PEDRO IVO VIEIRA GOOD GOD

MAPEAMENTO GENÉTICO EM FAMÍLIAS DE MEIO-IRMÃOS POR SIMULAÇÃO COMPUTACIONAL

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

PEDRO IVO VIEIRA GOOD GOD

MAPEAMENTO GENÉTICO EM FAMÍLIAS DE MEIO-IRMÃOS POR SIMULAÇÃO COMPUTACIONAL

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: 23 de julho de 2008.

Prof. Cosme Damião Cruz (Co-Orientador)	Prof. Maurílio Alves Moreira (Co-Orientador)
Dra. Cláudia Teixeira Guimarães	Prof. Fabyano Fonseca e Silva

Prof. Everaldo Gonçalves de Barros
(Orientador)

A Deus.

À minha mãe, Sueli Vieira Good God.

Ao meu pai, Hieroises Maria Good God.

Em especial à minha esposa, Liliane Evangelista Visôtto.

Ao meu filho Théo Visôtto Good God (In Memorian).

Dedico.

AGRADECIMENTOS

A Deus, sobretudo.

À Universidade Federal de Viçosa e ao Curso de Pós-Graduação em Genética e Melhoramento, pela oportunidade sem igual na realização do curso.

Ao Instituto de Biotecnologia Aplicada à Agropecuária (BIOAGRO), pela ótima infra-estrutura para a realização dos trabalhos científicos.

Ao CNPq e à CAPES, pela concessão da bolsa de estudos.

Ao Professor Everaldo Gonçalves de Barros, pela orientação, desde os tempos de graduação, pela extrema paciência, e, acima de tudo, pela amizade e respeito sempre demonstrados no decorrer dos anos.

Ao Professor Cosme Damião Cruz, pelo estímulo sempre presente e confiança irrestrita demonstrada.

Ao Professor Maurílio Alves Moreira, pela receptividade sempre presente, pela colaboração e incentivo na execução deste trabalho.

Aos demais professores do Curso de Pós-Graduação em Genética e Melhoramento, pela convivência e paciência, pelos ensinamentos e pela dedicação sempre externadas.

Ao Professor Ronan Xavier Corrêa, pelos primeiros e essenciais passos ensinados na árdua carreira de pesquisador.

À Liliane Evangelista Visôtto, pelo amor e carinho dedicados, pelo incentivo e ajuda disponíveis, sem os quais não passaria.

A todos os demais que colaboraram e contribuíram para o êxito deste trabalho.

BIOGRAFIA

PEDRO IVO VIEIRA GOOD GOD, filho de Sueli Vieira Good God e Hieroises Maria Good God, nasceu em 16 de agosto de 1977, na cidade de Belo Horizonte, estado de Minas Gerais.

Em fevereiro de 1993 ingressou no Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG), tendo concluído o curso de Técnico Industrial em Edificações em dezembro de 1995.

À partir de março de 1997, iniciou o curso de graduação em Agronomia na Universidade Federal de Viçosa (UFV), em Viçosa, Minas Gerais, tendo colado grau como Engenheiro Agrônomo em setembro de 2002.

Em setembro de 2002, iniciou o curso de Mestrado em Genética e Melhoramento, submetendo-se à defesa de tese em 04 de agosto de 2004.

Em agosto de 2004, iniciou o curso de Doutorado em Genética e Melhoramento, submetendo-se à defesa de tese em 23 de julho de 2008.

SUMÁRIO

LISTA DE FIGURAS E TABELAS..... vii
LISTA DE ABREVIATURAS..... x
RESUMO..... xi
ABSTRACT..... xiii
1. INTRODUÇÃO..... 01
2. REVISÃO DE LITERATURA..... 05
2.1. Construção de Mapas Genéticos de Ligação..... 05
2.2. Delineamentos Experimentais no Mapeamento Genético e na Análise de QTL..... 09
2.2.1. Populações Endogâmicas..... 10
2.2.2. Populações Exogâmicas..... 11
2.3. Informatividade no Mapeamento Genético e na Análise de QTL..... 14
2.3.1. Medidas de Informatividade..... 18
2.4. Mapeamento de QTL – Definição e Modelo..... 19
2.5. Métodos de Mapeamento de QTLs em Populações Exogâmicas..... 20
2.6. Eficiência e Poder Estatístico na Estimação no Mapeamento de QTL..... 23
2.7. Intervalos de Confiança e Nível de Significância Genômico..... 25
2.8. Mapas Genéticos e Mapeamento de QTL – Limitações..... 26
3. METODOLOGIA..... 30
3.1. Simulação de Genomas..... 30
3.1.1. Genoma de Referência..... 30
3.1.1.1. Simulação do Nível de Saturação de Marcas nos Grupos de Ligação..... 30
3.1.1.2. Simulação do Nível de Informação Alélica na População de Referência..... 31
3.1.1.3. Simulação do Tamanho Amostral da População de Mapeamento..... 33
3.2. Procedimentos de Simulação do Genitor P_1 e dos Indivíduos das Progênies de Meios -Irmãos..... 33

3.3. Análise Genômica – Mapeamento..... 35
3.3.1. Análise de Segregação de Locos Individuais..... 35
3.3.2. Análise de Pares de Marcas – Estimação da Fase de Ligação e Percentagem de Recombinação..... 35
3.3.3. Determinação dos Grupos de Ligação e Ordenamento das Marcas..... 35
3.4. Comparação de Genomas..... 36
3.4.1. Número de Grupos de Ligação Recuperados..... 36
3.4.2. Tamanho do Grupo de Ligação..... 37
3.4.3. Distância Média de Dois Marcadores Adjacentes no Grupo de Ligação..... 37
3.4.4. Variância das Distâncias entre Marcas Adjacentes..... 37
3.4.5. Estresse..... 38
3.4.6. Correlação de Spearman..... 39
3.4.7. Testes de Comparação Múltipla de Médias..... 39
4. RESULTADOS E DISCUSSÃO..... 40
4.1. Aplicação do Mapeamento Genético em Famílias de Meios-Irmãos... 40
4.1.1. Informatividade e Análise de Loco Único..... 40
4.1.2. Informatividade e Análise de Pares de Locos..... 46
4.2. Efeitos do Tamanho da População, Níveis de Informatividade e Saturação de Marcas no Mapeamento Genético..... 53
4.2.1. Número de Grupos de Ligação Recuperados..... 54
4.2.2. Tamanho dos Grupos de Ligação e Distância Média..... 57
4.2.3. Variância das Distâncias entre Marcas Adjacentes..... 66
4.2.4. Estresse..... 71
4.2.5. Correlação de Spearman..... 74
5. CONCLUSÕES..... 92
6. REFERÊNCIAS BIBLIOGRÁFICAS..... 95

LISTA DE FIGURAS E TABELAS

Figura 1. Principais etapas metodológicas utilizadas na construção de mapas genéticos..... 06

Figura 2. Populações de mapeamento e delineamentos experimentais utilizados no mapeamento genético e na análise de QTL. Adaptado de Weller (2001)..... 10

Figura 3. Esquema demonstrando o processo de obtenção de delineamentos experimentais utilizados na construção de mapas genéticos e na análise de QTL em cruzamentos exogâmicos..... 12

Figura 4. Genoma de referência com três grupos de ligação sob diferentes níveis de saturação. Cada grupo de ligação apresenta 11 marcas codominantes e eqüidistantes em 5, 10 e 15 cM..... 31

Figura 5. LOD escore médio em função do nível de informação alélica (s) para os diferentes tamanhos de FMI (N). (A) Alta Saturação – 5 cM; (B) Média Saturação – 10 cM; (C) Baixa Saturação – 15 cM..... 52

Figura 6. Número de grupos de ligação recuperados em função do tamanho da população N e o grau de saturação para diferentes informatividades – A (s = 2); B (s = 3); C (s = 4); D (s = 5); E (s = 6)..... 56

Figura 7. Box-Plot para a dispersão amostral de estimativas de tamanho dos grupos de ligação e distâncias médias entre marcas adjacentes para tamanho da população (N) sob diferentes graus de saturação..... 61

Figura 8. Redução da variância em função do tamanho da população N e o grau de saturação para diferentes informatividades – A (s = 2); B (s = 3); C (s = 4); D (s = 5); E (s = 6)..... 66

Figura 9. Estresse percentual médio em função do Tamanho da População (N) para os diferentes níveis de informatividade. (A) Alta Saturação; (B) Média Saturação; (C) Baixa Saturação..... 74

Figura 10. Correlação de Spearman média em função do Tamanho da População (N) para os diferentes níveis de informatividade. (A) Alta Saturação; (B) Média Saturação; (C) Baixa Saturação..... 77

Figura 11. Mapa de Ligação recuperado para tamanho da população (N=50) e nível de informatividade (s = 3)..... 78

Figura 12. Mapa de Ligação recuperado para tamanho da população (N=50) e nível de informatividade (s = 4)..... 79

Figura 13. Mapa de Ligação recuperado para tamanho da população (N=50) e nível de informatividade (s = 5)..... 80

Figura 14. Mapa de Ligação recuperado para tamanho da população (N=50) e nível de informatividade (s = 6)..... 81

Figura 15. Mapa de Ligação recuperado para tamanho da população (N=200) e nível de informatividade (s = 2)..... 82

Figura 16. Mapa de Ligação recuperado para tamanho da população (N=200) e nível de informatividade (s = 3)..... 83

Figura 17 Mapa de Ligação recuperado para tamanho da população (N=200) e nível de informatividade (s = 4)..... 84

Figura 18. Mapa de Ligação recuperado para tamanho da população (N=200) e nível de informatividade (s = 5)..... 85

Figura 19. Mapa de Ligação recuperado para tamanho da população (N=200) e nível de informatividade (s = 6)..... 86

Figura 20. Mapa de Ligação recuperado para tamanho da população (N=1000) e nível de informatividade (s = 2)..... 87

Figura 21. Mapa de Ligação recuperado para tamanho da população (N=1000) e nível de informatividade (s = 3)..... 88

Figura 22. Mapa de Ligação recuperado para tamanho da população (N=1000) e nível de informatividade (s = 4)..... 89

Figura 23. Mapa de Ligação recuperado para tamanho da população (N=1000) e nível de informatividade (s = 5)..... 90

Figura 24. Mapa de Ligação recuperado para tamanho da população (N=1000) e nível de informatividade (s = 6)..... 91

Tabela 1. Tipos de cruzamentos quanto ao grau de informação na progênie..... 17

Tabela 2. Número de alelos, freqüência alélica e valor de PIC associado para as diferentes populações segregantes de meios-irmãos simuladas..... 33

Tabela 3. Teste do χ^2 para segregação das marcas C11 e C12 para N = 50, 200 e 1000 segundo os diferentes níveis de informatividade alélica (s = 2, 3, 4, 5 e 6)..... 41

Tabela 4. Segregação média de locos únicos e freqüências de genótipos informativos f (Ai- + Aj-) e não informativos f (AiAj) para uma repetição simulada tamanho populacional (N) e informatividade (s)..... 43

Tabela 5. Freqüência de genótipos informativos em cruzamentos exogâmicos para um loco único para acasalamentos com um genitor P_1 heterozigoto A_iA_j..... 44

Tabela 6. Freqüência de gametas recombinantes e parentais em progênies de Famílias de Meios-Irmãos..... 47

Tabela 7. Segregação conjunta das marcas C11 e C12 para N = 50, 200, 1000 para diferentes níveis de informatividade alélica (s = 2, 3, 4, 5 e 6) e tamanho efetivo da população de mapeamento em progênies de FMI..... 51

Tabela 8. Tamanho dos grupos ligação para alta (5 cM), média (10 cM) e baixa (15 cM) saturação para o tamanho populacional (N) e informatividade (S). Médias seguidas de mesma letra em cada coluna para cada valor de s não diferem significativamente pelo teste Tukey a 5 %..... 59

Tabela 9. Variância entre marcas adjacentes para alta (5 cM), média (10 cM) e baixa (15 cM) saturação para o tamanho populacional (N) e informatividade (s). Médias seguidas de mesma letra em cada coluna para cada valor de s não diferem significativamente pelo teste Tukey a 5 %..... 68

Tabela 10. Estresse médio percentual para alta (5 cM), média (10 cM) e baixa (15 cM) saturação para o tamanho populacional (N) e informatividade (S). Médias seguidas de mesma letra em cada coluna para cada valor de s não diferem significativamente pelo teste Tukey a 5 %..... 72

Tabela 11. Correlação de Spearman para alta (5 cM), média (10 cM) e baixa (15 cM) saturação para o tamanho populacional (N) e informatividade (S). Médias seguidas de mesma letra em cada coluna para cada valor de s não diferem significativamente pelo teste Tukey a 5 %..... 75

LISTA DE ABREVIATURAS

EST: *Expressed Sequence Tags*

RAPD: *Random Amplified Polymorphic DNA*

AFLP: *Amplified Fragment Length Polymorphism*

RFLP: *Restriction Fragment Length Polymorphism*

SSR: *Simple Sequence Repeat*

SNP: *Single Nucleotide Polymorphisms*

PCR: *Polymerase Chain Reaction*

QTL: *Quantitative Trait Loci*

RC: Retrocruzamentos

DH: Duplo-Haplóides

RIL: *Recombinant Inbred Lines* (Linhagens Endogâmicas Recombinantes)

FIC: Famílias de Irmãos Completos

FMI: Famílias de Meio Irmãos

NIL: *Near Isogenic Lines* (Linhagens Quase Isogênicas)

PIC: *Polymorphism Information Content* (Conteúdo de Informação Polimórfica)

LIC: *Linkage Information Content* (Conteúdo de Informação de Ligação)

RESUMO

GOOD GOD, Pedro Ivo Vieira, D. Sc., Universidade Federal de Viçosa, julho de 2008. **Mapeamento Genético em Famílias de Meio-Irmãos por Simulação Computacional.** Orientador: Everaldo Gonçalves de Barros. Co-Orientadores: Cosme Damião Cruz e Maurílio Alves Moreira.

O presente trabalho teve como objetivo central estudar a construção de mapas genéticos em Famílias de Meios-Irmãos (FMI) por meio de simulação computacional para elucidar suas limitações e vantagens na análise genômica. Para esse fim, os seguintes cenários de simulação foram empregados: (1) Estudo dos efeitos do tamanho populacional em progênies de Meios-Irmãos para N = 50, 100, 200, 300, 500 e 1000; (2) Estudo dos efeitos dos graus de saturação (5 cM, 10 cM e 15 cM) do genoma com marcas moleculares; (3) Estudo dos efeitos dos níveis de informação alélica ou o conteúdo de informação polimórfica (PIC), com s = 2, 3, 4, 5 e 6 alelos equi-freqüentes presentes na população de mapeamento. Os diferentes cenários foram combinados entre si e simulados em 100 repetições, gerando um total de 3000 populações para análise. Foram utilizados os seguintes critérios para a determinação de acurácia do mapeamento genético: (i) número de grupos de ligação recuperados; (ii) tamanhos dos grupos de ligação; (iii) distâncias médias entre marcadores adjacentes nos grupos de ligação; (iv) variâncias das distâncias entre marcas adjacentes nos grupos de ligação;(v) estresse; e, (vi) correlação de Spearman. Verificou-se que, aliado ao tamanho populacional e o grau de saturação, a informatividade teve papel fundamental no mapeamento em FMI. Populações com N = 50 e 100 não são recomendadas para o mapeamento em FMI, para quaisquer níveis de polimorfismo. Quando o nível de polimorfismo na população é mínimo (s = 2), tamanhos populacionais de N < 300 não são suficientes para a obtenção de mapas fidedignos, principalmente para mapas com baixo grau de saturação, quando ocorre um aumento relativo de inversões de marcas no mapa de ligação. Mapas menos saturados apresentaram menor recuperação de genomas, menores estimativas de correlação de Spearman e maiores variâncias das distâncias entre marcas adjacentes, principalmente para baixo nível de polimorfismo. Como conclusão

geral, pode-se dizer que não é recomendável o uso de populações com N = 50 e 100 mesmo com alto nível de polimorfismo. Para N = 200 é possível obter mapas com certa fidelidade desde que o número de alelos segregando na população base seja igual ou maior do que 4, à semelhança do que ocorre em Famílias de Irmãos Completos para acasalamentos entre genitores completamente informativos.

ABSTRACT

GOOD GOD, Pedro Ivo Vieira, D. Sc., Universidade Federal de Viçosa, July, 2008. **Genetic mapping in half-sib families by computer simulation.** Adviser: Everaldo Gonçalves de Barros. Co-Advisers: Cosme Damião Cruz and Maurílio Alves Moreira.

This study aimed at investigating the genetic mapping in half-sib families (HSF) through computer simulation to elucidate its limitations and advantages in genomic analysis. The following scenarios of simulation were used: (1) Study the effects of population size in HSF for N = 50, 100, 200, 300, 500 and 1000; (2) Study the effects of saturation degrees (5 cM, 10 cM and 15 cM) of the genome with molecular markers; (3) Study the effects of levels of information by polymorphism information content (PIC), with s = 2, 3, 4, 5 and 6 alleles equi-frequents in the mapping populations. All the possible combinations among these various scenarios were simulated in 100 repetitions each. This generated a total of 3000 populations for analysis. We used the following criteria for determining the accuracy of genetic mapping: (i) number of linkage groups recovered, (ii) sizes of linkage groups, (iii) average distance between adjacent markers in linkage groups, (iv) variance of the distances between adjacent markers in the linkage groups, (v) stress, and (vi) Spearman's correlation. The results show that besides population size and degree of saturation, PIC has a critical role in the mapping of HSF. Populations with N = 50 and 100 are not recommended for the mapping in HSF, for all levels of PIC. When the level of PIC in the population is minimal (s = 2), population sizes of N <300 are not sufficient to obtain reliable maps, especially for those with a low degree of saturation, in which case there occurs an increase in inversions of molecular markers in linkage map. Maps less saturated showed lower genome recovery, lower estimates of correlation of Spearman and higher variance of distances between adjacent markers, especially for low level of PIC. As a general conclusion, we do not recommended the use of populations with N = 50 and 100 even with a high level of PIC. For N = 200 one can get maps with some fidelity provided that the number of segregating alleles in the base population is

equal to or greater than 4, similarly to that which occurs in Full-Sib Families from matings between fully informative parents.

1. INTRODUÇÃO

A análise de dados genômicos é considerada uma das grandes vertentes da genética moderna. Podem ser considerados como dados genômicos aqueles oriundos de bancos de dados genéticos tais como seqüências de nucleotídeos, bancos de cDNA, ESTs e *microarrays*, além de dados provenientes de marcadores polimórficos de DNA tais como RAPD, AFLP, RFLP, SSR e SNP. Dada à grande quantidade de informação existente com respeito a estes dados, torna-se imprescindível o desenvolvimento e a validação de ferramentas e metodologias adequadas para a sua análise. Tais ferramentas são desenvolvidas com base em modelos genéticos e fundamentadas a partir de pressuposições matemático-estatísticas.

Na genética moderna, o ramo que se ocupa com a sondagem e a validação de conceitos estatísticos na análise de genomas é conhecido como *estatística genômica* (LIU, 1998). De acordo com SCHUSTER & CRUZ (2004), genômica é a denominação dada à ciência que estuda o genoma de forma completa e que integra as áreas da genética mendeliana, citogenética, genética molecular, genética de populações, genética quantitativa e a bioinformática. Mais especificamente, a área tratada como *genômica clássica* envolve o estudo de marcadores genéticos e suas aplicações na construção de mapas genéticos, como a análise de ligação, o ordenamento de genes e a formação de grupos de ligação, o mapeamento de locos para características qualitativas e quantitativas e a seleção assistida por marcadores (SAM).

O estudo de genomas inteiros, com a utilização de marcadores genéticos na análise de caracteres quantitativos, teve início a partir do século XX, quando foram estabelecidos os primeiros mapas genéticos intra-específcos em *Drosophila melanogaster* (PAYNE, 1918) e em *Phaseolus vulgaris* (SAX, 1923). O conceito de mapa genético permitiu uma visão global do arranjo linear de genes e locos genéticos nos grupos de ligação, que por sua vez correspondiam ao arranjo linear das seqüências nucleotídicas nos cromossomos.

O conceito básico para o estabelecimento de mapas genéticos encontra-se na teoria de ligação ou da segregação não independente entre locos que estão proximamente localizados em um cromossomo (LIU, 1998). Estabelecer grupos de ligação com apenas dois locos tornou-se, a princípio, tarefa de relativa facilidade. No entanto, para o estabelecimento de mapas genéticos mais complexos, os procedimentos são computacionalmente dispendiosos e exigem um aparato estatístico mais avançado.

Dentre as principais dificuldades encontradas para a confecção dos primeiros mapas genéticos, pode-se destacar a limitação em se obter marcadores genéticos consistentes e adequados para a análise de ligação. Os marcadores morfológicos foram os primeiros a serem utilizados para esse fim. Embora os marcadores morfológicos tenham possibilitado o desenvolvimento de mapas genéticos no início do século passado, eles apresentam uma série de restrições para este tipo de abordagem, incluindo baixo nível de polimorfismo, pouca estabilidade ambiental, e número limitado de locos disponíveis para estudos de mapeamento (FERREIRA & GRATTAPAGLIA, 1995).

Este panorama mudou com o desenvolvimento de marcadores moleculares. Os primeiros marcadores moleculares desenvolvidos concentraram-se na detecção do polimorfismo de isoenzimas. Em seguida, marcadores moleculares baseados na análise direta de seqüências polimórficas de DNA foram desenvolvidos. O marcador RFLP, baseado na análise de fragmentos de restrição, foi um dos primeiros a ser utilizado no mapeamento de características de interesse e no estabelecimento de mapas genéticos (FERREIRA & GRATTAPAGLIA, 1995).

A partir do estabelecimento da técnica de PCR, uma grande variedade de marcadores moleculares foi desenvolvida. Os marcadores RAPD, AFLP, SSR e SNP são atualmente os mais utilizados, muito embora diferentes tipos de marcadores estejam em desenvolvimento constante. A grande vantagem dos marcadores moleculares é possibilitar a detecção de um alto nível de polimorfismo se comparados aos marcadores morfológicos, além da estabilidade genética inerente destes marcadores. Em paralelo aos avanços na tecnologia de marcadores moleculares, o desenvolvimento de softwares

analíticos e a proliferação de metodologias genético-estatísticas permitiram o desenvolvimento de mapas genéticos cada vez mais fidedignos.

No melhoramento genético de plantas e animais, os mapas genéticos podem ser utilizados como ferramentas poderosas para a análise genômica. Mapas genéticos podem ter importantes aplicações práticas no melhoramento, através da detecção e mapeamento de regiões e locos controladores de características qualitativas e/ou QTLs (LANDER & BOTSTEIN, 1989). Através da seleção assistida por marcadores moleculares (SAM) é possível aumentar a acurácia da seleção, principalmente para características de baixa herdabilidade, difícil mensuração ou fortemente influenciadas por fatores ambientais. Atualmente, a detecção de locos controladores de características de interesse também está associada a estudos de clonagem e caracterização de genes, através do mapeamento fino de regiões genômicas, além da identificação de genes candidatos, relacionados a rotas metabólicas específicas.

