU, V., MERDINOGLU, D., MEREDITH, C. P., RIAZ, S., ROUX,C., THIS, P. An integrated SSR map of grapevine based on five mapping populations. Theor Appl Genet 113: 369–382. 2006.

FALCÃO, C.L., PAPPAS, M.C.R., LOURENÇO, R.T., ALENCAR, M.M., BATISTA, A.R.S., PAPPAS Jr, G.J., GRATTAPAGLIA, D. **Desenvolvimento e mapeamento de microssatélites derivados de ESTs em *Eucalyptus*.** Brasília, DF: EMBRAPA, Circular Técnica 32, Dezembro, 2004.

FALCONER, D.S. **Introdução à genética quantitativa**. Tradução de M.A. Silva e j.c. SILVA. Viçosa, MG: UFV, Impr. Univ., 1987. 279p.

FALEIRO, F.G., RAGAGNIN, V.A., SCHUSTER, I., CORRÊA, R.X., GOOD-GOD, P.I., BROMMONSHENKEL, S.H., MOREIRA, M.A., BARROS, E.G. Mapeamento de genes de resistência do feijoeiro à ferrugem, antracnose e mancha-angular usando marcadores RAPD. **Fitopatologia Brasileira** 28: 059-066. 2003.

FRISH, M., BOHN, M., MELCHINGER, A. E. Comparison of selection strategies for marker-assisted backcrossing of a gene. **Crop Sci**. 39: 1295-1301. 1999.

GANGA, R.M.D., RUGGIERO, C., LEMOS, E.G. de M., GRILI, G.V.G., GONÇALVES, M.M., CHAGAS, E.A., WICKERT, E. Diversidade genética em maracujazeiro-amarelo

utilizando marcadores moleculares fAFLP. **Rev. Bras. Frutic**., Jaboticabal - SP, 26: n. 3, p. 494-498, Dezembro 2004

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

GODDARD, K.A.B.; GOODE, E.L.; ROZEK, L.S.. Impact of structure on the power of linkage tests using sib-pair methods. **Genet. Epidemiol**., 17: Suppl.1, p.S575-S579, 1999.

GUO, X., ELSTON, R.C. Two-stage global search designs for linkage analysis II: Including discordant relative pairs in the study. **Genet. Epidemiol**., 18: p.111-127, 2000.

HALDANE, J.B.S. The combination of linkage values and the calculation of distances between the loci of linked factors. **J. Genet.** 8: 299-309. 1919.

HASEMAN, J.K., ELSTON, R.C. The investigation of linkage between a quantitative trait and a marker locus. **Behav. Genet.** 2: 3±19.1972.

HOSPITAL, F., CHEVALET, C., MULSANT, P. Using markers in gene introgression breeding programs. **Genetics**. 132: 1199-1210. December, 1992.

HU X-S, GOODWILLIE C, RITLAND K., Joining genetic linkage maps using a joint likelihood function. **Theor Appl Genet** 109: 996–1004. 2004.

KNOTT, S. A., HALEY, C. S. Aspects of maximum likelihood methods for the mapping of quantitative trait loci in line crosses. **Genet. Res.**, Camb. 60: 139-151. 1992.

KOSAMBI, D.D. The estimation of map distances from recombination values. **Ann Eugen**. 12: p.172-175. 1944.

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

LANDER, E.S., GREEN, P., ABRAHAMSON, J. BARLOW, A., DALY, M.J., LINCOLN, S.E., NEWBURG, L. MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. **Genomics.** Oct, 1(2): 174-81. 1987.

LANNES, S.D., ZIMMER, P.D., OLIVEIRA, A.C. de, IRAJÁ, F., CARVALHO, F. de, VIEIRA, E.A, JUNIOR, A.M. de M., KOPP, M.M., FREITAS, F.A. de. Regeneração *in*

vitro de anteras de arroz irrigado (*Oryza sativa* L.) e mapeamento de QTL associado. **Ciência Rural**. 34: n.5, p.1355-1362, set-out 2004.

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, USA: Sinauer Associates, Inc. Ed., 1998.

MALIEPAARD, C.; JANSEN, J.; VAN OOIJEN, J.W. Linkage análisis in a full-sib family of an outbreeding plant species: overview and consequences. **Genet. Res. Camb.** 70: 237-250. 1997.

MARTINEZ, M. L., VUKASINOVIC, N., FREEMAN, A.E. Random model approach for QTL mapping in half-sib families. **Genet. Sel. Evol.** Elsevier, Paris. 31: p.319-340, 1999.