Para o mapeamento genético, os delineamentos genéticos-experimentais utilizados podem ser classificados de acordo com o tipo de amostragem feita em uma população base. Dados derivados da geração F_1 heterozigota, oriunda do cruzamento entre linhagens divergentes, produzem os chamados cruzamentos controlados ou populações endogâmicas, que incluem os retrocruzamentos, a geração F_2 e F_n, RILs, NILs e o uso de duplo-haplóides via cultura de anteras. Dados obtidos de famílias derivadas de genitores amostrados a partir de uma população referência, incluindo famílias de meios-irmãos e irmãos completos, são denominados de populações exogâmicas ou populações naturais.

O mapeamento genético em genealogias exogâmicas apresenta determinadas peculiaridades em relação aos cruzamentos gerados a partir de linhagens endogâmicas (MALIEPAARD *et al.* 1997). Em populações segregantes derivadas de linhagens endogâmicas todos os locos estarão segregando para apenas dois alelos. Com relação à fase de ligação do duplo heterozigoto, esta pode ser claramente determinada com base na análise da segregação dos

gametas recombinantes na população ou diretamente através da genotipagem dos genitores.

Contrariamente, na descendência de cruzamentos entre dois indivíduos não idênticos de uma população exogâmica o número de alelos segregando por loco marcador poderá variar além de dois, não sendo também uniforme o número de alelos para diferentes locos em uma mesma análise. Da mesma forma, a fase de ligação usualmente é desconhecida, sendo possível determiná-la apenas de forma estocástica (MALIEPAARD *et al.* 1997).

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 Meios-Irmãos e Famílias de Irmãos Completos.

O presente trabalho teve como objetivo central estudar a construção de mapas genéticos em Famílias de Meios-Irmãos por meio de simulação computacional para elucidar suas limitações e vantagens em diferentes cenários. Determinados aspectos foram abordados de acordo com os objetivos específicos dispostos a seguir:

(1) Estudar os efeitos do tamanho populacional;
(2) Estudar os efeitos dos níveis de saturação do genoma com marcas;
(3) Estudar os efeitos dos níveis de informação alélica – através do conteúdo de informação polimórfica (PIC) (BOTSTEIN *et al.* 1990) – na população de mapeamento;
(4) Determinar o relacionamento entre diferentes variáveis que medem a acurácia de mapas genéticos nas condições simuladas;
(5) Comparar a acurácia e condições de mapeamento em Famílias de Meios-Irmãos com outros delineamentos experimentais.

2. REVISÃO DE LITERATURA

2.1. Construção de Mapas Genéticos de Ligação

A análise genômica exige elevado número de marcadores genéticos. Na construção de mapas de ligação estão envolvidas várias etapas metodológicas necessárias para a sua obtenção (Figura 1). O ponto inicial (i) é a determinação de qual delineamento genético-experimental será utilizado para a detecção de ligação entre marcas. A discussão com relação aos diferentes tipos de delineamentos experimentais que podem ser utilizados e suas implicações no mapeamento genético será abordada de forma pormenorizada no próximo tópico. Na segunda etapa (ii) são obtidos os escores dos marcadores moleculares nos indivíduos da população de mapeamento, sendo fundamental levar em conta o conhecimento de técnicas de genética molecular e as bases genéticas de funcionamento dos diferentes tipos de marcadores moleculares, conforme discutido mais pormenorizadamente em FERREIRA & GRATTAPAGLIA (1995) e .

Na terceira etapa (iii), também denominada de análise de loco único ou modelo de loco único, são determinadas as freqüências alélicas e genotípicas dos marcadores na população de mapeamento. Em populações exogâmicas, o teste de equilíbrio de Hardy-Weinberg e o valor do conteúdo de informação polimórfica (PIC) podem ser importantes para a caracterização individual dos locos marcadores. Como será destacado ao longo deste trabalho, em Famílias de Meios-Irmãos, o valor do PIC assume papel relevante na determinação da qualidade dos dados para mapeamento.

Ainda na terceira etapa, o teste de χ^2 é realizado para avaliar a segregação individual das marcas (SCHUSTER & CRUZ, 2004). O objetivo do teste é verificar se as freqüências genotípicas observadas para cada loco marcador seguem um padrão de segregação mendeliano. A ocorrência de segregação distorcida ou não mendeliana pode indicar que o modelo de segregação adotado é inapropriado, que os dados são de baixa qualidade ou

que o processo de amostragem foi não aleatório (LIU, 1998; FERREIRA, 2006). A utilização de marcas com segregação distorcida proporciona estimativas viesadas das distâncias e reduz a acurácia de mapas genéticos, sendo recomendada a sua eliminação na análise de ligação (KAO *et al.* 1999) ou a adoção de procedimentos biométricos apropriados para a sua utilização, como proposto por FERREIRA (2006).

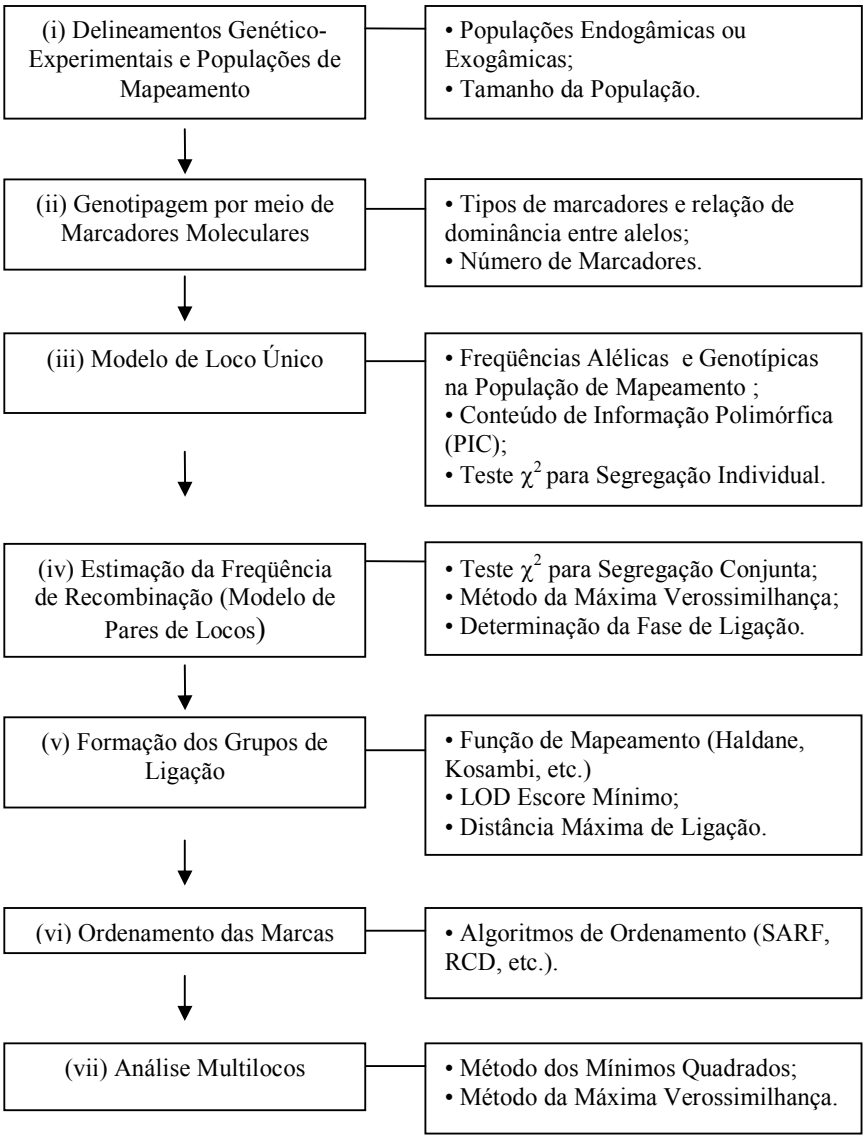


Figura 1. Principais etapas metodológicas utilizadas na construção de mapas genéticos

Após a análise individual de marcas, é necessário verificar a hipótese de ligação gênica entre pares de marcas e efetuar o cálculo das distâncias entre marcas (iv). O teste apropriado para verificar a hipótese de ligação é o teste χ^2 conjunto. Entretanto, este teste é apenas qualitativo e quando detectada a evidência de ligação deve-se calcular a porcentagem de recombinação entre os pares de marcadores. Como citado por ROCHA *et al.* (2004), embora outras metodologias possam ser utilizadas para obter estimativas das freqüências de recombinação, o Método da Máxima Verossimilhança é o mais utilizado para a análise genética de ligação. 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 & CRUZ, 2004).

A formação de grupos de ligação (v) é feita com base em um processo iterativo que utiliza como critérios a freqüência máxima de recombinação (r_{max}) e o LOD escore mínimo (LOD_{min}) para inferir se dois locos estão ligados. Com base na matriz de distâncias entre pares de marcas e no valor de LOD escore relativo ao cálculo da distância, os marcadores são agrupados em um processo em que se utiliza a propriedade transitiva de ligação em que: se o loco A está ligado ao loco B e este estiver ligado ao loco C então A estará ligado a C independente da distância entre A e C. Embora sejam critérios empíricos, a freqüência máxima de ligação e o LOD escore têm sido sistematicamente utilizados para a construção de grupos de ligação. Segundo SCHUSTER & CRUZ (2004) a utilização do r_{max} e LOD_{min} apresenta a vantagem de ser possível obter um padrão de agrupamento único e consistente, de modo que cada marcador se apresente em apenas um único grupo de ligação ou esteja segregando de forma independente em relação aos demais marcadores. Destaca-se ainda que, 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, cobertura

insuficiente de marcadores no genoma ou a utilização de um número reduzido de indivíduos (SCHUSTER & CRUZ, 2004).

Outro 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 as de HALDANE (1919) e 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 marcas, resolvendo o problema da aditividade. Cabe ressaltar que diferentes funções de mapeamento correspondem a diferentes graus de interferência assumidos na permuta entre regiões adjacentes (SCHUSTER & CRUZ, 2004). A função de Haldane é obtida considerando interferência nula. A função de Kosambi admite a existência de interferência.

Após a formação preliminar dos grupos de ligação, uma questão que surge é verificar se os marcadores encontram-se na melhor ordem possível. Nesta etapa (vi) o melhor ordenamento pode ser obtido por meio de critérios e algoritmos de ordenamento. O principal critério utilizado para se identificar a melhor ordem consiste em se adotar a ordem que proporciona a menor soma das distâncias, conhecido como SARF (*Sum of Adjacent Recombination Fractions*). Um dos processos para se identificar a menor soma é gerar todas as ordens possíveis entre marcadores no mesmo grupo de ligação. Entretanto, embora a análise de todas as ordens seja conclusiva, ela se torna inviável quando o número de marcadores se torna excessivamente grande. Um algoritmo que contorna este problema foi desenvolvido por DOERGE (1996), denominado de RCD (*Rapid Chain Delineation*). Neste método, permutas sucessivas são realizadas agrupando-se e invertendo (permutas) intervalos de dois, três ou quatro marcadores. A ordem será alterada se, após a permuta, a SARF for reduzida. Neste caso, a permuta é dita eficiente e o processo é reiniciado com base na nova ordem. Se a permuta não reduz a SARF, ela é dita ineficiente e o processo prossegue baseando-se na ordem inicial,

avançando sucessivamente em permutas duplas, triplas e quádruplas. Com base neste método é possível reduzir consideravelmente o número de ordens geradas para a obtenção do ordenamento ótimo entre marcas, em um mesmo grupo de ligação.

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, resultando em uma análise única para cada grupo de ligação (LYNCH & WALSH, 1998). A vantagem de se utilizar análise multilocos reside na obtenção de estimativas mais precisas das distâncias entre marcas, uma vez que se utiliza simultaneamente a informação de todos os locos envolvidos. Para a obtenção das estimativas de distâncias com base em modelos multilocos são empregados o Método dos Mínimos Quadrados ou o Método Máxima Verossimilhança. 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 à respeito das distribuições de permuta, do grau de interferência e do comprimento do segmento cromossômico analisado, conforme descrito por SCHUSTER & CRUZ (2004) e LIU (1998).

2.2. Delineamentos Experimentais no Mapeamento Genético e na Análise de QTL

Uma população utilizada para o mapeamento genético e a detecção de QTL é denominada de população de mapeamento. De acordo com a literatura, as populações de mapeamento são organizadas de acordo com a amostragem genética feita para se detectar gametas recombinantes e a maneira como o desequilíbrio de ligação é gerado entre o marcador e o QTL (LYNCH & WALSH, 1998; WELLER, 2001). Assim, as populações de mapeamento são classificadas em dois tipos: (1) Populações segregantes derivadas da geração F_1 proveniente de cruzamentos entre linhagens endogâmicas divergentes – denominadas de populações derivadas de cruzamentos controlados ou

populações endogâmicas; (2) Populações segregantes derivadas de amostragens tomadas a partir de uma população base, sem a utilização da geração F_1 e linhas divergentes – denominadas de populações naturais ou exogâmicas (Figura 2).

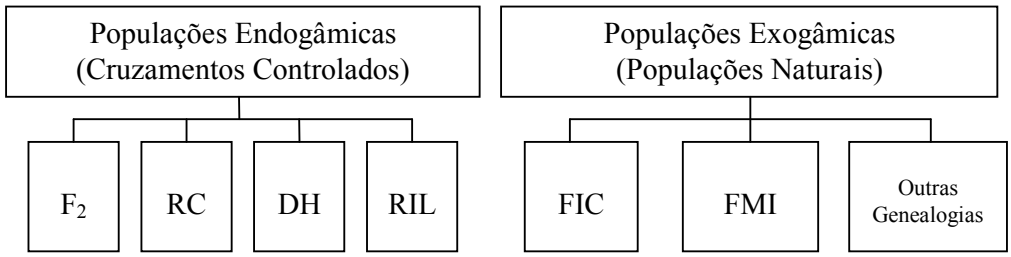


Figura 2. Populações de mapeamento e delineamentos experimentais utilizados no mapeamento genético e na análise de QTL. Adaptado de WELLER (2001)

2.2.1. Populações Endogâmicas

As populações derivadas de cruzamentos controlados apresentam como característica marcante o cruzamento entre linhagens endogâmicas divergentes. Dessa forma é obtida uma geração F_1 heterozigota para todos os locos marcadores. Nessas condições os genótipos dos parentais e a fase de ligação podem ser determinados precisamente, além do número de alelos segregantes ser fixado em dois (MALIEPAARD *et al.* 1997).

Cruzamentos entre linhagens endogâmicas oferecem a condição ideal para a detecção e o mapeamento de QTLs. Essa condição é dada por meio de uma F_1 idêntica geneticamente em todos os indivíduos e por apresentar, em conseqüência, desequilíbrio de ligação completo em todos os genes que diferem entre as duas linhagens. Dessa forma, a associação entre os marcadores genéticos e a característica pode ser estudada pela comparação do desempenho fenotípico de indivíduos provenientes de gerações subseqüentes à F_1. Em adição, tais dados são obtidos de delineamentos experimentais que podem ser considerados como provenientes de uma única grande família, pois todos os indivíduos partilham os mesmos genótipos dos parentais. Como resultado, o efeito de substituição do QTL e a dominância

podem ser diretamente estimados. Segundo XU & ATCHLEY (1995) o modelo linear que descreve tais conjuntos de dados é chamado de modelo fixo.

A partir da geração F_1, diferentes delineamentos genéticos podem ser utilizados, destacando-se os retrocruzamentos, duplo-haplóides, RILs, NILs, geração F_2 e F_n. Assim, o uso de populações endogâmicas é mais aplicável em espécies autógamas, principalmente em plantas, casos em que é possível a obtenção de famílias numerosas, derivadas de autofecundação (MALIEPAARD *et al.*, 1997).

2.2.2. Populações Exogâmicas

Para muitas espécies, como espécies florestais ou de ciclo longo ou mesmo a espécie humana ou espécies alógamas, a facilidade de obtenção e manipulação de linhagens endogâmicas a partir de cruzamentos controlados está fora de alcance. Em tais casos deve-se então voltar para a análise e utilização de populações exogâmicas.

As populações de mapeamento derivadas de cruzamentos exogâmicos são obtidas por meio da escolha de genitores a partir de uma população segregante (Figura 3). A descendência obtida dos genitores de cruzamentos exogâmicos é estruturada em famílias para a análise de ligação genética e a detecção de QTLs. Se indivíduos são amostrados de maneira aleatória diretamente da população segregante é bem provável que o loco marcador e o QTL estejam em equilíbrio de ligação. Entretanto, se os indivíduos são organizados em famílias derivadas dos genitores, o desequilíbrio de ligação é garantido dentro das famílias.

Dentre os delineamentos experimentais utilizados para o mapeamento de QTL em cruzamentos exogâmicos pode-se destacar o uso de famílias de irmãos completos e de famílias de meios-irmãos. De acordo com WELLER (2001) três tipos básicos de delineamentos experimentais são propostos: (1) o delineamento de pares de irmãos (*sib-pair analysis*) que emprega a análise de várias famílias de irmãos completos, pouco numerosas, que tem sido aplicado em larga escala na espécie humana; (2) o delineamento de irmãos completos,

utilizado para a análise de famílias de irmãos completos com progênie numerosa; e, (3) o delineamento de meio-irmãos, também conhecido como delineamento de filhas (*daugther design*) no melhoramento animal, que emprega a análise de famílias de meio-irmãos de grande tamanho e que tem sido largamente utilizado no mapeamento de QTLs em gado de leite.

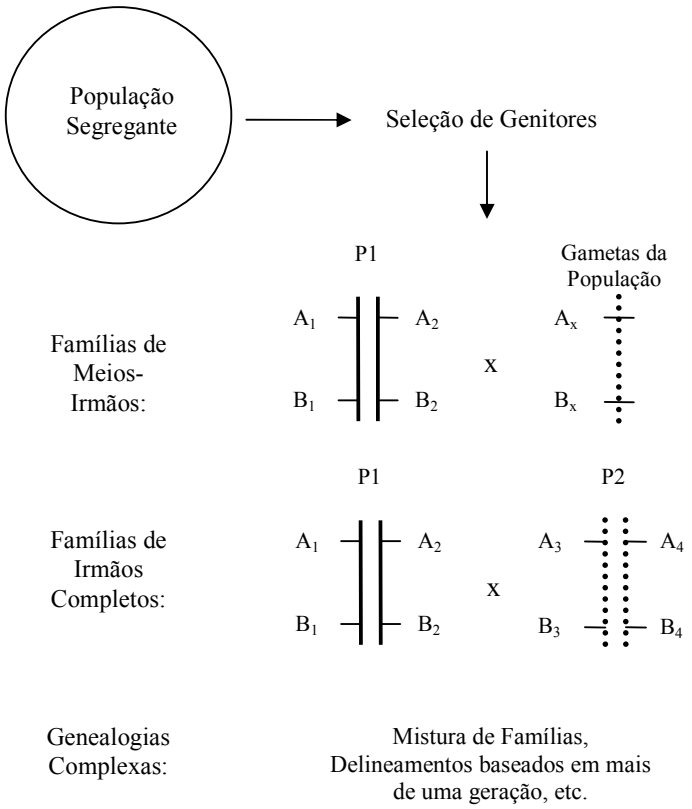


Figura 3. Esquema demonstrando o processo de obtenção de delineamentos experimentais utilizados na construção de mapas genéticos e na análise de QTL em cruzamentos exogâmicos

No melhoramento de plantas, famílias de meio-irmãos e de irmãos completos prestam-se não somente a estudos genômicos, mas também podem ser utilizadas como populações de melhoramento. Em geral, no melhoramento de plantas, a utilização de famílias de meio-irmãos é obtida por meio da descendência de uma árvore ou planta matriz, quando o enfoque está em

espécies frutíferas ou perenes, ou em espécies que apresentam alogamia. Alternativamente, a utilização de famílias de irmãos completos tem grande potencial de uso em plantas perenes, pelo acasalamento de clones elite.

Pode-se considerar que a utilização de uma única e numerosa família informativa é a condição ideal para se detectar QTLs, entretanto, quando isso não é possível pode-se utilizar a informação combinada de várias famílias para se proceder a análise, como utilizado na análise de pares de irmãos (WELLER, 2001).

O delineamento conhecido como pseudocruzamento-teste foi utilizado com grande impacto pela primeira vez no trabalho de GRATTAPAGLIA & SEDEROFF (1994). Este delineamento baseia-se no uso de uma única família de irmãos completos derivada de genitores heterozigóticos. Entretanto, nesta estratégia as meioses são analisadas separadamente para cada genitor, de maneira que mapas genéticos diferentes são obtidos quando considerados os diferentes genitores.

Genealogias mais complexas podem ser utilizadas na análise de QTLs em cruzamentos exogâmicos. Em alguns casos, o uso de mais de uma geração permite aumentar a precisão experimental no mapeamento de QTL. O delineamento de netas (*grand-daugther design*), como é conhecido no melhoramento animal, utiliza a primeira geração de meio-irmãos para se realizar a genotipagem dos indivíduos por meio de marcadores moleculares e, na segunda geração familiar, ou netos, são avaliadas as características de interesse. Este procedimento permite uma melhor avaliação das características quantitativas, levando a estimativas mais precisas na detecção de QTLs (WELLER, 2001).

As populações exogâmicas são também de interesse por razões outras que as limitações experimentais. Por exemplo, QTLs detectados por cruzamentos controlados usualmente representam diferenças fixas entre linhas, ou muitas vezes entre espécies, e a relevância destes resultados para QTLs segregando dentro de populações permanece incerta. Esta fundamental distinção de que os cruzamentos controlados detectam QTLs responsáveis por diferenças entre populações enquanto cruzamentos exogâmicos detectam

QTLs responsáveis por variação dentro de populações tornam tais abordagens mais complementares do que competitivas (LYNCH & WALSH, 1998).

A utilização da variação dentro da população, em oposição às diferenças fixas entre populações, resulta em uma redução significante do poder de detecção de QTLs. Com linhagens endogâmicas, todos parentais F_1 têm genótipos idênticos (incluindo a mesma fase de ligação), dessa forma todos os indivíduos são informativos, e o desequilíbrio de ligação é maximizado. Adicionalmente, os efeitos dos QTLs são expressos como médias (o valor médio de cada genótipo do QTL). Em contrapartida, os efeitos dos QTLs são expressos como variâncias em cruzamentos exogâmicos. Uma vez que a variância é estimada com precisão menor do que a média, as estimativas para populações exogâmicas são tidas como menos precisas (LYNCH & WALSH, 1998).

2.3. Informatividade no Mapeamento Genético e na Análise de QTL

A importância da aplicação de delineamentos experimentais para estudos de ligação reside na obtenção do máximo conteúdo de informação das estimativas de ligação de acordo com os recursos disponíveis (DA & LEWIN, 1995). A principal diferença que pode ser estabelecida entre cruzamentos controlados e cruzamentos exogâmicos é a uniformidade genética dos indivíduos que formam a geração parental. Nos cruzamentos controlados são utilizados indivíduos contrastantes entre si, porém, geneticamente uniformes. Nos cruzamentos exogâmicos a geração parental é geneticamente variável, pois é oriunda da amostragem feita a partir de uma população de referência. Esta distinção tem muitas conseqüências, como já mencionadas. No entanto, pode-se destacar o conteúdo de informação obtido nos cruzamentos exogâmicos. Se nas populações derivadas de linhagens endogâmicas o conteúdo de informação é o mesmo para um determinado delineamento experimental, nas populações exogâmicas apenas uma fração dos cruzamentos efetuados podem ser totalmente informativos (LYNCH & WALSH, 1998).

De acordo com HASEMAN & ELSTON (1972) e MARTINEZ & VULKANOSIVIC (2000), 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. 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 de 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 (MARTINEZ & VULKANOSIVIC, 2000). 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.

Além da heterozigosidade observada entre genitores de um cruzamento exogâmico, deve ser verificada adicionalmente a existência de indivíduos informativos na prole. Segundo DA & LEWIN (1995) um marcador é considerado informativo quando a origem dos alelos nos indivíduos em uma progênie pode ser inequivocamente determinada. Em outras palavras, um marcador é

considerado informativo quando é possível associar os alelos da progênie a cada genitor que lhe originou.

Para que um genitor seja informativo para a análise de QTL, ele deve ser heterozigoto tanto para o marcador quanto para um QTL ligado a esse marcador. Somente nesta situação a associação marcador-QTL pode ser detectada na progênie (BEARZOTI, 2003). Assim, apenas uma fração da amostra da geração parental em cruzamentos exogâmicos deve observar a condição de duplos heterozigotos. Ao contrário, em cruzamentos controlados, a geração F_1 é heterozigota para todos os locos que diferem entre as linhas parentais, o que implica na informatividade total dos cruzamentos.

Um segundo ponto a ser destacado é de que existem apenas dois alelos segregando em qualquer loco nas populações derivadas de cruzamentos controlados, ao passo que em cruzamentos exogâmicos podem estar segregando até quatro alelos em um determinado loco. Adicionalmente, nas populações exogâmicas, os indivíduos podem diferir quanto à fase de ligação entre o marcador e o QTL. Assim, em um parental portador da marca A_i pode estar associado o alelo Q_i do QTL. De outro modo, um outro indivíduo poderá produzir gametas com a associação $A_i - Q_j$, caracterizando uma mistura de fases de ligação. Assim, em populações exogâmicas o estudo da associação marcador-QTL deve ser efetuado examinando separadamente cada cruzamento parental reduzindo poder e precisão no mapeamento (BEARZOTI, 2003).

Em cruzamentos exogâmicos – para a análise de um loco único – podem ser estabelecidas três categorias de acasalamentos segundo o grau de informatividade nas progênies (LYNCH & WALSH, 1998) (Tabela 1): (1) famílias derivadas de cruzamentos completamente informativos; (2) famílias derivadas de 'retrocruzamentos' ou família com loco informativo; e (3) famílias derivadas de 'intercruzamentos' ou com loco parcialmente informativo. Deve-se ressaltar que o grau de informação pode se referir ao marcador, ao QTL ou a ambos. Para que uma quantidade mínima de indivíduos seja informativa quanto ao marcador, é necessária a utilização de marcadores altamente polimórficos. Embora em algumas situações seja possível escolher indivíduos informativos

quanto aos marcadores, não há garantias de que esses indivíduos sejam informativos quanto aos QTLs.

Como descrito por LYNCH & WALSH (1998) uma família completamente informativa é derivada do cruzamento A_iA_j vs. A_kA_l. Nesse tipo de cruzamento, os genitores são heterozigotos para diferentes alelos, possibilitando que a progênie seja completamente informativa na distinção da origem destes alelos. Assim, é possível comparar os valores da característica na progênie por meio do contraste A_i vs. A_j e do contraste A_k vs. A_l. Nesse tipo de cruzamento a razão esperada é de 1:1:1:1.

Tabela 1. Tipos de cruzamentos quanto ao grau de informação na progênie

Categoria	Cruzamento	Característica
Completamente Informativo	A_iA_j x A_kA_l	Toda progênie é informativa na distinção dos alelos de cada genitor
Retrocruzamento	A_iA_j x A_kA_k	A progênie é informativa somente na distinção dos alelos do genitor heterozigoto
Intercruzamento	A_iA_j x A_iA_j	Apenas indivíduos homozigotos são informativos na progênie

Fonte: LYNCH & WALSH (1998)

Em uma família de retrocruzamento (A_iA_j vs. A_kA_k) apenas os alelos de um único genitor podem ser distinguidos. Assim, apenas o contraste entre A_i vs. A_j pode ser determinado na progênie. Nesse caso, a razão de segregação esperada é de 1:1.

No último caso, em famílias de intercruzamentos (A_iA_j vs. 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 razão de segregação esperada para estes cruzamentos é de 1:2:1.

O caso especial em que ambos os genitores podem ser genotipados refere-se apenas a Delineamentos de Irmãos Completos. Entretanto, em Delineamentos de Meios-Irmãos, apenas o genitor comum pode ser genotipado. Assim, apenas genitores heterozigotos do tipo A_iA_j são informativos para a análise de ligação e detecção de QTL. Nesse sentido, a razão de segregação e a informatividade para Meios-Irmãos assemelha-se à

famílias derivadas de retrocruzamentos, como descrito acima. Na descendência de cruzamentos em Famílias de Meios-Irmãos, apenas as progênies A_i- e A_j- serão informativas e as progênies A_iA_j serão não informativas. A razão de segregação esperada em FMI para o genitor informativo é de 1:1.

Na análise de dois de locos em Famílias de Irmãos Completos, diferentes combinações de razão de segregação podem surgir, pois cada marcador poderá segregar de forma diferente. Por corolário, diferentes conteúdos de informação serão obtidos na análise de dois locos. Assim, para cada situação deverá ser utilizada uma função de verossimilhança específica para o cálculo das freqüências de recombinação. Estas funções estão descritas em detalhes no trabalho de BHERING *et al.* (2008). Em Famílias de Meios-Irmãos, a análise de ligação para dois locos envolve apenas a combinação única de segregação do tipo (1:1)(1:1). Ao contrário do observado para em FIC, apenas uma única função de verossimilhança será necessária para o cálculo das freqüências de recombinação.

2.3.1. Medidas de Informatividade

Como mencionado anteriormente, em cruzamentos exogâmicos nem sempre toda a progênie derivada de um acasalamento pode ser útil para a análise de ligação. O conteúdo de informação está relacionado ao número e freqüências dos alelos amostrados, ao sistema de acasalamentos e a estrutura das famílias utilizadas (DA & LEWIN, 1995). Assim, diferentes trabalhos foram desenvolvidos com o objetivo de se determinar o conteúdo de informação derivado de diferentes delineamentos e para diferentes genealogias populacionais (DA & LEWIN, 1995; GUO & ESLSTON, 1999; GUO *et al.* 2002; RIJSDIJK & SHAM, 2002; ROCHA *et al.* 2004).

Várias medidas são utilizadas para se determinar o grau de informatividade em cruzamentos exogâmicos. Neste trabalho será destacado o PIC e o Conteúdo de Informação de Fischer. O PIC ou Conteúdo de Informação Polimórfica (BOTSTEIN *et al.* 1980) permite avaliar o grau de

polimorfismo ou heterozigosidade de um loco único e é calculado com base na freqüência alélica do loco em análise. O PIC apresenta um papel relevante na seleção de genitores e marcadores informativos, sendo mais utilizado para avaliar a qualidade a priori de dados para o mapeamento. Segundo DA & LEWIN (1995), esse índice pode não ter aplicabilidade direta na avaliação do conteúdo de informação em mapas genéticos uma vez que é necessário utilizar pelo menos dois locos para que seja feita a análise de ligação.

O Conteúdo de Informação de Fischer (WEIR, 1996), ou Conteúdo Médio de Informação, corresponde ao negativo da derivada segunda da função de verossimilhança para o cálculo da freqüência de recombinação. O inverso da informação de Fischer fornece a variância da estimativa da freqüência de recombinação. A informação de Fisher é calculada com base no valor da freqüência de recombinação e assume estimadores particulares para diferentes populações de mapeamento, como pode ser constatado no trabalho de ROCHA *et al.* (2004). O Conteúdo de Informação de Fischer tem sido usado sistematicamente para avaliar o grau de informatividade em populações de mapeamento e em trabalhos de simulação para verificar a acurácia de mapas genéticos (LIU, 1998; ROCHA *et al.* 2004).

2.4. Mapeamento de QTL – Definição e Modelo

Um caráter quantitativo é tradicionalmente definido como uma característica que está sob controle de um sistema poligênico e que demonstra variação contínua entre e dentro de populações sob o efeito do ambiente. Devido ao fato de uma grande quantidade das características de importância ser de herança quantitativa, a procura por genes controlando características complexas ou quantitativas é vital no desenvolvimento e aplicação da informação genômica (LIU, 1998). Os locos que controlam características quantitativas e que servem para inferir a respeito da arquitetura genética destas características têm sido comumente referidos como QTL (*Quantitative Trait Loci*). Ao processo de procura e localização de QTL é chamado de mapeamento de QTL.

O mapeamento de QTL envolve a construção de mapas genômicos e a análise de associação entre características e marcadores polimórficos. Uma associação significativa entre as características e os marcadores pode evidenciar a presença de um QTL próximo de um marcador. Dessa forma, o mapeamento de QTL é a combinação do mapeamento de ligação e a análise tradicional da genética quantitativa. O modelo genético utilizado para características quantitativas na análise de QTL emprega uma abordagem um pouco modificada, pois leva em conta a possibilidade de se explicar características complexas ao nível de locos individuais ou segmento cromossômico, ao invés de fazer unicamente o uso de parâmetros como as variâncias entre e dentro de populações (FALCONER & MACKAY, 1996).

A adoção do modelo genético na identificação de QTL permite estudar questões específicas da arquitetura genética dos locos relacionados à característica quantitativa, desde a estimação de parâmetros como médias e variâncias a inferências do tipo de ação gênica, fenômenos de ligação, interações epistáticas e genótipo-ambiente. Embora os genótipos do QTL não possam ser observados, as estimativas dos parâmetros do modelo podem ser estimadas pelo contraste entre médias dos indivíduos nas classes genotípicas dos marcadores (LIU, 1998).

2.5. Métodos de Mapeamento de QTLs em Populações Exogâmicas

Segundo CHURCHILL & DOERGE (1998) há três relevantes hipóteses utilizadas na detecção de QTL, sendo:

H_0^1: não há QTL presente;

H_0^2: o QTL está presente, porém não ligado ao marcador;

H_A: o QTL está presente e ligado ao marcador.

O problema estatístico do mapeamento de QTL pode ser visto sob três aspectos. O primeiro é a detecção dos fatores genéticos que possuem efeitos na característica de interesse e que estão segregando na população. O segundo se refere à localização de QTLs relativos aos locos dos marcadores genéticos. O terceiro diz respeito à estimação dos efeitos dos QTLs e suas

interações. Esses problemas são interdependentes, mas a distinção entre eles é útil para facilitar o entendimento dos procedimentos de inferência utilizados no mapeamento de QTL (CHURCHILL & DOERGE, 1998).

Metodologia de HASEMAN & ELSTON (1972)

A utilização de modelos aleatórios foi proposta, inicialmente, por HASEMAN & ELSTON (1972) para a utilização em conjuntos de dados que correspondem a pares de irmãos, obtidos de uma população de referência. Esta metodologia, baseada em marcas simples, é efetuada por meio de uma regressão linear simples do quadrado das diferenças fenotípicas entre dois irmãos, em função da proporção de genes idênticos por descendência (IBD) do QTL entre os dois irmãos (π_{ij}). O princípio envolvido no método é de que quanto maior π_{ij}, menor será a diferença entre o par de irmãos. Assim, uma associação negativa significativa indica um QTL ligado ao marcador em questão.

O método de pares de irmãos de HASEMAN & ELSTON (1972) é considerado um método robusto tanto no sentido genético quanto no sentido estatístico, uma vez que o modelo genético do QTL não precisa ser conhecido em detalhes, bem como os testes estatísticos não são grandemente alterados quando a distribuição da variância residual não é normalmente distribuída (AMOS & ELSTON, 1989). Entretanto, o método pode ser considerado limitado, pois o efeito genético do QTL e a taxa de recombinação entre o QTL e o loco marcador estão confundidos, de forma que não podem ser estimados individualmente. Assim, o método permite apenas que se detecte a ligação entre o QTL e o marcador, não permitindo, entretanto, estimar se a variação observada é devida a um QTL de grande efeito localizado distante do marcador ou devido a um QTL de pequeno efeito localizado próximo ao marcador (MARTINEZ *et al.*, 1999).

Atualmente, numerosas variações da metodologia de HASEMAN & ELSTON (1972) podem ser encontradas na literatura e incluem: uso de diferenças, uso de somas, uso de somas ao quadrado, uso de observações

individuais. Alternativamente pode-se usar na estimação a metodologia de máxima verossimilhança em substituição ao método dos quadrados mínimos (VISSCHER & HOPPER, 2001).

Metodologia de FULKER & CARDON (1994)

A metodologia de FULKER & CARDON (1994) baseia-se no procedimento de mapeamento por intervalo. Nesta metodologia são utilizados dois marcadores flanqueando o QTL para estimar separadamente a posição do QTL e o seu efeito sobre a característica analisada. Assim, esta metodologia procura separar a variância devido ao QTL e o parâmetro de ligação e, ainda, localizar o QTL em uma posição específica do cromossomo.

Na metodologia de FULKER & CARDON (1994), aplica-se o mesmo princípio de HASEMAN & ELSTON (1972) para cada um dos dois marcadores que flanqueiam o QTL. Assim, são utilizadas duas equações e duas incógnitas devem então ser estimadas. O primeiro passo da análise é estimar a proporção de genes idênticos por descendência do QTL ($\hat{\pi}_q$) em um par de irmãos, a partir da proporção IBD dos marcadores que estão flanqueando o intervalo em que se supõe estar o QTL. No segundo passo da análise, uma regressão é efetuada da maneira que HASEMAN & ELSTON (1972) propuseram, exceto que a proporção de IBD utilizada é aquela estimada no primeiro passo. O mapeamento por intervalo é então conduzido para todos os $n-1$ intervalos, considerando-se um cromossomo com n marcadores.

Embora a metodologia de FULKER & CARDON (1994) apresente maior poder estatístico o método de estimação utilizado é o dos mínimos quadrados, à semelhança da metodologia de HASEMAN & ELSTON (1972), e não otimiza, portanto, a utilização de todas as informações contidas nos dados. Metodologias mais poderosas, como o método da máxima verossimilhança, têm a vantagem de considerar as informações contidas nos dados, com base na pressuposição de suas propriedades distribucionais.

Metodologia de GOLDGAR (1990)

GOLDGAR (1990) desenvolveu um método de IBD múltiplos baseado na máxima verossimilhança para estimar a variância genética causada por uma determinada região do cromossomo. Embora tal método considere as vantagens relacionadas às funções de distribuições dos dados, além de considerar marcadores flanqueadores para definir um segmento cromossômico, ele não se baseia na teoria de mapeamento por intervalo e, portanto, tal método pode apenas estimar a variância devido ao QTL, mas não a sua exata posição (MARTINEZ, 1999).

Metodologia de XU & ATCHLEY (1995)

Para incrementar o poder de detecção do QTL e fornecer estimativas de sua posição, XU & ATCHLEY (1995) desenvolveram o mapeamento por intervalo para a metodologia de GOLDGAR (1990) em famílias de irmãos completos. A metodologia foi desenvolvida com base em um procedimento geral para o mapeamento do QTL, assumindo uma única distribuição normal dos valores genotípicos do QTL e colocando no modelo os efeitos do QTL e o efeito do sistema poligênico (devido aos demais genes) ambos como variáveis ao acaso. Assim, é possível mapear um QTL com sucesso e estimar a sua variância com precisão.

2.6. Eficiência e Poder Estatístico na Estimação no Mapeamento de QTL

Segundo LIU (1998), a inferência estatística inclui testes de hipóteses e a estimação de parâmetros. Para se avaliar o desempenho e a acurácia de uma metodologia de estimação é necessária a análise da qualidade das estimativas fornecidas. Assim, o teste de hipóteses é usualmente considerado como uma inferência qualitativa enquanto que a estimação um processo quantitativo.

Dentre as possíveis maneiras de se avaliar a qualidade das inferências pode-se usar a construção de intervalos de confiança e a determinação do poder do teste estatístico.

O poder do teste é definido como a probabilidade de se rejeitar a hipótese de nulidade quando a hipótese alternativa é verdadeira. Ele também é comumente denominado de poder estatístico. O poder é definido como:

$$\text{Poder} = 1 - \beta$$

onde, β: é a probabilidade de um falso negativo (erro tipo II) ou a probabilidade de aceitar a hipótese de nulidade como verdadeira quando ela é falsa. Ela é a proporção do teste estatístico que cai na região de aceitação quando a hipótese de nulidade é falsa e o teste é repetido inúmeras vezes. Para exemplificar, pode-se considerar como verdadeira a hipótese nula de que não há ligação entre dois marcadores quando na verdade eles estão ligados.

O poder estatístico de um teste é definido em termos da hipótese alternativa enunciada, das condições experimentais e do nível de significância escolhido para o teste (erro tipo I ou probabilidade de falso positivo). Assim, o poder estatístico, sendo um complemento, pode ser considerado como a probabilidade de se rejeitar uma hipótese nula, sendo ela falsa. Na prática, o poder estatístico empírico de detecção de um QTL pode ser definido como o percentual das repetições em que a hipótese nula é rejeitada, dado um nível de significância, quando o teste é repetido inúmeras vezes. Em geral, utiliza-se o poder estatístico para a determinação do tamanho amostral mínimo esperado para certo nível de confiança.

Vários exemplos simulados podem ser extraídos da literatura, os quais utilizam o poder de detecção empírico para quantificar o grau de precisão de uma dada metodologia de estimação, em diferentes condições, no mapeamento de QTL (GOLDGAR, 1990; MARTINEZ *et al.*, 1999; SILVA *et al.*, 2004). Em populações exogâmicas, o poder de detecção é altamente influenciado pelo número de famílias amostradas, pelo tamanho de famílias, pela proporção da variância genética explicada pelo QTL e a sua posição. Além destes fatores, o número de alelos do marcador é um fator que contribui para o poder de detecção. Em geral, pode-se obter maior poder de detecção com o

aumento do número marcadores informativos, do número de famílias e o número de progênies por famílias (SILVA *et al.*, 2004; VAN DER BEEK *et al.*, 1995).

Em relação à proporção da variância genética explicada pelo QTL e a posição do QTL no grupo de ligação, o que se espera é que o poder de detecção seja reduzido, quando inúmeros QTLs de menor efeito estejam condicionando o caráter (GOLDGAR, 1990; SILVA *et al.*, 2004). Em geral, a existência de dois QTLs no mesmo intervalo leva à conclusão da existência de um QTL único no intervalo. Da mesma forma, metodologias que avaliam um intervalo de cada vez, sem considerar o restante da informação do genoma podem apresentar poder de detecção reduzido, especialmente quando houver outros QTLs de grande efeito no mesmo grupo ou em outros grupos de ligação (SCHUSTER & CRUZ, 2004). Assim, caso haja a presença de mais de um QTL no grupo de ligação a posição e os efeitos dos QTLs identificados por metodologias de intervalo simples podem estar viesados e o poder do teste é reduzido.

De acordo com SCHUSTER & CRUZ (2004), o ideal para o mapeamento de QTL é de que os testes realizados em um intervalo sejam independentes dos efeitos de outros QTLs localizados fora deste intervalo. Uma abordagem que considera um teste deste tipo faz uso da combinação do mapeamento por intervalo com a análise de regressão múltipla *stepwise*, denominado mapeamento por intervalo composto (CIM) (JANSEN, 1992, 1993; ZENG, 1993).

2.7. Intervalos de Confiança e Nível de Significância Genômico

A determinação de intervalos de confiança (IC) para um QTL pode ser obtida por meio de métodos de reamostragem como o *bootstrap*, *jacknife*, permuta e métodos de simulação Monte Carlo, facilmente utilizados em conjunto com o mapeamento por intervalo baseado em regressão (LIU, 1998). Tais procedimentos são utilizados principalmente devido à estrutura complexa dos dados genômicos, que não permitem uma estimação paramétrica confiável ou devido ao fato da estatística utilizada não seguir a distribuição padrão.

A amplitude do IC depende do tamanho da população e do efeito do QTL. Em geral espera-se que quanto maior a evidência de um QTL estar presente em uma região, menor será o intervalo de confiança (VISSCHER *et al.*, 1996).

O mapeamento de QTLs é um procedimento de testes múltiplos com o intuito de verificar qual ou quais marcadores estão ligados a QTLs. O problema neste procedimento refere-se ao nível de significância conjunto destes testes, chamado de nível de significância genômico (NONES *et al.*, 2006).

Segundo LANDER & KRUGLYAK (1995), é importante fazer a distinção entre o nível de significância pontual e o genômico. O nível de significância genômico é obtido utilizando a correção de Bonferroni, que consiste em determinar o valor do nível de significância no cromossomo que proporcionará o nível de significância conjunto desejado. Entretanto, a correção de Bonferroni é obtida considerando-se a utilização de t testes independentes, o que não ocorre no mapeamento de QTLs, pois os marcadores podem estar localizados em um mesmo cromossomo. Assim, o verdadeiro valor do nível de significância genômico obtido é um pouco menor do que o desejado, aumentando a ocorrência de falsos positivos (erro tipo I).

Alternativamente à utilização da correção de Bonferroni, pode-se lançar mão de procedimentos de permuta dos dados, um procedimento empírico que permite determinar o nível de significância genômico. Além desta abordagem, o nível de significância genômico pode ser expresso na forma do valor LOD, que indica na escala logarítmica a razão entre a probabilidade dos dados observados terem surgido assumindo a presença de um QTL sobre a probabilidade de sua ausência (CHURCHILL & DOERGE, 1994).

2.8. Mapas Genéticos e Mapeamento de QTL – Limitações

De acordo com LIU (1998), a precisão com que os QTLs podem ser detectados depende de uma série de questões, as quais devem ser respondidas adequadamente para delinear experimentos de mapeamento de

QTLs com eficiência. Sob essa abordagem, os seguintes fatores que afetam o poder de mapeamento de QTLs podem ser enumerados:

(a) número de genes controlando o caráter e sua posição no genoma;

(b) distribuição dos efeitos genéticos e a existência de interação gênica;

(c) herdabilidade do caráter;

(d) número e tipo de marcadores genéticos segregando na população de mapeamento;

(e) tipo e tamanho da população de mapeamento;

(f) densidade do mapa de ligação e cobertura;

(g) metodologia estatística empregada e nível de significância utilizado.

Os efeitos do número de marcadores e seu tipo, o número e tamanho da amostra têm sido relatados vastamente na literatura (DOERGE *et al.* 1997; LIU, 1998; LYNCH & WALSH, 1998; PATTERSON, 1998; DOERGE, 2002). Essencialmente, a detecção de QTL depende do tipo de marcador empregado, sua distribuição ao longo dos cromossomos (incluindo a fase de ligação: acoplamento ou repulsão), o delineamento experimental utilizado e a magnitude do QTL.