MARTÍNEZ, O., CURNOW, R. N. Estimating the locations and the size of the effects of quantitative trait using flanking markers. **Theor. Appl. Genet.** 85: 480-488. 1992.

MORGAN, T. H., Sex limited inheritance in Drosophila. **Science** v.32: p120-122, 1910.

NOVAES, E. Detecção de QTls para qualidade de madeira em Eucalyptus grandis x Eucalyptus urophylla e ancoragem de clones BAC no mapa genético. **Dissertação de mestrado** (Apresentada a Universidade Federal de Viçosa (UFV) no programa de Genética e Melhoramento). 171p. 2006.

OOIJEN, J.W.V. Accuracy of mapping quantitative trait loci in autogamous species. **Theor. Appl. Genet** 84: 803-8011. 1992.

PATERSON, A.H., LANDER, E., HEWITT, J.D., PETERSON, S., LINCLN, S.E., TANKSLEY, S.D. Resolution of quantitative traits into mendelian factors by using a complete linkage map of restriction fragment length polymerphisms. **Nature**. 355: 721-726. 1988.

ROCHA, R.B., PEREIRA, J. F., CRUZ, C.D., QUEIROZ, M.V., ARAÚJO, E. F.O Mapeamento Genético no Melhoramento de plantas. **Revista Biotecnologia Ciência e desenvolvimento**, 30: p27-31, 2003.

SCHUSTER, I., CRUZ, C.D. **Estatística genômica aplicada a populações derivadas de cruzamentos controlados**. Viçosa: UFV, 2004. 568p.

SILVA, A.R. da. **Análise genética de caracteres quantitativos em milho com delineamento III e marcadores moleculares. Piracicaba,** SP: ESALQ, 2002. 143p. (Dissertação Doutorado). Escola Superior de Agricultura Luiz de Queiroz, 2002.

SILVA, L. da C. e. **Simulação do tamanho da população e da saturação do genoma para mapeamento genético de RILs**. Viçosa, MG: UFV, 2005. 120p. (Dissertação de mestrado). Universidade Federal de Viçosa, 2005.

SILVA, M.V.G.B; MARTINEZ, M.L.; TORRES, R.A.; LOPES, P.S.; EUCLYDES, R.F.; MACHADO, M.A.; ARBEX, W. Mapeamento de QTL em famílias de irmãos completos por meio de modelos aleatórios. **Arq. Bras. Med. Vet. e Zootec**. 56: nº2. 2004.

SILVA, M.V.G.B. Utilização de modelos aleatórios na estimação da localização de QTL em famílias de meios-irmãos. **Tese**(Doutorado em Genética e Melhoramento). Universidade Federal de Viçosa, Viçosa, MG. 112p 2002.

STUTERVANT, A. H. **A History of Genetics.** New York: Harper e Row, 1965. 156p.

STUTERVANT, A.H. The linear arrangement of six-linked factors in Drosophila, as shown by their mode of association. **Journal of Experimental Zoology**, 14: p. 43-59. 1913.

TELLES, M.P. de C., MONTEIRO, M.S.R., RODRIGUES, F.M., SOARES, T.N., RESENDE, L.V., AMARAL, A. das G., MARRA, P.R. Marcadores RAPD na análise da divergência genética entre raças de bovinos e número de locos necessários para a estabilidade da divergência estimada. **Ciência Animal Brasileira** 2(2): 87-95, jul./dez. 2001.

UKRAINETZ, N.K., RITLAND, K., MANSFIELD, S.D. An AFLP linkage map for Douglas-fir based upon multiple full-sib families. **Tree Genetics & Genomes** 4: 181–191. 2008.

VIDOR, M.A., RUIZ, C.P., MORENO, S.V., FLOSS, P.A. Marcadores moleculares em estudos de caracterização de erva-mate (*Ilex paraguariensis* St. Hil.): O sabor. **Cienc. Rural** 32: n.3 Santa Maria May/June 2002

VISSCHER, P.M., HALEY, C.S., THOMPSON, R. Marker-assisted introgression in backcross breeding programs. **Genetics.** 144: 1923-1932. 1996.

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

YOUNG, N.D. **Constructing a plant genetic linkage map with DNA makers**. In: PHILLIPS, R.L., VASIL, I.K. DNA-Based Markers in Plants. Dordrecht, The Netherlands: Kluwer Academic Publisher, 1994. p.39-57.