Em geral, em estudos de mapeamento de QTL que empregam delineamentos tradicionais como aqueles derivados de linhagens endogâmicas e que utilizam grandes tamanhos amostrais deverão apresentar maior facilidade na detecção de QTL. Como regra, pode-se dizer que experimentos que empregam um tamanho amostral menor do que 300 indivíduos poderão fornecer estimativas viesadas da verdadeira distribuição dos efeitos do QTL (ERICKSON *et al.* 2004). Segundo DOERGE *et al.* (1997), sob condições ideais, um QTL perfeitamente aditivo, não exibindo nenhuma dominância, pode ser detectado usando 206 indivíduos em uma população F_2 utilizando marcadores codominantes. Entretanto, devido às interações ambientais, à baixa herdabilidade e acurácia incompleta na estimativa do genótipo e do fenótipo, é sugerido uma amostra razoável de 300 indivíduos.

Considerando a necessidade de uma correta estimativa da magnitude de um QTL, um problema estatístico associado com o emprego de pequenas amostras é a superestimação do efeito do QTL. Quando amostras de tamanho

menor do que 300 indivíduos são utilizadas, estimativas do efeito do QTL serão exageradas e o poder de identificar QTLs de pequeno efeito decresce drasticamente, levando a uma redução do poder global de detecção de QTLs. Assim, um experimento com baixo poder de detecção não apenas falha em identificar o verdadeiro QTL, como também pode falsamente sugerir um QTL ou aumentar exageradamente a magnitude do efeito de um QTL verdadeiro (XU, 2003).

Os tipos de marcadores genéticos utilizados também têm efeitos importantes na resolução do mapeamento de QTL. Os marcadores genéticos podem ser classificados em dominantes e codominantes. Marcadores dominantes, tais como RAPD e AFLP, irão produzir apenas duas classes genotípicas em uma população F_2, não sendo possível distinguir entre as classes de homozigotos dominantes e heterozigotos. Dado o número menor de classes genotípicas detectadas, um menor número de eventos de recombinação será observado dentro de um intervalo, resultando em um menor conteúdo de informação quando marcadores dominantes são utilizados (LIU, 1998). Marcadores codominantes como microssatélites e SNP oferecem maior poder de inferência para as freqüências de recombinação entre marcadores adjacentes e, portanto, têm maior conteúdo de informação permitindo o alcance de maior resolução na detecção de QTL.

A distribuição dos marcadores genéticos ao longo dos grupos de ligação é de fundamental importância e irá afetar tanto o poder de detecção de um QTL, tanto quanto a sua localização. Quanto maior o número de marcadores disponíveis e, conseqüentemente, menores os intervalos entre marcadores adjacentes, mais precisas serão as estimativas dos efeitos do QTL e sua posição. Entretanto, um balanço entre o tamanho dos intervalos entre marcadores ao longo de um cromossomo e o tamanho das amostras precisa ser estabelecido. Estimativas a priori sugerem que para cerca de 300 indivíduos, intervalos de 10-15 cM entre marcadores são apropriados para a maioria dos experimentos.

Dessa forma, o que se tem sugerido é trabalhar com mapas iniciais aproximados e menos saturados. Assim, em uma abordagem inicial, deve-se

procurar maximizar o tamanho amostral às expensas de uma maior densidade de marcadores, levando a identificação de intervalos maiores (20 cM) contendo um eventual QTL. Em uma segunda etapa, mais marcadores podem ser incluídos nas regiões de interesse para refinar o efeito de posição do QTL. Tal procedimento pode reduzir o tempo e os custos relacionados ao mapeamento por evitar a genotipagem de regiões cromossômicas nas quais nenhum QTL está localizado.

3. MATERIAL E MÉTODOS

3.1. Simulação de Genomas

Todas as populações (famílias de meio-irmãos) foram geradas com base na espécie diplóide hipotética. Foi utilizado o módulo de simulação do aplicativo computacional GQMOL (CRUZ, 2004), que permite gerar informações sobre genomas, genótipos de genitores, indivíduos de diferentes tipos de populações e dados de características quantitativas.

3.1.1. Genoma de Referência

Uma espécie diplóide (2n= 2x= 6) hipotética, cujo comprimento total do genoma é estimado em 200 cM, foi tomada como base de simulação. Neste trabalho foram adotados os seguintes fatores de influência, denominados cenários de simulação, a serem mensurados no Mapeamento em Famílias de Meios-Irmãos:

(i) Nível de Saturação de Marcas nos Grupos de Ligação;

(ii) Nível de Informação Alélica na População de Referência;

(iii) Tamanho Amostral da População de Mapeamento (N).

A seguir serão descritos pormenorizadamente cada um dos itens citados acima.

3.1.1.1. Simulação do Nível de Saturação de Marcas nos Grupos de Ligação

Como a espécie diplóide apresenta 6 cromossomos, foram simulados três grupos de ligação com diferentes graus de saturação (Figura 4), conforme detalhado abaixo. As marcas foram assumidas como co-dominantes e multialélicas.

a) Grupo de Ligação 1 – Alta Saturação:

Constituído por 11 marcadores espaçados de forma eqüidistante a 5 cM, totalizando em comprimento de 50 cM.

b) Grupo de Ligação 2 – Média Saturação:

Constituído por 11 marcadores espaçados de forma eqüidistante a 10 cM, totalizando em comprimento de 100 cM.

c) Grupo de Ligação 3 – Baixa Saturação:

Constituído por 11 marcadores espaçados de forma eqüidistante a 15 cM, totalizando em comprimento de 150 cM.

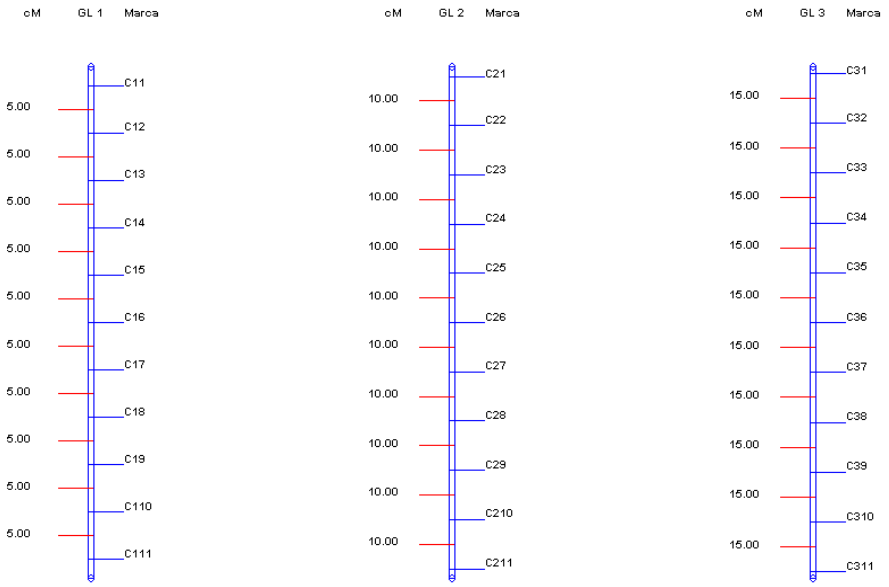


Figura 4. Genoma de referência com três grupos de ligação sob diferentes níveis de saturação. Cada grupo de ligação apresenta 11 marcas codominantes e eqüidistantes em 5, 10 e 15 cM.

3.1.1.2. Simulação do Nível de Informação Alélica na População de Referência

Na obtenção de famílias de meio-irmãos, em geral, apenas o genótipo do genitor comum é conhecido. A progênie gerada é dada como resultado do acasalamento ao acaso do genitor comum com a população segregante. O genótipo do genitor comum, que será denominado doravante de P_1 poderá ser homozigoto (A_iA_i) ou heterozigoto (A_iA_j), sendo este último de interesse no

mapeamento por permitir o estudo da ligação fatorial entre os marcadores. O genitor comum heterozigoto produz metade de gametas A_i e metade de A_j enquanto que a população segregante é capaz de contribuir com qualquer tipo de alelo A_k (k = 1, 2, ... , s; sendo s o número de alelos por loco) com probabilidade p_k. Neste trabalho, o nível de informação alélica será definido pelo número de alelos e sua freqüências, simulados para as populações segregantes, que darão origem às Famílias de Meios-Irmãos.

As populações segregantes, aqui denominadas de Populações de Referência, foram simuladas com s = 2, 3, 4, 5 e 6. O índice s será utilizado para denotar o nível de informação alélica. Para cada valor de s, os valores de freqüência p_k foram simulados equitativamente. Assim, para o valor de s = 2 os valores das freqüências são $p_1 = p_2 = 0{,}50$. Para s = 3 os valores das freqüências são $p_1 = p_2 = p_3 = 0{,}33$, e assim sucessivamente.

Os valores de s para cada população foram relacionados com valor de PIC (BOTSTEIN et al. 1980). O valor de PIC é dado por:

$$PIC = 1 - \sum_{i=1}^{n} p_i^2 - \sum_{i=1}^{n-1} \sum_{j=1+1}^{n} 2 p_i^2 p_j^2 \leq \frac{(n-1)^2 (n+1)}{n^3}$$

onde: p_i é a freqüência do i-ésimo alelo de um total de n alelos amostrados no loco marcador. O termo à direita da desigualdade ocorre para a aplicação do presente trabalho, quando todos os alelos têm a mesma freqüência alélica ($p_i = 1/n$ onde n = s), sendo este o valor máximo para um dado valor de i. Na Tabela 2 estão demonstrados os valores dos níveis de informação alélica (s) utilizados, os valores das freqüências alélicas p_s e os valores de PIC associados.

Tabela 2. Número de alelos (s), freqüência alélica (p_s) e valor de PIC associado para as diferentes populações segregantes de meios-irmãos simuladas

Número de Alelos na População de Mapeamento	Freqüência Alélica (p_s)	PIC
s = 2	0,50	0,3750
s = 3	0,33	0,5926
s = 4	0,25	0,7031
s = 5	0,20	0,7679
s = 6	0,17	0,8103

3.1.1.3. Simulação do Tamanho Amostral da População de Mapeamento (N)

Para os diferentes níveis de informação alélica (s = 2, 3, 4, 5 e 6) e os diferentes níveis de saturação foram geradas FMI com tamanho amostral variando de N = 50, 100, 200, 300, 500 e 1000 indivíduos. Para cada combinação de N x s foram simuladas 100 repetições, totalizando 3000 populações de mapeamento a serem analisadas (6 (N) x 5 (s) x 100 repetições).

3.2. Procedimento de Simulação do Genitor P_1 e dos Indivíduos das Progênies de Meios-Irmãos

Os genitores P_1 e os indivíduos da população foram gerados assumindo-se o equilíbrio de Hardy-Weinberg para locos individuais. Para o genitor P_1 foram assumidos para a análise somente genótipos heterozigotos do tipo A_1A_2, uma vez que esta é a condição necessária e suficiente para que seja possível a análise de ligação com dados de FMI.

Para cada indivíduo das progênies de FMI foram produzidos os dados genotípicos referentes aos marcadores de acordo com a informação do genoma de referência segundo os diferentes cenários já descritos (nível de

saturação, informatividade alélica (s) e tamanho populacional (N)). A estratégia básica de simulação foi baseada em permutas nos intervalos entre marcas adjacentes em cada cromossomo, de acordo com as distâncias dos marcadores. O processo de simulação pode ser sumarizado através dos seguintes passos:

i) A partir do genoma simulado foram construídos os cromossomos do genitor P_1 conforme especificado acima.

ii) Os gametas do genitor P_1 foram gerados com base no genótipo de P_1 obtido do Genoma de Referência para a formação dos indivíduos nas progênies FMI. A simulação dos gametas foi obtida através do pareamento dos cromossomos homólogos do genoma de referência, seguido por permutas ao longo dos cromossomos nas regiões delimitadas por dois marcadores adjacentes. A probabilidade de ocorrência de recombinação numa região entre marcadores adjacentes será dada de acordo com a distância destes marcadores no genoma simulado. Por exemplo, se a distância entre os dois primeiros marcadores em um cromossomo for de 10 cM no genoma simulado a probabilidade de recombinação nesta região será de 10%. Após este primeiro passo, o processo de simulação seguirá para a próxima região, delimitada pelo segundo e terceiro marcadores, prosseguindo sucessivamente até que todas as regiões entre marcadores adjacentes no cromossomo sejam contempladas. Para a formação de cada indivíduo da população foram simulados 5000 gametas parentais, na forma como apresentado acima, sendo sorteado apenas um único gameta para compor o genótipo do indivíduo.

iii) Os gametas oriundos da População Referência foram construídos por amostras aleatórias de alelos para cada loco, segundo os diferentes níveis de informatividade alélica.

iv) O genótipo dos indivíduos nas progênies de FMI foi obtido por meio da junção dos gametas do genitor P_1 com os gametas gerados a partir da População de Referência.

3.3. Análise Genômica – Mapeamento

Obtidos os dados simulados, o mapeamento genético em Famílias de Meios-Irmãos foi efetuado como descrito a seguir.

3.3.1. Análise de Segregação de Locos Individuais

Foi aplicado o teste Qui-quadrado $\left(\chi^2\right)$ para verificar a razão de segregação de cada marca em todas as populações geradas.

A estatística Qui-quadrado é dada por:

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

onde: χ^2 é valor de qui-quadrado calculado; Obs_i e Esp_i, são os valores observado e esperado, para a i-ésima classe fenotípica.

A hipótese (H_0) de segregação específica para cada loco foi testada a 5% de probabilidade (erro tipo I).

3.3.2. Análise de Pares de Marcas – Estimação da Fase de Ligação e Percentagem de Recombinação

Efetuada a análise de locos individuais, as freqüências de recombinação entre pares de marcas foram determinadas. O Método da Máxima Verossimilhança foi utilizado para este fim. Os detalhes da aplicação do Método da Máxima Verossimilhança em Famílias de Meios-Irmãos serão descritos no item Resultados e Discussão.

3.3.3. Determinação dos Grupos de Ligação e Ordenamento das Marcas

Para a formação dos grupos de ligação foram utilizados os critérios de agrupamento baseado na freqüência máxima de recombinação (r_{max} = 30%) e LOD escore mínimo (LOD_{min} = 3). Após a formação dos grupos de ligação, a melhor ordem das n marcas nos grupos foi determinada pelo SARF (*Sum of*

Adjacent Recombination Fractions), em que a melhor ordem é aquela que apresenta a menor soma das recombinações adjacentes. Neste método, considera-se a ordem original estabelecida pelo processo de agrupamento e aplica-se o algoritmo RCD (*Rapid Chain Delineation*), que consiste em realizar permutas entre dois marcadores vizinhos ou distantes envolvendo três ou quatro marcadores. A ordem é alterada se, após a permuta, a soma das distâncias adjacentes for reduzida. Após todas as permutas conclui-se que a melhor ordem é aquela que apresentar menor soma de distâncias adjacentes (SCHUSTER & CRUZ, 2004).

3.4. Comparação dos Genomas

Foi utilizado o módulo 'Comparação de Genomas', do aplicativo computacional GQMOL (CRUZ, 2004), para se realizar as comparações entre o Genoma de Referência e o mapa genético obtido a partiras das progênies de Meios-Irmãos simuladas nos diferentes cenários já descritos. Em cada situação foram estudados nas progênies de FMI simuladas, o número de grupos de ligação recuperados, os tamanhos dos grupos de ligação, as distâncias médias entre marcadores adjacentes nos grupos de ligação, as variâncias das distâncias entre marcas adjacentes nos grupos de ligação, o estresse, e a correlação de Spearman. Estes critérios, descritos por FERREIRA et al. (2006) foram levados em conta para se determinar a acurácia do mapeamento genético em FMI. A seguir, cada critério será descrito individualmente.

3.4.1. Número de Grupos de Ligação Recuperados

Em cada repetição de simulação, foi considerado como grupo de ligação recuperado, particular para cada nível de saturação simulado, aquele que exibir as mesmas marcas e o mesmo número de marcas expostas no Genoma de Referência. O número máximo de grupos corretamente recuperados será igual a 100, que correspondem a 100 repetições para cada cenário simulado. Um

grupo de ligação com menor número de marcas será considerado não recuperado.

3.4.2. Tamanho do Grupo de Ligação

Dado pelo somatório das distâncias entre marcas adjacentes no grupo de ligação analisado, como segue:

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

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

3.4.3. Distância Média de Dois 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:

$$\bar{d} = \frac{L}{m-1}$$

3.4.4. Variâncias das Distâncias entre Marcas Adjacentes

Esta medida faz-se útil uma vez que o genoma simulado possui marcas eqüidistantes, apresentando variância nula. Portanto, será avaliada a variância do genoma analisado, partindo-se do pressuposto que o ideal é ter variância nula. A estimativa é dada por:

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

em que: $\hat{\sigma}^2$ é a variância das distâncias entre marcas adjacentes; d_k é a distância entre as marcas adjacentes m_k e m_{k+1} no grupo de ligação do

genoma analisado ($k = 1, \ldots, m-1$); $\bar{d}$ é a distância média entre marcadores adjacentes no grupo de ligação do genoma analisado; I é o número de intervalos entre marcas adjacentes, dado por $m-1$, onde m é o número de marcadores no grupo de ligação.

3.4.5. Estresse

O coeficiente de estresse (S) é utilizado como medida de adequação das distâncias estimadas em relação às verdadeiras distâncias determinadas no Genoma Referência. Conforme FERREIRA et al. (2006), o coeficiente de estresse é dado por:

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

em que: d_{ok} é a distância entre marcas adjacentes m_k e m_{k+1} no Genoma Referência; e, d_k é a distância entre marcas adjacentes m_k e m_{k+1} no genoma analisado ($k = 1, \ldots, m-1$).

3.4.6. Correlação de Spearman

A Correlação de Spearman será empregada para avaliar o grau de concordância do ordenamento das marcas nos genomas analisados e nos genomas simulados. É também conhecida como correlação de *rank* e sua aplicação é importante quando não é possível mensurar variação contínua de duas variáveis nos membros de uma população. Entretanto, é possível mensurar as variáveis por meio de notas (*rank*), onde cada nota pode ser colocada em ordem para os membros de uma população. Esta correlação expressa o grau de concordância nas notas das duas variáveis, desta forma a sua utilização na análise de genomas foi proposta FERREIRA et al. (2006), conforme descrito a seguir:

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

em que: r_s é o valor estimado da correlação de Spearman para um grupo de ligação do genoma analisado $\left(-1 \leq r_S \leq 1\right)$; Δ_k é a diferença da nota do marcador m_k ($k = 1,\ldots,m$) na $k-ésima$ posição do grupo de ligação do genoma simulado e a nota do marcador m_k na $k-ésima$ posição do grupo de ligação do genoma analisado; m é o número de marcadores no grupo de ligação do genoma simulado.

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.

3.4.7. Testes de comparação múltipla de médias

As médias das variáveis (i) tamanho de grupo de ligação, (ii) distância média de marcas adjacentes, (iii) variância e (iv) estresse obtidas para os diferentes tamanhos de população, dentro de cada nível de saturação e de informatividade alélica, foram comparadas pelo teste de comparação múltipla de médias de Tukey, a 5% de probabilidade (erro tipo I), com o auxílio do aplicativo computacional GENES (CRUZ, 2001).

4. RESULTADOS E DISCUSSÃO

4.1. Aplicação do Mapeamento Genético em Famílias de Meios-Irmãos

Neste tópico será abordada a análise de locos individuais e o cálculo da freqüência de recombinação para pares de marcas na construção de mapas genéticos com base em progênies de Meios-Irmãos. Para tanto, como ilustração, será utilizada uma única repetição das populações de tamanho N = 50, 200 e 1000, para todos os níveis de informatividade s = 2, 3, 4, 5 e 6. Para a análise de locos individuais, a ênfase será dada em aspectos da segregação de marcas em cruzamentos exogâmicos e os efeitos da informatividade em locos individuais comparando-se FIC e FMI. Na análise de pares de marcas para o cálculo das freqüências de recombinação, será demonstrada a função de verossimilhança utilizada no cálculo das distâncias em FMI e a influência da informatividade nos valores de LOD escore calculados com respeito a todos os cenários de simulação.

4.1.1. Informatividade e Análise de Loco Único

A análise de locos individuais, também denominada na literatura de *single-locus model*, é utilizada para o controle de qualidade dos dados no mapeamento genético e para identificar o modelo de segregação mendeliana para um único loco. Um loco simples pode ser caracterizado pelo número de alelos presentes, suas freqüências na população, seu modo de herança e seu estado de equilíbrio segundo o modelo de Hardy-Weinberg (LIU, 1998). Em populações endogâmicas, as freqüências alélicas e o estado de equilíbrio são únicos e definidos pelo delineamento experimental utilizado. Para dados de cruzamentos exogâmicos esta etapa da análise assume uma importante função, principalmente tendo em vista que o número e a freqüência dos alelos pode ser variável para cada loco em análise.

Como já mencionado, em FIC a segregação de marcas únicas na progênie pode ser variável, assumindo as configurações de (1:1:1:1), (1:2:1) ou (1:1), definidos pela informatividade do loco marcador. Em cruzamentos completamente informativos, quando todos os alelos são distintos em ambos os genitores, a razão de segregação esperada é de (1:1:1:1). Em cruzamentos informativos, que assume uma configuração de retrocruzamentos, a razão de segregação esperada é de (1:1). Em cruzamentos parcialmente informativos, que assume a configuração de intercruzamentos, a razão de segregação esperada é de (1:2:1). Em FMI apenas o genitor comum heterozigoto do tipo A_iA_j é informativo e na progênie apenas os genótipos Ai- e Aj- são informativos, estabelecendo uma razão de segregação de 1:1. Genótipos do tipo A_iA_j na progênie são não informativos.

Na Tabela 3 é apresentado o teste de χ^2 para segregação das marcas C11 e C12, tomadas como ilustração, localizadas no grupo de ligação 1 de alta saturação, para três diferentes tamanhos populacionais (N = 50, 200 e 1000) e segundo os diferentes níveis de informatividade (s = 1, 2, 3, 4, 5 e 6). Pelo exposto, nenhuma das marcas apresentou segregação distorcida. Este fato é esperado uma vez que se tratam de dados simulados.

Tabela 3. Teste do χ^2 para segregação das marcas C11 e C12 para N = 50, 200 e 1000 segundo os diferentes níveis de informatividade alélica (s = 2, 3, 4, 5 e 6)

N	Marcador	Ai-	Aj-	A_iA_j	Hipótese	χ^2	Prob(%)
				s = 2			
	C11	11	12	27	1:1	0,043	83,4827 ^{ns}
	C12	11	19	20	1:1	2,133	14,4127 ^{ns}
				s = 3			
	C11	18	17	15	1:1	0,029	86,5772 ^{ns}
	C12	20	20	10	1:1	0	100,0 ^{ns}
				s = 4			
50	C11	21	18	11	1:1	0,231	63,0954 ^{ns}
	C12	20	19	11	1:1	0,026	87,278 ^{ns}
				s = 5			
	C11	18	20	12	1:1	0,105	74,5603 ^{ns}
	C12	17	23	10	1:1	0,9	34,2782 ^{ns}
				s = 6			
	C11	20	21	9	1:1	0,024	87,5896 ^{ns}
	C12	20	25	5	1:1	0,556	45,6056 ^{ns}

Tabela 3. Continuação.: Teste do χ^2 para segregação das marcas C11 e C12 para N = 50, 200 e 1000 segundo os diferentes níveis de informatividade alélica (s = 2, 3, 4, 5 e 6)

N	Marcador	Ai-	Aj-	A_iA_j	Hipótese	χ^2	Prob(%)
				s = 2			
	C11	49	45	106	1:1	0,17	67,9923 ^{ns}
	C12	53	48	99	1:1	0,248	61,8823 ^{ns}
				s = 3			
	C11	83	62	55	1:1	3,041	8,1167 ^{ns}
	C12	64	63	73	1:1	0,008	92,9292 ^{ns}
				s = 4			
200	C11	79	75	46	1:1	0,104	74,7203 ^{ns}
	C12	70	76	54	1:1	0,247	61,9497 ^{ns}
				s = 5			
	C11	71	86	43	1:1	1,433	23,1256 ^{ns}
	C12	76	81	43	1:1	0,159	68,9861 ^{ns}
				s = 6			
	C11	86	82	32	1:1	0,095	75,7621 ^{ns}
	C12	83	86	31	1:1	0,053	81,7494 ^{ns}
				s = 2			
	C11	248	241	511	1:1	0,1	75,1584 ^{ns}
	C12	234	251	515	1:1	0,596	44,0156 ^{ns}
				s = 3			
	C11	332	345	323	1:1	0,25	61,7335 ^{ns}
	C12	325	358	317	1:1	1,594	20,6693 ^{ns}
				s = 4			
1000	C11	381	332	287	1:1	3,367	6,6496 ^{ns}
	C12	386	370	244	1:1	0,339	56,0624 ^{ns}
				s = 5			
	C11	386	432	182	1:1	2,587	10,7758 ^{ns}
	C12	397	399	204	1:1	0,005	94,3487 ^{ns}
				s = 6			
	C11	398	430	172	1:1	1,237	26,6105 ^{ns}
	C12	403	426	171	1:1	0,638	42,4393 ^{ns}

Deve ser observado que na medida em que o nível de informatividade aumenta, a freqüência de genótipos não informativos do tipo A_iA_j apresenta redução substancial. Por exemplo, tomando-se a marca C11 para N = 50 e s = 2, verifica-se que o número de indivíduos na progênie com genótipo não informativo do tipo A_iA_j é superior ao somatório dos indivíduos informativos A_i- e A_j-. O mesmo ocorre para a marca C11 em s = 2 nos tamanhos N = 200 e 1000. De fato, considerando a segregação média para as para as 33 marcas

do genoma para as repetições usadas como ilustração, em populações de N = 50, 200 e 1000, verifica-se que as freqüências médias de genótipos informativos (Ai- + Aj-) para s = 2, 3, 4, 5 e 6 convergem para os seguintes valores (Tabela 4):

f_{s2} (Ai- + Aj-) = 0,50

f_{s3} (Ai- + Aj-) = 0,66

f_{s4} (Ai- + Aj-) = 0,75

f_{s5} (Ai- + Aj-) = 0,79

f_{s6} (Ai- + Aj-) = 0,84

As freqüências de genótipos não informativos f_s (A_iA_j) são dadas por 1 - f_s (Ai- + Aj-).

Tabela 4. Segregação média de locos únicos e freqüências de genótipos informativos f (Ai- + Aj-) e não informativos f (AiAj) para uma repetição simulada de tamanho populacional (N) e informatividade (s)

N - s	Ai-	Aj-	AiAj	Total	f (AiAj)	f (Ai- + Aj-)
50-2	11,7878	13,3030	24,9091	50	0,4982	0,5018
50-3	16,6969	17,7879	15,5152	50	0,3103	0,6897
50-4	19,1515	19,0303	11,8182	50	0,2364	0,7636
50-5	18,6667	20,7879	10,5455	50	0,2109	0,7891
50-6	20,6364	21,6364	7,7273	50	0,1546	0,8455
200-2	50,2727	49,7576	99,9697	200	0,4998	0,5002
200-3	68,4242	64,8485	66,7273	200	0,3336	0,6664
200-4	78,5152	72,4546	49,0303	200	0,2453	0,7548
200-5	82,6061	77,3030	40,0909	200	0,2005	0,7995
200-6	83,8788	83,6364	32,4849	200	0,1624	0,8376
1000-2	252,6061	248,9697	498,4242	1000	0,4984	0,5016
1000-3	337,6970	326,6970	335,6061	1000	0,3356	0,6644
1000-4	370,4242	379,8485	249,7273	1000	0,2497	0,7503
1000-5	398,4848	399,7576	201,7576	1000	0,2018	0,7982
1000-6	411	422,8788	166,1212	1000	0,1661	0,8339

A explicação para este fato pode ser dada da seguinte maneira: na medida em que o valor de s aumenta na População de Referência – com s variando de 2 a 6 – a freqüência de acasalamentos dos tipos A_iA_j x A_iA_i, A_iA_j x A_jA_j, A_iA_j x A_iA_j é reduzida. Veja que estas classes de cruzamentos são logicamente associadas a níveis de baixa informatividade assim como em FIC.

Por outro lado a freqüência de cruzamentos do tipo A_iA_j x A_kA_l, A_iA_j x A_iA_k, A_iA_j x A_jA_k e A_iA_j x A_kA_k aumenta com o aumento de s, o que reduz a freqüência de genótipos não informativos A_iA_j na progênie. Da mesma forma, estes cruzamentos são associados à idéia de informatividade para locos individuais em FIC, tal qual proposto por LYNCH & WALSH (1998).

Dessa forma, um modelo geral de informatividade é dado a seguir (Tabela 5), adaptado de DA & LEWIN (1995), para a freqüência esperada de genótipos informativos em um loco individual, considerando progênies de FIC e FMI em cruzamentos exogâmicos.

Tabela 5. Freqüência de genótipos informativos em cruzamentos exogâmicos para um loco único para acasalamentos com um genitor P_1 heterozigoto A_iA_j. Adaptado de DA & LEWIN (1995)

Genitor (P_2)	**Freqüência Genotípica na Progênie**	**Genótipos Informativos**		**Freqüências de Genótipos Informativos**		**Classe de Cruzamentos segundo Lynch & Walsh (1998)**
Genótipo		**FIC**	**FMI**	**FIC**	**FMI**	
A_kA_l	$1/4A_iA_k+1/4A_iA_l$ $+1/4A_jA_k+1/4A_jA_l$	Todos	Todos	1	1	Completamente Informativa
A_iA_k	$1/4A_iA_i+1/4A_iA_j$ $+1/4A_iA_k+1/4A_jA_k$	Todos	A_iA_i A_iA_k A_jA_k	1	3/4	Completamente Informativa
A_jA_k	$1/4A_jA_j+1/4A_iA_j$ $+1/4A_iA_k+1/4A_jA_k$	Todos	A_iA_i A_iA_k A_jA_k	1	3/4	Completamente Informativa
A_kA_k	$1/2A_iA_k+$ $1/2A_jA_k$	Todos	Todos	1	1	Informativa
A_iA_j	$1/4A_iA_i+1/2A_iA_j$ $+1/4A_jA_j$	A_iA_i A_jA_j	A_iA_i A_jA_j	1	1/2	Parcialmente Informativa
A_iA_i	$1/2A_iA_i+1/2A_iA_j$	Todos	A_iA_i	1	1/2	Informativa
A_jA_j	$1/2A_jA_j+1/2A_iA_j$	Todos	A_jA_j	1	1/2	Informativa

Considere uma população em que o loco A_i apresenta i = 1, 2, ..., até n alelos. O i-ésimo alelo A_i terá freqüência de p_i na população. Admitindo-se o equilíbrio de Hardy-Weinberg, as freqüências de cruzamentos do genitor A_iA_j heterozigoto-informativo com os genótipos tomados ao acaso na população são definidas pelas freqüências dos diferentes alelos encontrados na população.

Analisando-se a Tabela 5, pode-se facilmente deduzir as freqüências de genótipos informativos na progênie de FIC e FMI segundo as diferentes configurações de cruzamentos. Para se determinar os genótipos informativos nas progênies de FIC as mesmas regras descritas pela Tabela 1, segundo LYNCH & WALSH (1998), são utilizadas. Em progênies de FIC as freqüências de genótipos informativos é determinada pela exclusão do genótipo não informativo A_iA_j. Dessa forma, por exemplo, a freqüência de genótipos informativos na progênie do acasalamento A_iA_j x A_iA_k será igual a 1, para FIC, uma vez que a origem dos alelos de cada genitor pode ser definida precisamente. Em FMI, somente 3/4 da progênie será informativa, uma vez que só é conhecido o genótipo do genitor comum A_iA_j. Os alelos nos genótipos A_iA_j na progênie são indistinguíveis com relação a sua origem e, portanto, são não informativos.

Outro fato que pode ser verificado é que em progênies de Meios-Irmãos ocorre uma maior redução de genótipos informativos para determinados acasalamentos que em FIC. A causa desta redução é devida, logicamente, a impossibilidade de se genotipar ambos os parentais em FMI, além do teste para locos individuais ser de natureza gamética e não genotípica. Em outras palavras, em progênies de Irmãos Completos é possível realizar os testes de segregação de natureza gamética e genotípica. Nesse caso, para o teste de segregação genotípica toda a progênie é informativa.

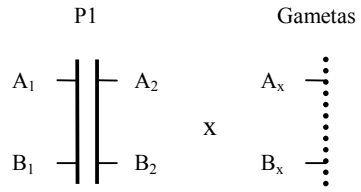
Pode-se depreender deste modelo geral, que o nível de informatividade, ou mais especificamente, o nível de polimorfismo amostrado na população terá importante papel na definição das freqüências de genótipos informativos nas progênies de cruzamentos exogâmicos, tanto para FIC, quanto para FMI. Como demonstrado para os dados simulados sumarizados na Tabela 4, as

freqüências de genótipos informativos observados convergem para as freqüências esperadas com base na Tabela 5. Da mesma forma pode-se concluir que à medida que o nível de informatividade ou o PIC da população aumenta as freqüências de genótipos informativos tendem a aumentar.

4.1.2. Informatividade e Análise de Pares de Locos

A análise de pares de locos, ou *two-locus model*, inclui a detecção de ligação e a estimação da freqüência de recombinação. Os resultados da análise de pares de locos são a base para os passos subseqüentes na construção de mapas de ligação genômicos, como o agrupamento das marcas em grupos de ligação e o seu ordenamento (LIU, 1998). Ademais, a teoria para a análise de ligação entre marcas é também fundamental para a análise e detecção de QTLs.

Mapas genéticos envolvendo informações de indivíduos meios-irmãos, provenientes de um genitor comum cuja informação genotípica é conhecida, são facilmente estabelecidos considerando apenas a segregação gamética deste genitor. Admitindo que os genótipos do genitor comum (P_1), para todos os marcadores em estudo, são heterozigotos, pode-se avaliar o número de ocorrência de indivíduos na progênie considerando o seguinte esquema de cruzamento:



As freqüências de gametas parentais e recombinantes são definidas na progênie somente com base na segregação dos alelos oriundos do genitor P_1 (Tabela 6). As progênies A_1A_2 e/ou B_1B_2, como já demonstrado para locos únicos são não informativas.

Tabela 6. Freqüência de gametas recombinantes (R) e parentais (P) em progênies de Famílias de Meios-Irmãos

Gameta do Genitor P_1	Progênie*	Número de indivíduos	Freqüência
A_1B_1	A_1-B_1-	n_{11}	(1-r)/2 (P)
A_1B_2	A_1-B_2-	n_{12}	r/2 (R)
A_2B_1	A_2-B_1-	n_{21}	r/2 (R)
A_2B_2	A_2-B_2-	n_{22}	(1-r)/2 (P)

(*) Exclui-se as informações da progênie A_1A_2 e/ou B_1B_2 que não são informativas quanto a origem dos alelos parentais.

Um fator complicador na análise de ligação em cruzamentos envolvendo 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 pode ser detectada, antes de se realizar o mapeamento, a partir das informações dos genitores. Entretanto, no caso da análise de Família de Meios-Irmãos ela poderá ser facilmente estabelecida a partir da análise da progênie, computando o número de ocorrência de cada classe gamética do genitor comum. Assim, entre as duas fases possíveis, admite-se que:

A_1B_1// A_2B_2 se $n_{11} + n_{22} > n_{12} + n_{21}$

A_1B_2// A_2B_2 se $n_{12} + n_{21} > n_{11} + n_{22}$

Teve-se ter em mente que ao serem avaliados N indivíduos na progênie, tem-se que:

$$N = n_{11} + n_{22} + n_{12} + n_{21} + n_{00}$$

em que n_{00} é o número de indivíduos da progênie com genótipo idêntico ao do genitor comum. Portanto, o número efetivo utilizado para a estimativa da freqüência de recombinação entre pares de locos será inferior ou número de indivíduos genotipados na população.

No presente trabalho foi aplicado o método da Máxima Verossimilhança (MV) para o cálculo das freqüências de recombinação (SCHUSTER & CRUZ, 2004; LIU, 1998). A função de verossimilhança utilizada para o cálculo MV baseia-se na função densidade de probabilidade (fdp) da distribuição de uma variável aleatória x. Para o cálculo das freqüências de recombinação, esta função é construída com base na distribuição multinomial (SCHUSTER & CRUZ, 2004; LIU, 1998). Função de Verossimilhança para a distribuição multinomial assume a seguinte configuração:

$$L(p_i;n_i) = \lambda {p_1}^{n_1} {p_2}^{n_2} \ldots {p_n}^{n_n}$$

onde: n_i é o número de observações na i-ésima classe; p_i: a probabilidade de ocorrência da i-ésima classe; e, $\lambda = \frac{N!}{n_1!n_2!\ldots n_n!}$, em que $N = \sum_i n_i$.

Portanto, considerando as informações dispostas na Tabela 5 em que é considerado o genitor P_1 em fase de aproximação, temos que:

$$L(r;n_i) = \lambda \left[\frac{1}{2}(1-r)\right]^{n_{11}} \left[\frac{1}{2}(1-r)\right]^{n_{22}} \left[\frac{1}{2}r\right]^{n_{12}} \left[\frac{1}{2}r\right]^{n_{21}}$$

A função suporte é dada por:

$$\ell(r;n_i) = \ln\lambda + N\ln(\frac{1}{2}) + (n_{11}+n_{22})\ln(1-r) + (n_{12}+n_{21})\ln(r)$$

Pela função suporte é obtida a função escore dada pela derivada em função de r:

$$\frac{\partial \ell(r;n_i)}{\partial r} = \frac{(n_{11}+n_{22})}{1-r}(-1) + \frac{(n_{12}+n_{21})}{r} = \frac{(n_{12}+n_{21})-Nr}{r(1-r)}$$

A função escore é maximizada pela igualdade:

$$\frac{(n_{12}+n_{21})-Nr}{r(1-r)} = 0$$

Obtendo-se:

$$r = \frac{(n_{12}+n_{21})}{N}$$

Este estimador também expressa o conceito de distância entre locos gênicos dado pela freqüência de gametas recombinantes.

A variância da estimativa da porcentagem de recombinação é obtida pela derivada segunda da função suporte ou a derivada primeira da função escore, de maneira que:

$$f'(r) = \frac{(n_{12}+n_{21})-Nr}{r(1-r)}$$

Denotando o denominador e numerador, respectivamente, por duas novas funções, tem-se:

$$H(r) = r(1-r)$$

$$G(r) = (n_3+n_4)-Nr$$

Logo:

$$G'(r) = -N$$

De forma que o índice de informatividade será:

$$I(r) = -f''(r) = \frac{N}{r(1-r)}$$

O conteúdo médio de informação, que é característico da população de mapeamento, será expresso, para progênies de Meios-Irmãos, por:

$$c(r) = \frac{1}{r(1-r)}$$

Logo, a variância será dada por:

$$V(r) = \frac{1}{Nc(r)}$$

Na Tabela 7 pode ser observada a segregação conjunta das marcas C11 e C12 para N = 50, 200 e 1000 para os diferentes níveis de informatividade estudados. Os valores das freqüências de recombinação e o LOD escore são apresentados. O objetivo é evidenciar a redução do tamanho da população de mapeamento em FMI. A redução do tamanho N é ocasionada pela perda de informação dada pela presença de genótipos não informativos A_1A_2 ou B_1B_2 na progênie. O tamanho real, ou aqui denominado tamanho efetivo (n_{ef}), refere-se à soma de genótipos informativos na progênie que são efetivamente utilizados no cálculo de r. Caracteristicamente, o tamanho efetivo tende a aumentar na medida em que o nível de informatividade aumenta na população, uma vez que a freqüência de genótipos não informativos reduz com o aumento de s.

Tabela 7. Segregação conjunta das marcas C11 e C12 para N = 50, 200, 1000 segundo os diferentes níveis de informatividade alélica (s = 2, 3, 4, 5 e 6) e tamanho efetivo da população de mapeamento em progênies de FMI. Valor esperado de (r = 0,05)

N - s	M1	M2	A1-B1-	A1-B2-	A2-B1-	A2-B2-	n_{ef}	r (%)	LOD escore
50-2	C11	C12	9	0	0	5	14	0	0
50-3	C11	C12	11	2	1	14	28	10,71	4,29
50-4	C11	C12	14	1	1	16	32	6,25	6,38
50-5	C11	C12	17	0	2	14	33	6,06	6,57
50-6	C11	C12	19	1	2	15	37	8,11	6,16
LOD Máximo*			23,75	1,25	1,25	23,75	50	5	10,74
200-2	C11	C12	22	0	1	30	53	1,89	13,80
200-3	C11	C12	39	1	3	49	92	4,35	20,55
200-4	C11	C12	51	5	4	51	111	8,11	19,85
200-5	C11	C12	65	1	5	52	123	4,88	26,62
200-6	C11	C12	67	2	3	70	142	3,52	33,35
LOD Máximo*			95	5	5	95	200	5	42,96
1000-2	C11	C12	119	5	3	115	242	3,31	57,59
1000-3	C11	C12	225	6	11	219	461	3,69	107,16
1000-4	C11	C12	257	5	14	271	547	3,47	128,83
1000-5	C11	C12	304	30	16	300	650	7,08	123,51
1000-6	C11	C12	336	20	18	310	684	5,56	142,17
LOD Máximo*			475	25	25	475	1000	5	214,82

r = freqüência de recombinação; $LOD = \log_{10} \frac{L(r; n_i)}{L(r = 0{,}5; n_i)}$; n_{ef} = Tamanho efetivo; *LOD Máximo calculado com base nas freqüências esperadas para r = 5% quando n_{ef} = N.

A redução da informatividade pode também ser evidenciada por meio do LOD Máximo, calculado quando N = n_{ef} e assumindo-se que as freqüências esperadas para r = 5 são iguais às freqüências observadas. Na medida em que o nível de informatividade aumenta, o tamanho efetivo tende a se aproximar de N e o LOD escore calculado aproxima-se do LOD Máximo. Deve-se levar em conta, que o LOD escore é calculado em função do tamanho do tamanho da amostra (N) e da freqüência de recombinação (r) (LIU, 1998; SCHUSTER & CRUZ, 2004). Portanto, somente fazem sentido as comparações de LOD escores quando fixados estes valores (SILVA et al., 2007).

De acordo com a Figura 5 verifica-se que os valores médios de LOD escore em cada tamanho de população aumentam na medida em que o nível

de informatividade aumenta. Exceto para a saturação de 10 cM, quando LOD escore para o nível de informatividade s = 3 é menor que para s = 2, o aumento do LOD escore reflete perfeitamente o aumento da freqüência de genótipos informativos na progênie e o aumento do tamanho efetivo de mapeamento.

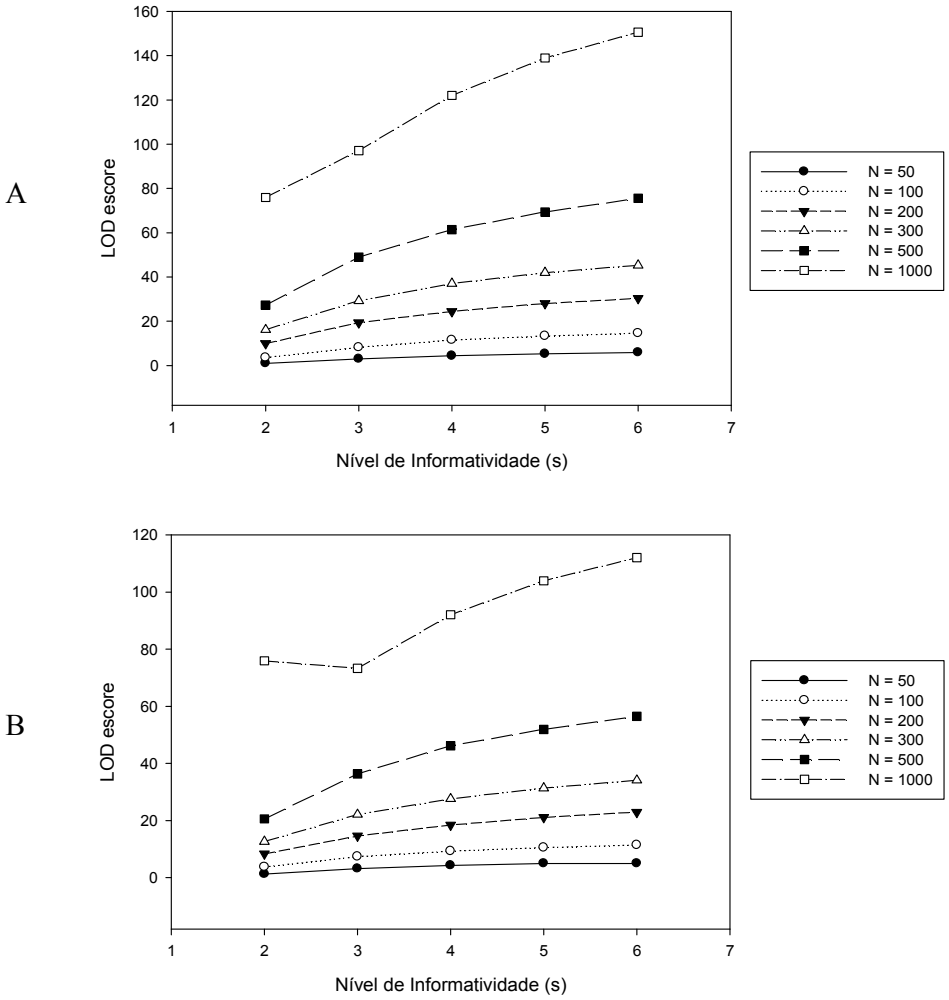


Figura 5. LOD escore médio em função do nível de informação alélica (s) para os diferentes tamanhos de FMI (N). (A) Alta Saturação – 5 cM; (B) Média Saturação – 10 cM; (C) Baixa Saturação – 15 cM

C

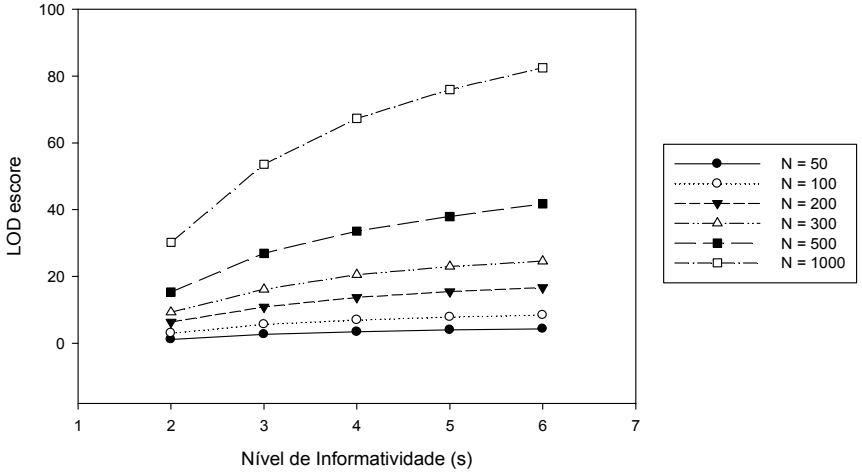


Figura 5. Continuação: LOD escore médio em função do nível de informação alélica (s) para os diferentes tamanhos de FMI (N). (A) Alta Saturação – 5 cM; (B) Média Saturação – 10 cM; (C) Baixa Saturação – 15 cM

4.2. Efeitos do Tamanho da População, Níveis de Informatividade e Saturação de Marcas no Mapeamento Genético em FMI

A metodologia de análise de ligação e mapeamento genético em FMI tem sido pouco explorada na literatura. Determinar a acurácia de mapas genéticos é de fundamental importância tanto para o mapeamento de QTLs, quanto para a obtenção de mapas fidedignos representativos de genomas das mais variadas espécies (SILVA et al. 2007; FERREIRA et al. 2006; ROCHA et al., 2004). A acurácia de mapas genéticos irá depender de vários fatores tais como: o delineamento experimental utilizado e sua conseqüente estrutura populacional; o tamanho da população empregada; o tipo e o número de marcadores utilizados. Além desses fatores, em cruzamentos exogâmicos deve ser levada em conta a influência do grau de informatividade para cada particular loco na análise de ligação e na acurácia do mapa obtido. Por exemplo, em FIC é sabido que genitores completamente informativos do tipo A_iA_j x A_kA_l fornecem mapas mais acurados (BHERING et al. 2008a).

Diferentes critérios podem ser utilizados para se determinar a acurácia de mapas genéticos. Neste trabalho, foram utilizados aqueles descritos por FERREIRA et al (2006), quais sejam: número de grupos de ligação recuperados, os tamanhos dos grupos de ligação, as distâncias médias entre marcadores adjacentes nos grupos de ligação, as variâncias das distâncias entre marcas adjacentes nos grupos de ligação, o estresse, e a correlação de Spearman. Cada um destes critérios foi empregado para avaliar a acurácia de mapas genéticos em FMI sob os efeitos dos diferentes cenários de simulação utilizados: (i) Tamanho da População, (ii) Níveis de Informatividade e (iii) Saturação de Marcas.

4.2.1. Número de Grupos de Ligação Recuperados

Na Figura 6, pode ser observado o número de grupos de ligação recuperados em função do tamanho da população para cada nível de informatividade. A recuperação dos grupos de ligação também foi analisada em separado para cada grau de saturação simulado, considerado de alta saturação para distâncias entre marcas de 5 cM; de média saturação para distâncias entre marcas de 10 cM; e, de baixa saturação para distâncias entre marcas de 15 cM. Os valores numéricos para o número de grupos de ligação recuperados também podem ser observados na Tabela 7.

O número de grupos de ligação recuperados referem-se àquelas repetições em que, pela análise de ligação e mapeamento genético, foram recuperados grupos de ligação com as mesmas marcas e o mesmo número de marcas que o Genoma de Referência. Este critério permite avaliar, indiretamente, a presença de marcas não ligadas e a junção indesejada de segmentos de grupos de ligação diferentes. O número máximo de grupos de ligação recuperados é igual a 100, que corresponde ao número de repetições efetuados para cada cenário. Para ilustrar, nas Figuras de 11 a 14 (Pág. 78 a 81) podem ser observados mapas genéticos com grupos de ligação não recuperados. As figuras de 15 a 24 indicam os casos em que os todos os grupos de ligação foram recuperados.

Verifica-se que para o nível de informatividade s = 2, não é recuperado nenhum grupo de ligação para N = 50, independente do grau de saturação do grupo. De maneira geral, a recuperação de grupos é maior quanto maior o valor de N. Da mesma forma, à medida que a informatividade aumenta, o percentual de recuperação aumenta. Entretanto, pode-se observar que para o tamanho N = 50 condições mais restritivas para a recuperação de grupos de ligação ocorrem na medida que o grau de saturação é reduzido.

O grau de saturação pode ser relacionado aqui com o nível de resolução de marcas no genoma. Na medida que o grau de saturação é reduzido, uma menor cobertura é obtida, e a possibilidade de se encontrar marcas não ligadas aumenta. SILVA et al. (2007) mencionam o nível de resolução de mapas como condicionante para sua aplicabilidade sob diferentes objetivos. Mapas de alta resolução são utilizados em geral como base para a clonagem posicional de genes a partir da localização de QTL. Uma resolução em torno de 1 e 2 cM é necessária para a aplicação de mapeamento físico e procedimentos de clonagem, implicando na utilização de técnicas de mapeamento fino, como o desequilíbrio de ligação, para a obtenção de mapas de alta definição (DARVASI et al. 1993). No melhoramento genético, mapas de menor resolução são menos restritivos desde que esteja disponível, no processo de seleção assistida, marcadores que flanqueiem um QTL, cujo efeito possa ser facilmente detectado (VAN-OOIJEM, 1992).

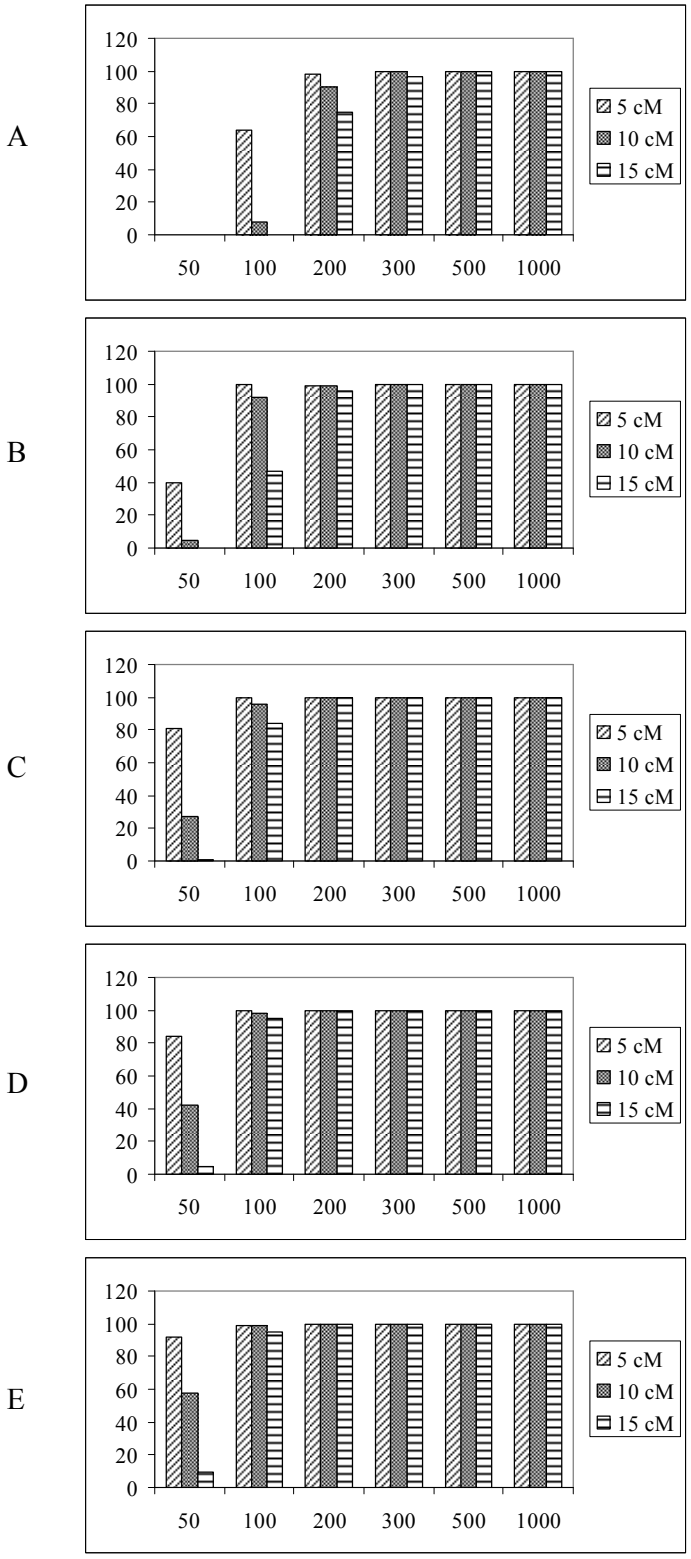


Figura 6. Número de grupos de ligação recuperados em função do tamanho da população N e o grau de saturação para diferentes informatividades – A (s = 2); B (s = 3); C (s = 4); D (s = 5); E (s = 6)

Como já mencionado, o nível de informatividade é fator fundamental na recuperação de grupos de ligação. Assim, para s = 2, populações com N = 50 e 100 não foram eficientes em recuperar os grupos de ligação. Pode-se notar que, com valores de s superiores a 3 o percentual de recuperação aumenta consideravelmente para populações com N = 100 para os três graus de saturação adotados. Para populações com N > 100 a informatividade teve pouca influência na redução de grupos de ligação recuperados, sob os diferentes níveis de informatividade.

A determinação do número de grupos de ligação recuperados também foi utilizada como estratégia de avaliação da acurácia no mapeamento genético em estudos simulados para diferentes delineamentos experimentais. De acordo com FERREIRA et al. (2006), para populações segregantes em cruzamentos endogâmicos com grau de saturação de 10 cM, a formação de grupos de ligação não é afetada a partir do tamanho amostral mínimo de 200 indivíduos. No presente estudo, para N = 200, verifica-se que para um baixo nível de informatividade (s = 2 e 3) ainda ocorre perturbação na recuperação de grupos de ligação.

SILVA et al. (2007) verificaram em populações de RILs que o grau de saturação está intrinsecamente ligado à recuperação de grupos de ligação fidedignos. Para graus de saturação de 10 cM e 20 cM, são necessárias populações com tamanho superiores a 154 indivíduos. No presente trabalho observa-se que para N = 100 com grau de saturação de 10 cM, ainda é possível haver perda de informação com relação ao número de grupos recuperados mesmo para a maior informatividade (s = 6) testada.

4.2.2. Tamanho dos Grupos de Ligação e Distância Média

O tamanho esperado dos grupos de ligação de acordo com o Genoma de Referência é estabelecido para cada grau de saturação. Assim, para alta saturação o tamanho esperado é de 50 cM. Para média saturação o tamanho esperado é de 100 cM e finalmente, o tamanho esperado para baixa saturação

é de 150 cM. Como o número de marcas em cada grupo é igual a 11, o número de intervalos obtido é igual a dez. Dessa forma, o tamanho dos grupos de ligação e as distâncias médias entre marcas adjacentes são múltiplos de 10, de modo que as inferências feitas para estes dois critérios são as mesmas. Ressalta-se que somente foram utilizados para o cômputo das médias dos tamanhos aquelas repetições que tiveram seus grupos de ligação recuperados.

Na Tabela 8, estão representados os tamanhos médios para 100 repetições segundo os diferentes cenários simulados. Foi utilizado o teste de Tukey, a 5 % de significância, para realizar comparações entre os tamanhos dos grupos de ligação obtidos nos diferentes tamanhos populacionais, independentemente dentro dos graus de saturação e níveis de informatividade. Percebe-se que para alta saturação, os tamanhos médios foram invariantes, ao passo que para os demais graus de saturação houve variação em termos de tamanho de grupos formados.

De maneira geral, na medida em que o tamanho amostral aumentou os valores de tamanho do grupo de ligação, e conseqüentemente das distâncias médias entre marcas adjacentes, aproximaram dos valores esperados. Esta aproximação e a invariância das estimativas para altos graus de saturação podem ser explicadas em termos da precisão das estimativas de recombinação em populações com maiores tamanhos amostrais e do nível de resolução para alta saturação. Conforme discutido por ROCHA et al. (2004), ao avaliar intervalos de confiança para estimativas de distâncias genéticas, além do tipo de população envolvida, o número de indivíduos genotipados em uma população de mapeamento influencia a acurácia das estimativas de freqüência de recombinação. Intuitivamente, maiores populações permitem uma melhor amostragem dos eventos de permuta nos gametas e geram dessa forma estimativas mais precisas e acuradas. Com base nas variâncias das estimativas de verossimilhança das freqüências de recombinação e no conteúdo de informação de Fischer, o trabalho de ROCHA et al. (2004) também demonstrou que melhores estimativas são obtidas quando o valor médio das distâncias são menores. Logicamente, uma maior resolução permite o

estabelecimento de mapas mais acurados, conforme também demonstrado no item anterior.

Tabela 8. Tamanho dos grupos ligação para alta (5 cM), média (10 cM) e baixa (15 cM) saturação para o tamanho populacional (N) e informatividade (s). Médias seguidas de mesma letra em cada coluna para cada valor de s não diferem significativamente pelo teste Tukey a 5 %

N	5 cM	Nr	10 cM	Nr	15 cM	Nr
			s = 2			
50	-	0	-	0	-	0
100	49,41 a	64	82,70 b	8	-	0
200	51,85 a	98	96,14 a	90	139,89 b	75
300	48,69 a	100	95,80 a	100	146,42 a	97
500	48,89 a	100	96,77 a	100	146,22 a	100
1000	48,67 a	100	97,32 a	100	146,79 a	100
			s = 3			
50	47,87 a	40	68,34 b	5	-	0
100	49,00 a	100	96,50 a	92	138,20 b	47
200	47,47 a	99	98,52 a	99	147,20 a	96
300	48,71 a	100	96,38 a	100	147,63 a	100
500	47,99 a	100	97,21 a	100	146,91 a	100
1000	48,66 a	100	96,42 a	100	146,97 a	100
			s = 4			
50	49,67 a	81	87,56 b	27	116,93c	1
100	49,28 a	100	99,17 a	96	144,21 b	84
200	48,94 a	100	98,15 a	100	147,32 ab	100
300	48,22 a	100	98,28 a	100	146,91 ab	100
500	48,70 a	100	96,92 a	100	148,85 a	100
1000	49,26 a	100	97,33 a	100	147,98 a	100
			s = 5			
50	49,99 a	84	89,85 b	42	116,50 b	5
100	49,43 a	100	99,46 a	98	146,58 a	95
200	48,37 a	100	97,42 a	100	147,88 a	100
300	49,05 a	100	98,24 a	100	149,52 a	100
500	49,64 a	100	98,93 a	100	149,62 a	100
1000	49,14 a	100	98,29 a	100	149,32 a	100
			s = 6			
50	50,32 a	92	92,16 b	58	124,17 b	9
100	50,64 a	99	100,36 a	99	149,06 a	95
200	49,30 a	100	97,15 a	100	149,34 a	100
300	49,78 a	100	97,74 a	100	151,10 a	100
500	49,05 a	100	98,49 a	100	147,67 a	100
1000	49,31 a	100	99,54 a	100	149,34 a	100

Nr: Número de Grupos de Ligação Recuperados

Com relação ao efeito da informatividade no tamanho dos grupos de ligação formados e nas distâncias médias entre marcas adjacentes, não se verifica nenhuma influência. Nesse caso, a informatividade tem pouco a contribuir, uma vez que a análise do tamanho dos grupos e das distâncias médias entre marcas adjacentes somente foi realizada com base em grupos de ligação recuperados. Verificar-se-á, mais adiante, que a informatividade tem maior influência nos critérios de variância das distâncias entre marcas adjacentes, estresse e correlação de Spearman.

Outro ponto a ser destacado é que o número de grupos de ligação recuperados não influenciou nas estimativas do tamanho e das distâncias médias. Este fato pode ser entendido justamente pelo conceito de grupos de ligação recuperados, que correspondem àqueles grupos de ligação que reproduzem as mesmas marcas ligadas do grupo de ligação correspondente no genoma de referência. Realmente, os grupos de ligação recuperados podem ser associados, em escala, à maior precisão na obtenção de mapas aproximados ao número haplóide do genoma, situação na qual todos os marcadores genéticos estão ligados. A obtenção de mapas genômicos fidedignos está justamente apoiada neste ponto (LIU, 1998; LANZA et al. 2000).

A dispersão amostral dos tamanhos dos grupos de ligação e, conseqüentemente, das distâncias médias de marcas adjacentes são apresentadas por meio de *Box-Plot* nas Figuras 7a, 7b, 7c, 7d, e 7e de acordo com o nível de informatividade s = 2, 3, 4, 5 e 6, respectivamente. Pode-se verificar que o aumento do tamanho populacional tem influência direta na redução da dispersão dos valores dos tamanhos dos grupos de ligação, Por outro lado, a saturação de marcas e o nível de informatividade também não parecem influenciar na dispersão amostral dos valores de tamanho do grupo de ligação e das distâncias médias similares ao genoma de referência. Verificar-se-á, a seguir, que as maiores dispersões amostrais dos tamanhos dos grupos de ligação e das distâncias médias de marcas adjacentes apresentam-se qualitativamente correlacionadas ao comportamento dos critérios de variância e estresse na avaliação da precisão e acurácia do mapeamento genético.

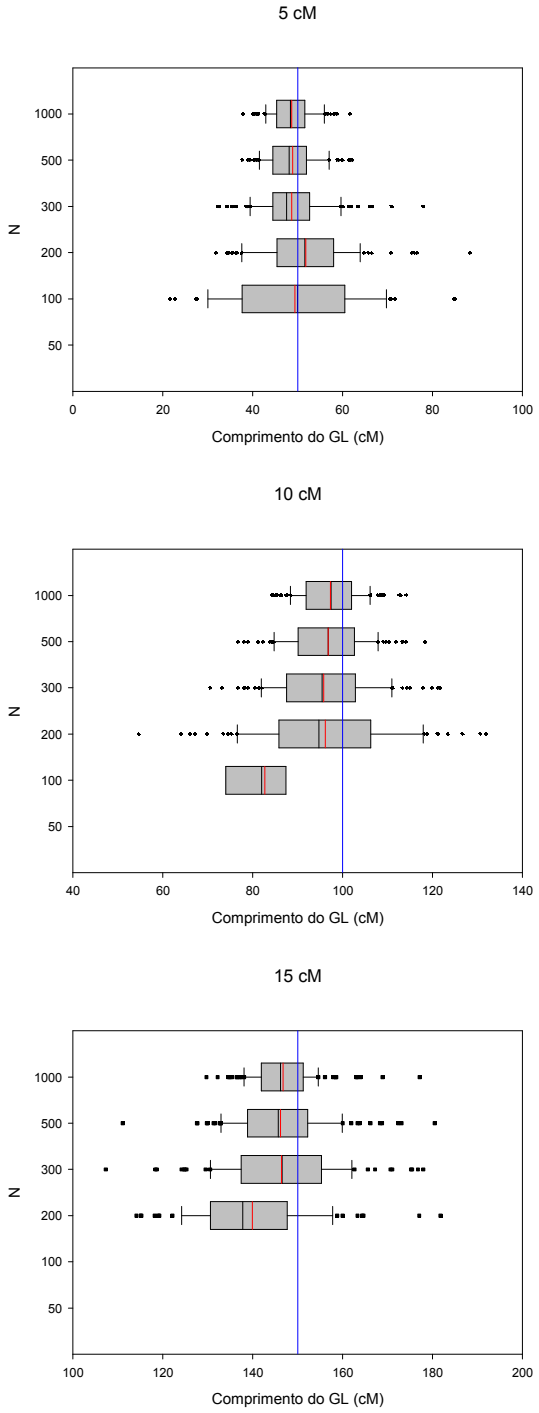


Figura 7a. Box-Plot para a dispersão amostral de estimativas de tamanho dos grupos de ligação e distâncias médias entre marcas adjacentes para tamanho da população (N) sob diferentes graus de saturação. s = 2. A linha azul indica o tamanho do grupo de ligação no genoma de referência

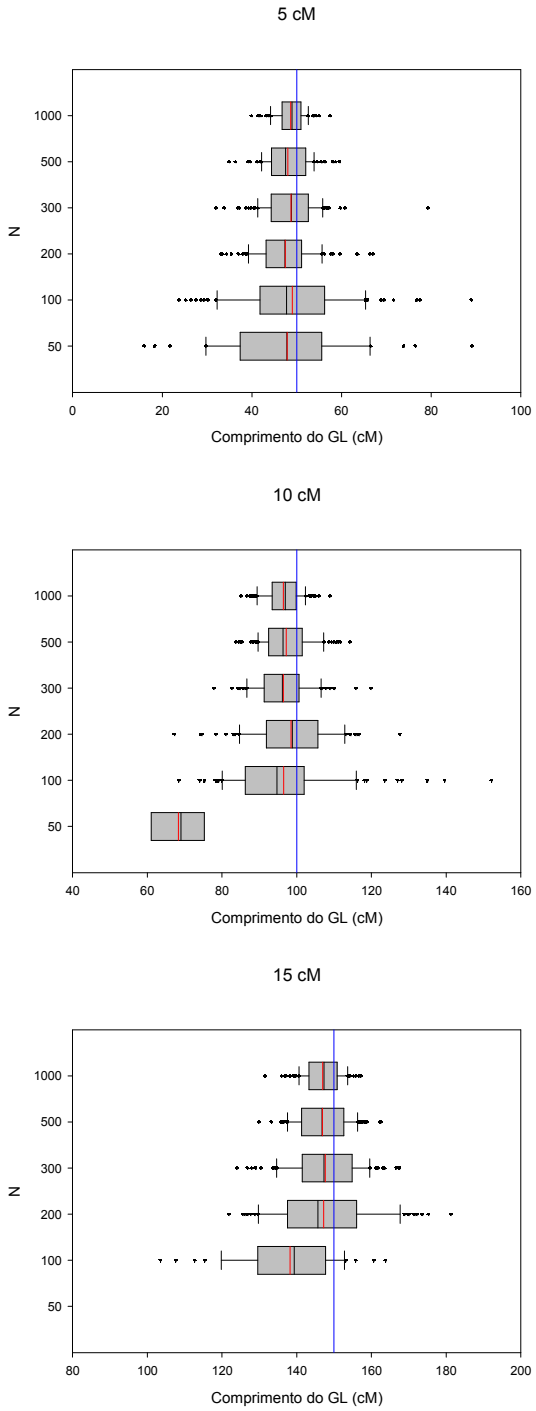


Figura 7b. Box-Plot para a dispersão amostral de estimativas de tamanho dos grupos de ligação e distâncias médias entre marcas adjacentes para tamanho da população (N) sob diferentes graus de saturação. s = 3. A linha azul indica o tamanho do grupo de ligação no genoma de referência

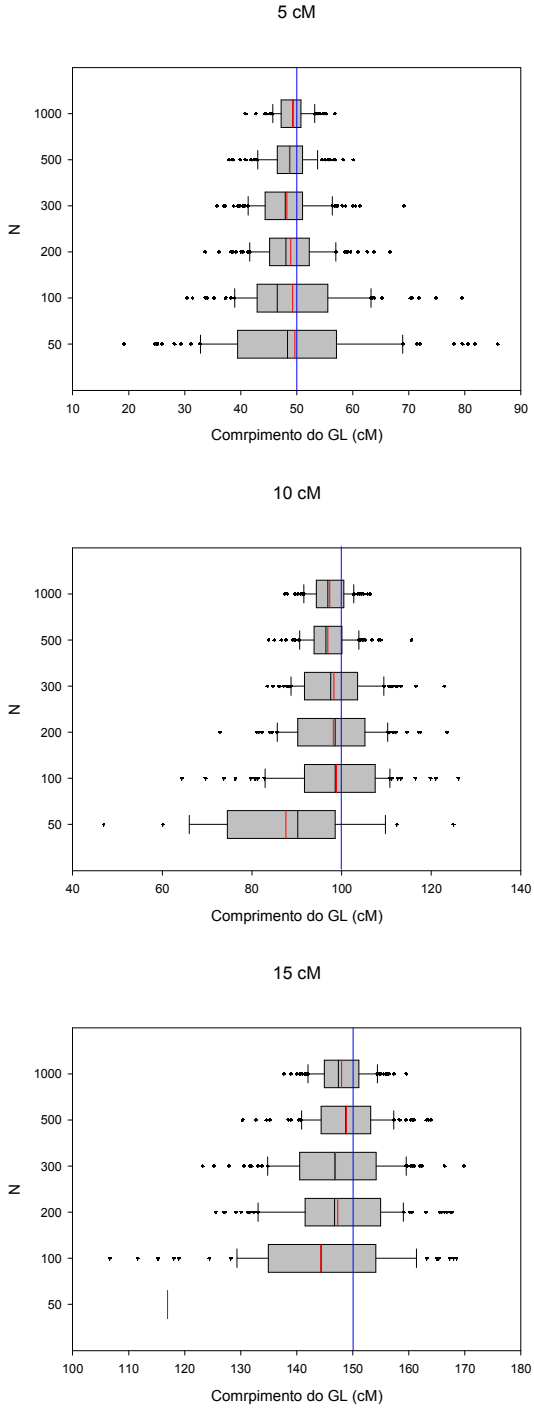


Figura 7c. Box-Plot para a dispersão amostral de estimativas de tamanho dos grupos de ligação e distâncias médias entre marcas adjacentes para tamanho da população (N) sob diferentes graus de saturação. s = 4. A linha azul indica o tamanho do grupo de ligação no genoma de referência

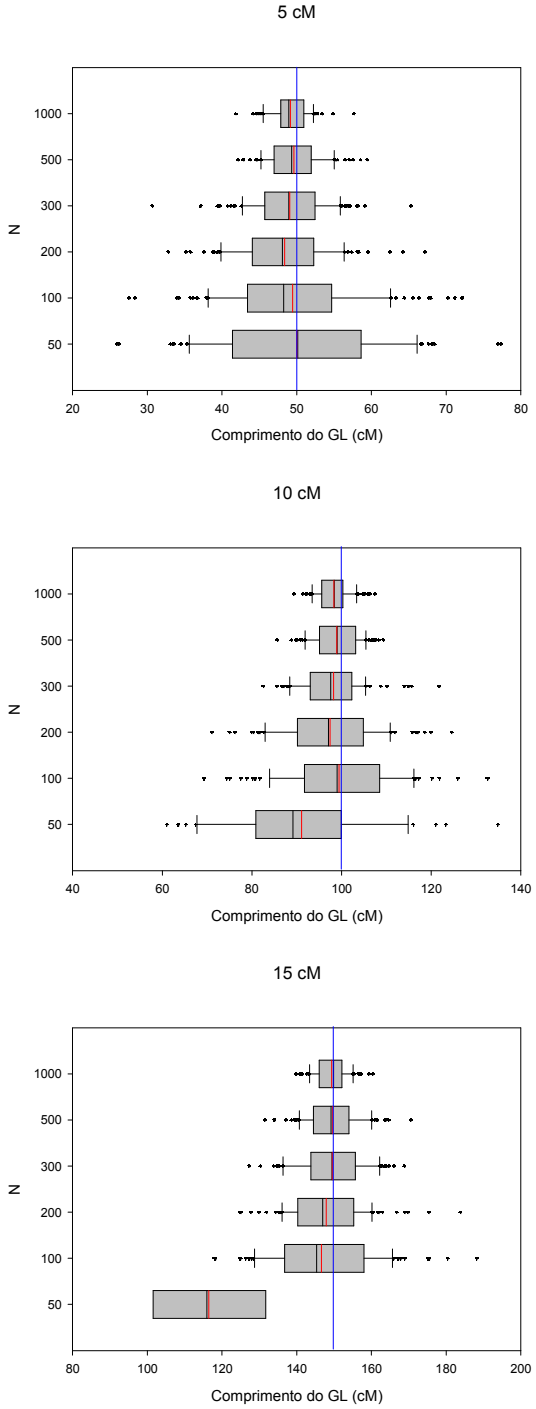


Figura 7d. Box-Plot para a dispersão amostral de estimativas de tamanho dos grupos de ligação e distâncias médias entre marcas adjacentes para tamanho da população (N) sob diferentes graus de saturação. s = 5. A linha azul indica o tamanho do grupo de ligação no genoma de referência

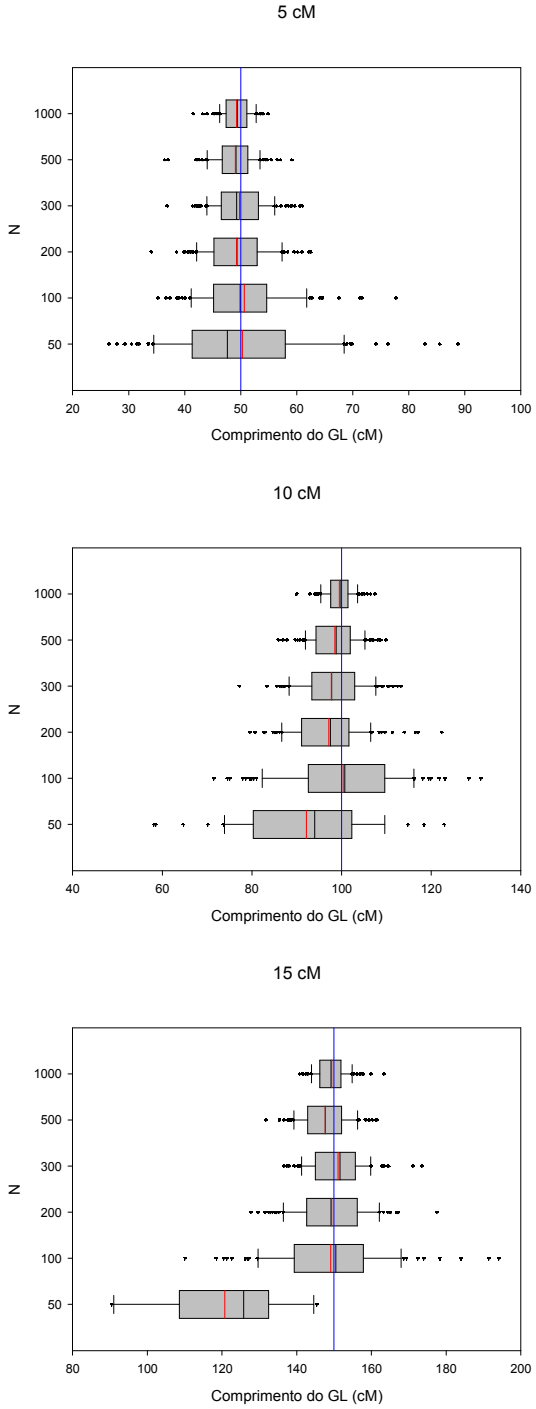
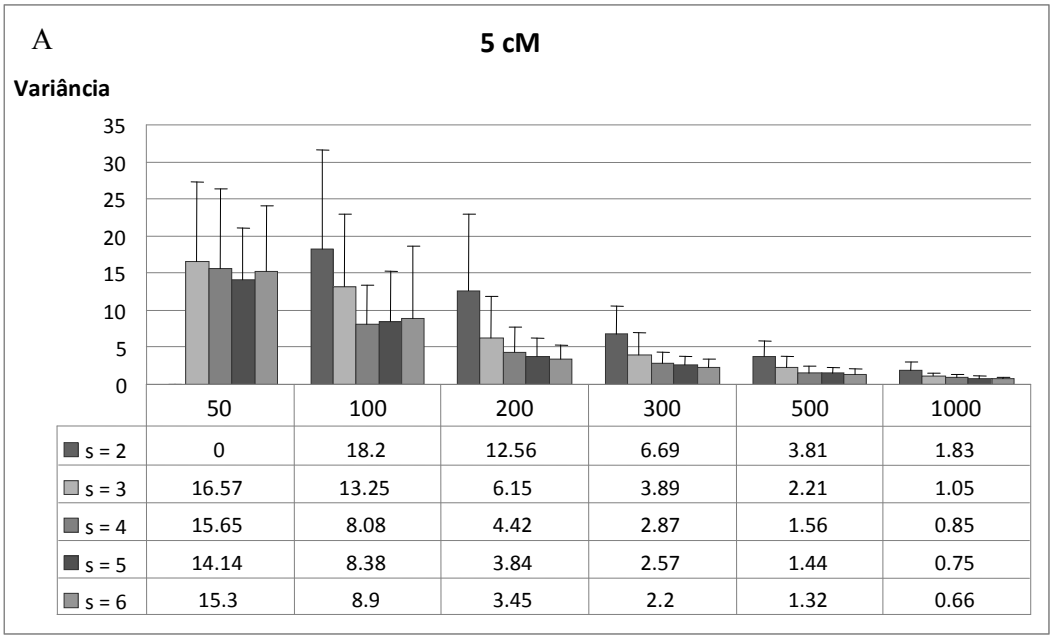


Figura 7e. Box-Plot para a dispersão amostral de estimativas de tamanho dos grupos de ligação e distâncias médias entre marcas adjacentes para tamanho da população (N) sob diferentes graus de saturação. s = 5. A linha azul indica o tamanho do grupo de ligação no genoma de referência

4.2.3. Variância das distâncias entre marcas adjacentes

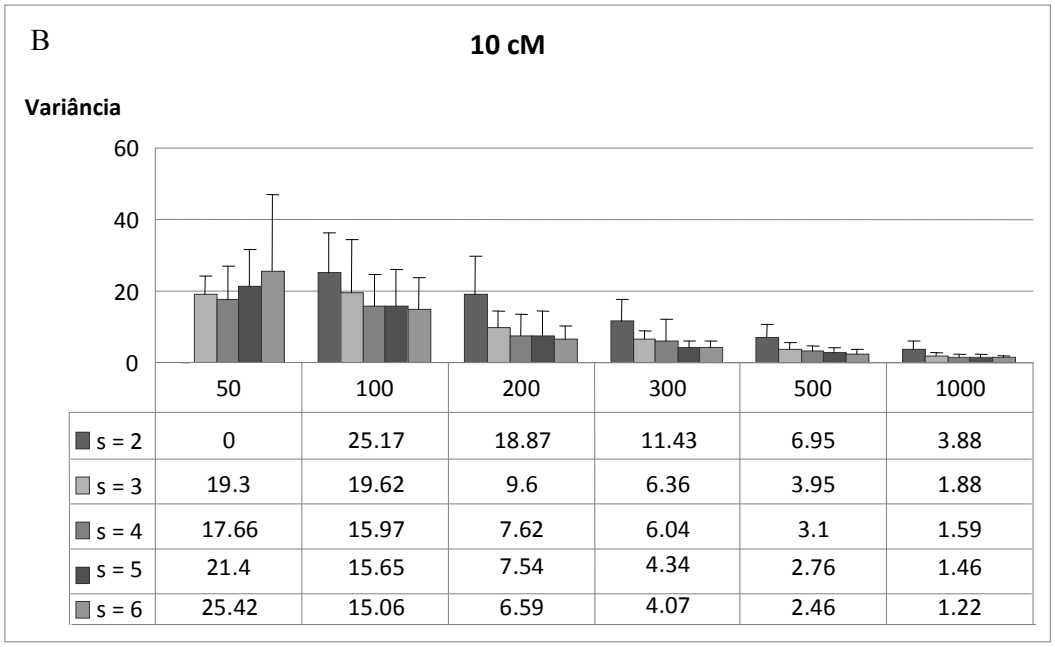
O critério de variância das estimativas das distâncias entre marcas adjacentes é também de interesse na análise de acurácia de mapas genéticos, assim como as estimativas de tamanho e distâncias médias entre marcas adjacentes. A medida da variância torna-se útil principalmente em estudos de simulação em que são utilizados no grupo de ligação original marcadores eqüidistantes (FERREIRA et al. 2006). Conseqüentemente, as variâncias nos grupos originais são nulas. Segundo SILVA (2005), os valores de variâncias refletem a eqüidistância entre marcas adjacentes nos grupos de ligação. Dessa forma, quanto menores os valores das variâncias estimadas, maior será a acurácia do mapa recuperado.

A Figura 8 ilustra o comportamento das variâncias das distâncias médias entre marcas adjacentes de acordo com os diferentes cenários de simulação. Novamente a análise foi realizada somente para os genomas recuperados com base no critério de grupos de ligação recuperados. O teste de Tukey ($p < 0.05$) realizado para as médias das variâncias obtidas para cada repetição é apresentado na Tabela 9.

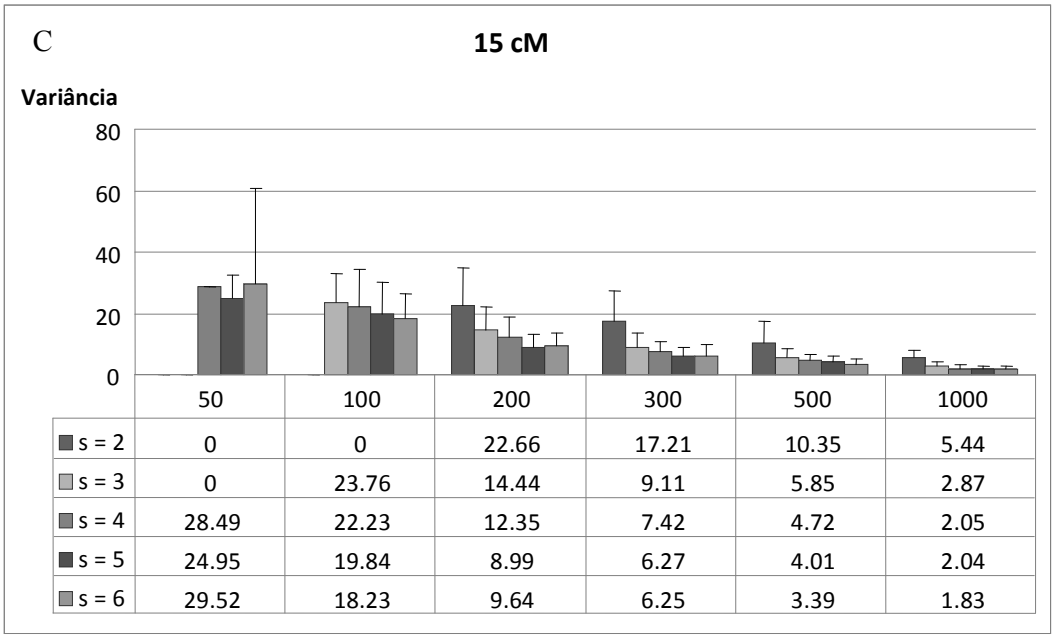


	50	100	200	300	500	1000
s = 2	0	18.2	12.56	6.69	3.81	1.83
s = 3	16.57	13.25	6.15	3.89	2.21	1.05
s = 4	15.65	8.08	4.42	2.87	1.56	0.85
s = 5	14.14	8.38	3.84	2.57	1.44	0.75
s = 6	15.3	8.9	3.45	2.2	1.32	0.66

Figura 8. Redução da variância em função do tamanho da população N (N = 50, 100, 200, 300, 500 e 1000) e do nível de informatividade s para diferentes graus de saturação – A (5 cM); B (10 cM); C (15 cM)



	50	100	200	300	500	1000
s = 2	0	25.17	18.87	11.43	6.95	3.88
s = 3	19.3	19.62	9.6	6.36	3.95	1.88
s = 4	17.66	15.97	7.62	6.04	3.1	1.59
s = 5	21.4	15.65	7.54	4.34	2.76	1.46
s = 6	25.42	15.06	6.59	4.07	2.46	1.22



	50	100	200	300	500	1000
s = 2	0	0	22.66	17.21	10.35	5.44
s = 3	0	23.76	14.44	9.11	5.85	2.87
s = 4	28.49	22.23	12.35	7.42	4.72	2.05
s = 5	24.95	19.84	8.99	6.27	4.01	2.04
s = 6	29.52	18.23	9.64	6.25	3.39	1.83

Figura 8. Continuação: Redução da variância em função do tamanho da população N (N = 50, 100, 200, 300, 500 e 1000) e do nível de informatividade s para diferentes graus de saturação – A (5 cM); B (10 cM); C (15 cM)

Tabela 9. Variância entre marcas adjacentes para alta (5 cM), média (10 cM) e baixa (15 cM) saturação para o tamanho populacional (N) e informatividade (s). Médias seguidas de mesma letra em cada coluna para cada valor de s não diferem significativamente pelo teste Tukey a 5 %

N	5 cM	10 cM	15 cM
	s = 2		
50	-	-	0
100	18,20 a	25,17 a	0
200	12,56 b	18,87 a	22,66 a
300	6,69 c	11,43 b	17,21 b
500	3,81 d	6,95 c	10,35 c
1000	1,83 d	3,88 d	5,44 d
	s = 3		
50	16,57 a	19,30 a	-
100	13,25 b	19,62 a	23,76 a
200	6,15 c	9,60 b	14,44 b
300	3,89 cd	6,36 c	9,11 c
500	2,21 de	3,95 cd	5,85 d
1000	1,05 e	1,88 d	2,87 e
	s = 4		
50	15,65 a	17,66 a	28,49a
100	8,08 b	15,97 a	22,23 b
200	4,42 c	7,62 b	12,35 c
300	2,87 cd	6,04 b	7,42 d
500	1,56 de	3,10 c	4,72 e
1000	0,85 e	1,59 c	2,05 f
	s = 5		
50	14,14 a	21,40 a	24,95 a
100	8,38 b	15,65 b	19,84 a
200	3,84 c	7,54 c	8,99 b
300	2,57 cd	4,34 d	6,27 c
500	1,44 de	2,76 de	4,01 d
1000	0,75 e	1,46 e	2,04 d
	s = 6		
50	15,30 a	25,42 a	29,52 a
100	8,90 b	15,06 b	18,23 b
200	3,45 c	6,59 c	9,64 c
300	2,20 cd	4,07 cd	6,25 d
500	1,32 cd	2,46 d	3,39 e
1000	0,66 d	1,22 d	1,83 e

Como esperado, as variâncias médias obtidas para genomas recuperados reduziram na medida que o tamanho das progênies de FMI aumentaram. Este fato tem sido observado sistematicamente no mapeamento genético para estudos simulados (BHERING, 2007; SILVA, 2005; BARROS, 2007). Em um trabalho conduzido por FERREIRA (2006) foram testados cenários de simulação que envolviam o uso de marcadores com segregação distorcida na estimativa das freqüências de recombinação, para tamanhos de N = 200, 400 e 1000 indivíduos em populações F_2. Da mesma forma, foi detectada a redução das variâncias na medida em que os tamanhos amostrais aumentaram. Fato interessante é que, quando o mapeamento incluiu a informação de marcas com segregação distorcida, as variâncias foram significativamente superiores em relação ao cenário em que estas marcas foram descartadas.

Com relação ao grau de saturação, constatou-se que as variâncias para os grupos de ligação com menor resolução – ou seja, menores graus de saturação – apresentaram maiores variâncias em relação aos grupos de ligação com maior resolução. Esta redução também é esperada, com já demonstrado para os critérios de tamanho de grupos de ligação e distâncias médias entre marcas adjacentes. Novamente, outros trabalhos embasam este comportamento, tanto no mapeamento em populações endogâmicas (SILVA, 2005) quanto em cruzamentos exogâmicos (BHERING & CRUZ, 2008).

Fato interessante foi observado para o comportamento da variância entre marcas adjacentes sob os efeitos dos níveis de informatividade. Observando-se a Figura 8, pode-se verificar que, nas progênies com reduzidos tamanhos amostrais (N = 50 e 100), as variâncias para informatividades variaram de forma mais errática que em progênies com maiores números de indivíduos. As variâncias para níveis de informatividade maiores podem inclusive exceder as variâncias para menores níveis de informatividade. Nota-se, nesse caso, que os efeitos do pequeno tamanho amostral foram mais importantes que os efeitos dos níveis de informatividade. Como já discutido, menores tamanhos amostrais limitam a detecção de permutas nos gametas, o que resulta em uma menor acurácia dos mapas recuperados. Nesse caso,

mesmo populações com alto grau de informatividade não são suficientes para contornar os efeitos do pequeno tamanho amostral.

A influência dos níveis de informatividade e do tamanho amostral na variância de marcas adjacentes em progênies de FIC também foi estudada no trabalho de simulação proposto por BHERING & CRUZ (2008). Da mesma forma que o observado no presente trabalho, as variâncias obtidas reduziram-se na medida em que o tamanho das progênies aumentou. Entretanto, fato interessante foi constatado no estudo de BHERING & CRUZ (2008) com relação à informatividade dos cruzamentos. Nos cruzamentos derivados de genitores completamente informativos as variâncias estimadas foram sensivelmente menores do que as variâncias estimadas para cruzamentos envolvendo genitores ao acaso. Foi detectado neste estudo que em média a estimativa da variância para N = 200 indivíduos em um cruzamento completamente informativo foi igual à estimativa para N = 600 em cruzamentos com genitores tomados ao acaso.

Pode-se recapitular neste ponto, os conceitos de informatividade para locos individuais discutidos anteriormente. Em FIC e FMI, o conceito de informatividade deve ser feito, se possível, de forma *a priori* para a caracterização de genitores na classificação dos cruzamentos que venham a ocorrer em um delineamento experimental. De fato, pode-se realizar uma triagem direta dos genitores em FIC para o estabelecimento de progênies segregantes para o mapeamento genético. Em FMI, apenas o genitor comum pode ser submetido a este tipo de análise, procurando-se tão somente aqueles que apresentem maior heterozigosidade geral. Entretanto a informatividade da população base deve ser investigada para que possa ser garantido um mapa acurado e preciso. Deve ser ressaltado, entretanto, que o grau de polimorfismo, que pode ser mensurado pelo PIC, da população base é importante não somente para FMI mas também para FIC.

4.2.4. Estresse

O valor do estresse médio é utilizado para verificar o grau de ajuste entre as distâncias no genoma original e aquelas obtidas a partir das populações simuladas (FERREIRA et al., 2006). Os valores de estresse são utilizados, conforme mencionado por SILVA (2005), para avaliar distorções de escalas multidimensionais pela simplificação do espaço n-dimensional em duas ou três dimensões. Deve-se ressaltar que, de acordo com SILVA et al. (2007), os valores de estresse entre graus de saturação diferentes não são comparáveis, uma vez que os desvios médios observados para grupos de ligação com baixa resolução são superiores aos desvios médios obtidos com alta resolução, levando a significados quantitativos diferentes sobre o valor do estresse.

Os valores de estresses médios são apresentados na Tabela 10 e podem ser visualizados para os diferentes graus de saturação, como demonstrado na Figura 9. De forma geral, o valor de estresse é reduzido na medida em que o tamanho populacional aumenta, fato comprovado em outros estudos de simulação para mapas genéticos (SILVA, 2005; BARROS, 2007; BHERING, 2007). A redução dos valores de estresse médio com o aumento do tamanho das progênies indica uma maior confiabilidade no mapeamento genético quando maiores tamanhos populacionais são utilizados.

Verifica-se pela Tabela 9 que os valores de estresse médio para alta saturação são maiores que o valor de estresse para baixa saturação. Como comentado, esta comparação direta não pode ser efetuada. Entretanto, o valor do estresse médio pode ser decodificado em termos de desvios médios das distâncias dos grupos de ligação simulados em relação ao genoma original. Nesse caso, os valores entre diferentes níveis de saturação podem ser comparados de forma cautelosa (SILVA, 2005). No trabalho de SILVA (2005), foram utilizados os desvio médios para a comparação dos graus de saturação em populações de RILs, sendo verificado que, na medida em que o grau de saturação foi reduzido, os valores dos desvios aumentaram.

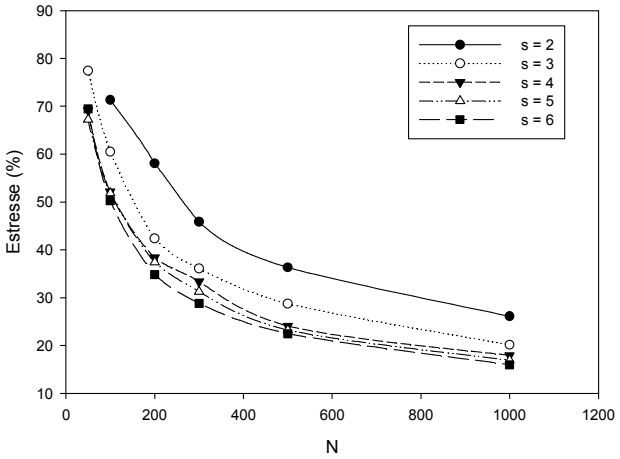
Tabela 10. Estresse médio percentual para alta (5 cM), média (10 cM) e baixa (15 cM) saturação para o tamanho populacional (N) e informatividade (S). Médias seguidas de mesma letra em cada coluna para cada valor de s não diferem significativamente pelo teste Tukey a 5 %

N	5 cM	10 cM	15 cM
	s = 2		
50	-	-	-
100	71,32 a	52,58 a	-
200	58,05 b	39,33 b	30,00 a
300	45,85 c	31,69 c	25,83 b
500	36,31 d	25,52 d	20,13 c
1000	26,10 e	19,27 e	15,42 d
	s = 3		
50	77,44 a	57,40 a	-
100	60,49 b	40,39 b	30,67 a
200	42,34 c	29,65 c	24,07 b
300	36,09 d	24,84 d	19,63 c
500	28,72 e	19,76 e	15,80 d
1000	20,14 f	13,98 f	11,19 e
	s =4		
50	69,15 a	38,70 a	24,25*
100	52,15 b	37,23 a	29,60 a
200	38,34 c	26,37 b	22,58 b
300	33,32 d	23,04 c	18,05 c
500	24,04 e	17,40 d	14,24 d
1000	17,92 f	12,62 e	9,39 e
	s = 5		
50	67,24 a	42,76 a	40,16 a
100	51,87 b	36,62 b	28,53 b
200	37,39 c	26,04 c	19,61 c
300	31,25 d	20,64 d	16,50 d
500	23,20 e	16,19 e	13,16 e
1000	16,91 f	11,78 f	9,32 f
	s = 6		
50	69,49 a	44,99 a	30,79 a
100	50,28 b	36,69 b	27,99 a
200	34,80 c	24,82 c	20,25 b
300	28,80 d	20,16 d	16,10 c
500	22,45 e	15,41 e	12,15 d
1000	15,97 f	10,65 f	8,85 e

O comportamento do estresse sob o efeito da informatividade pode ser observado na Figura 9. Conforme demonstrado, populações com nível de informatividade s = 2 apresentam maiores valores de estresse em relação aos

demais níveis para todos os tamanhos de população. Este comportamento foi claramente observado para os diferentes graus de saturação.

A



B

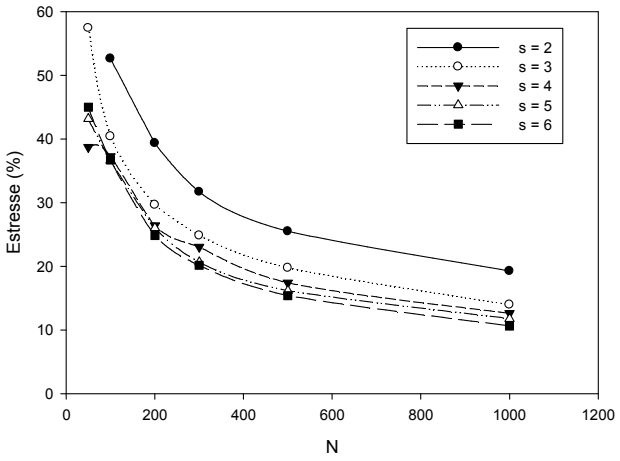


Figura 9. Estresse percentual médio em função do Tamanho da População (N) para os diferentes níveis de informatividade. (A) Alta Saturação; (B) Média Saturação; (C) Baixa Saturação

C

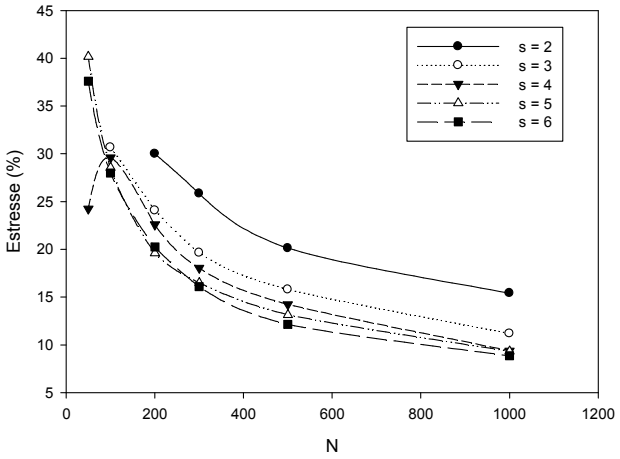


Figura 9 – Continuação. Estresse percentual médio em função do Tamanho da População (N) para os diferentes níveis de informatividade. (A) Alta Saturação; (B) Média Saturação; (C) Baixa Saturação

4.2.5. Correlação de Spearman

Os valores de correlação de Spearman são apresentados na Tabela 11 para os diferentes cenários simulados. A correlação de Spearman é utilizada para identificar inversões de posições de marcas nos grupos de ligação. As figuras de 11 a 14, 16 e 17 ilustram mapas de ligação em que ocorrem inversões e alterações no ordenamento de marcas. Valores iguais a unidade indicam que a ordem das marcas nos grupos de ligação simulados não foi alterada em relação ao genoma de referência (Figuras 15, 18 a 20). Conforme exposto na Tabela 11, as estimativas da correlação de Spearman convergem para a unidade na medida em que ocorre o aumento do tamanho populacional, independente do nível de informatividade.

Mais especificamente, a convergência das correlações de Spearman pode ser comprovada a partir da análise da Tabela 11, com o auxílio do teste de Tukey. Quando o nível de informatividade é s=2, somente populações com N > 300 apresentam correlação de Spearman iguais à unidade. No entanto, para o nível de informatividade s = 6, populações com N ≥ 100 apresentam correlação de Spearman igual à unidade.

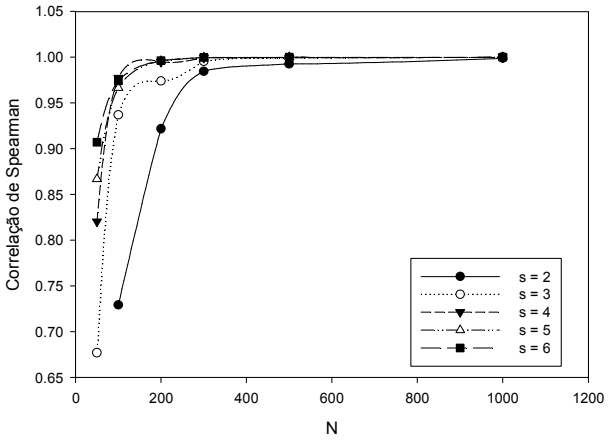
Tabela 11. Correlação de Spearman para alta (5 cM), média (10 cM) e baixa (15 cM) saturação para o tamanho populacional (N) e informatividade (S). Médias seguidas de mesma letra em cada coluna para cada valor de s não diferem significativamente pelo teste Tukey a 5 %

N	5 cM	10 cM	15 cM
	s = 2		
50	-	-	-
100	0,7476 c	0,8739 c	-
200	0,9218 b	0,9789 b	0,9882 b
300	0,9846 a	0,9903 ab	0,9933 ab
500	0,9927 a	0,9984 a	0,9970 a
1000	0,9989 a	0,9995 a	1,0000 a
	s = 3		
50	0,7873 c	0,8782 c	-
100	0,9369 b	0,9760 b	0,9874 b
200	0,9740 a	0,9966 a	0,9965 a
300	0,9955 a	0,9999 a	0,9999 a
500	0,9995 a	1,000 a	1,0000 a
1000	1,0000 a	1,000 a	1,0000 a
	s =4		
50	0,8203 b	0,9475 b	0,9909 c
100	0,9775 a	0,9919 a	0,9933 b
200	0,9951 a	0,9985 a	1,0000 a
300	0,9995 a	0,9985 a	1,0000 a
500	0,9994 a	1,0000 a	1,0000 a
1000	1,0000 a	1,0000 a	1,0000 a
	s = 5		
50	0,8669 b	0,9604 b	0,9782 c
100	0,9666 a	0,996 a	0,9949 b
200	0,9961 a	0,9976 a	0,9999 a
300	0,9997 a	1,0000 a	1,0000 a
500	1,0000 a	1,0000 a	1,0000 a
1000	1,0000 a	1,0000 a	1,0000 a
	s = 6		
50	0,9071 b	0,9389 b	0,9697 b
100	0,9745 a	0,9932 a	0,9969 a
200	0,9963 a	0,9985 a	1,0000 a
300	0,9998 a	1,0000 a	1,0000 a
500	1,0000 a	1,0000 a	1,0000 a
1000	1,0000 a	1,0000 a	1,0000 a

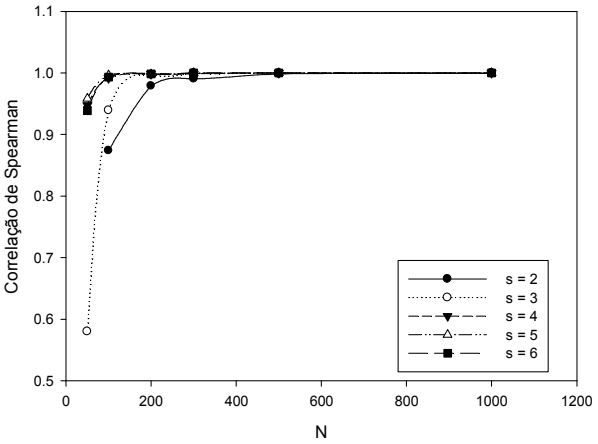
O comportamento de convergência da correlação de Spearman em direção à unidade, sob os efeitos do nível de informatividade e para diferentes graus de resolução pode ser mais apropriadamente observado na Figura 10. Pode-se inferir que o aumento da informatividade leva a uma convergência mais acentuada na medida em que o tamanho das progênies é aumentado. Conclui-se, portanto, que o nível de informatividade, o tamanho das progênies de meios-irmãos e grau de saturação são importantes aspectos a serem levados em conta para o ordenamento e a ocorrência de inversão de marcas em mapas genéticos.

O efeito do tamanho populacional no ordenamento e ocorrência de inversão de marcas moleculares em mapas genéticos também foi investigado, através de modelos simulados, em populações de RILs (SILVA et al. 2007), F_2 (BARROS, 2007) e em Famílias de Irmãos Completos (BHERING & CRUZ, 2008). Constata-se, nestes trabalhos, que assim como o aumento do tamanho da população de mapeamento favorece a ocorrência de um menor número de marcas invertidas, um maior do grau de saturação de marcas nos grupos de ligação resulta em valores da correlação de Spearman mais próximos da unidade. Em populações exogâmicas, além do tamanho da população de mapeamento e do grau de saturação, deve-se também levar em conta o nível de informatividade na população base, no caso de progênies de meios-irmãos, ou a informatividade dos genitores, no caso de progênies de irmãos completos.

A



B



C

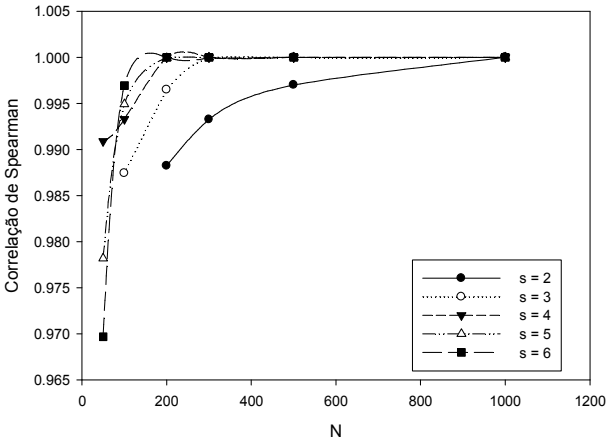


Figura 10. Correlação de Spearman média em função do Tamanho da População (N) para os diferentes níveis de informatividade. (A) Alta Saturação; (B) Média Saturação; (C) Baixa Saturação

ANEXO – MAPAS DE LIGAÇÃO

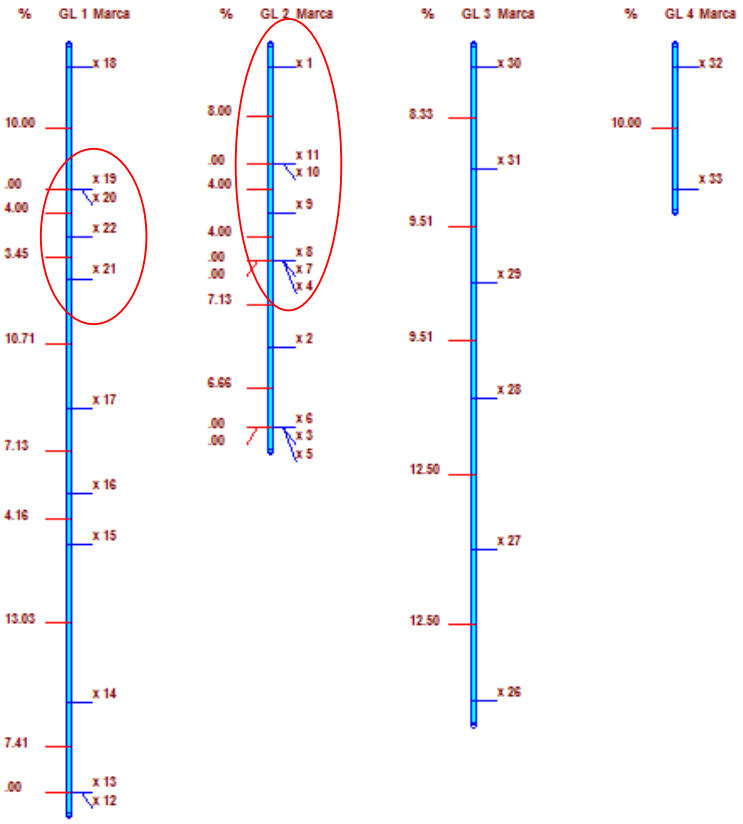


Figura 11. Mapa de Ligação recuperado para tamanho da população (N = 50) e nível de informatividade (s = 3). A figura ilustra a disjunção das marcas X 32 (=C32) e X 33 (=C33) do grupo de ligação 3, formando o grupo de ligação 4. As elipses vermelhas indicam casos de inversão do ordenamento de marcas.

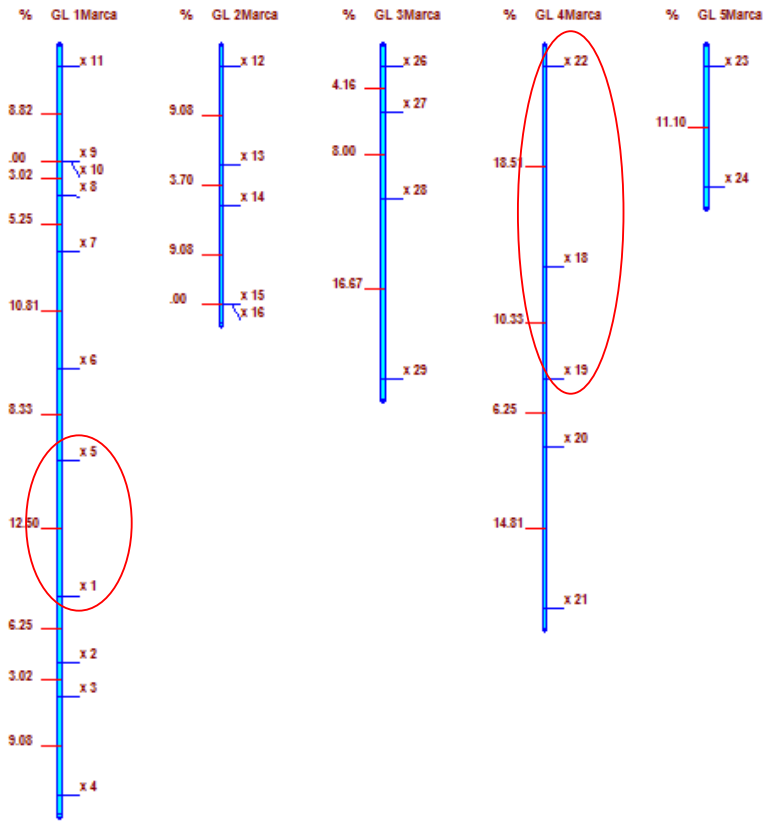


Figura 12. Mapa de Ligação recuperado para tamanho da população (N = 50) e nível de informatividade (s = 4). A figura ilustra diferentes disjunções de grupos que ocorrem para este cenário. As elipses vermelhas indicam casos de inversão do ordenamento de marcas.

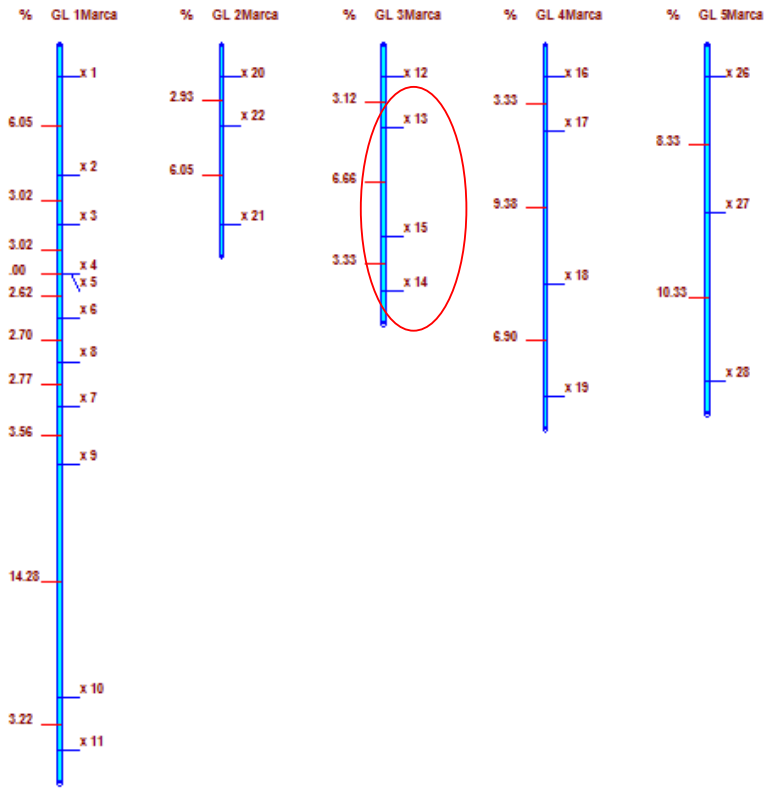


Figura 13. Mapa de Ligação recuperado para tamanho da população (N = 50) e nível de informatividade (s = 5). A figura ilustra diferentes disjunções de grupos que ocorrem para este cenário. A elipse vermelha indica um caso de inversão do ordenamento de marcas.

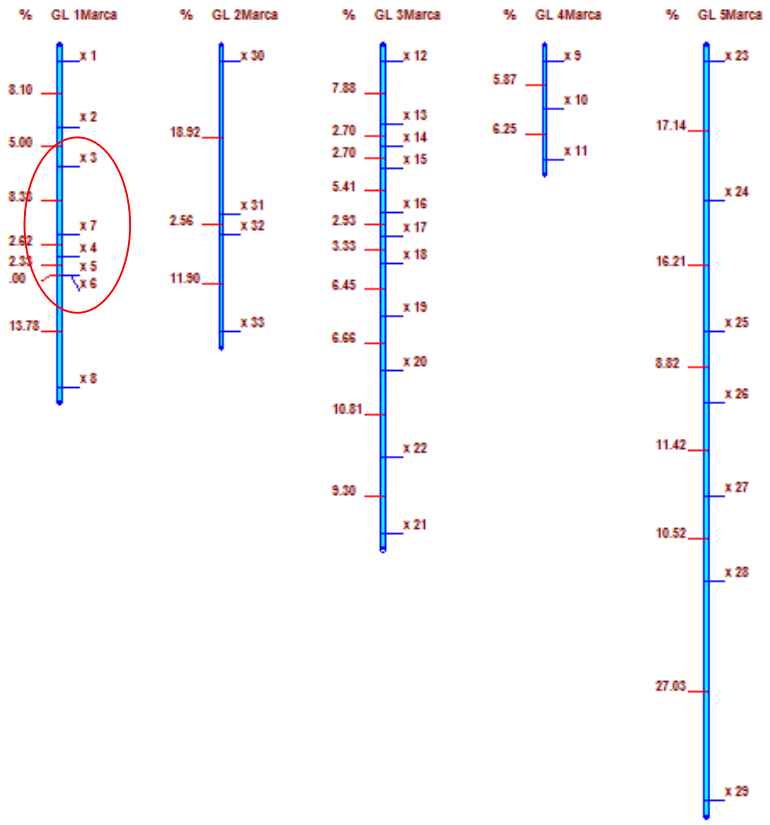


Figura 14. Mapa de Ligação recuperado para tamanho da população (N = 50) e nível de informatividade (s = 6). A figura ilustra diferentes disjunções de grupos que ocorrem para este cenário. A elipse vermelha indica um caso de inversão do ordenamento de marcas.

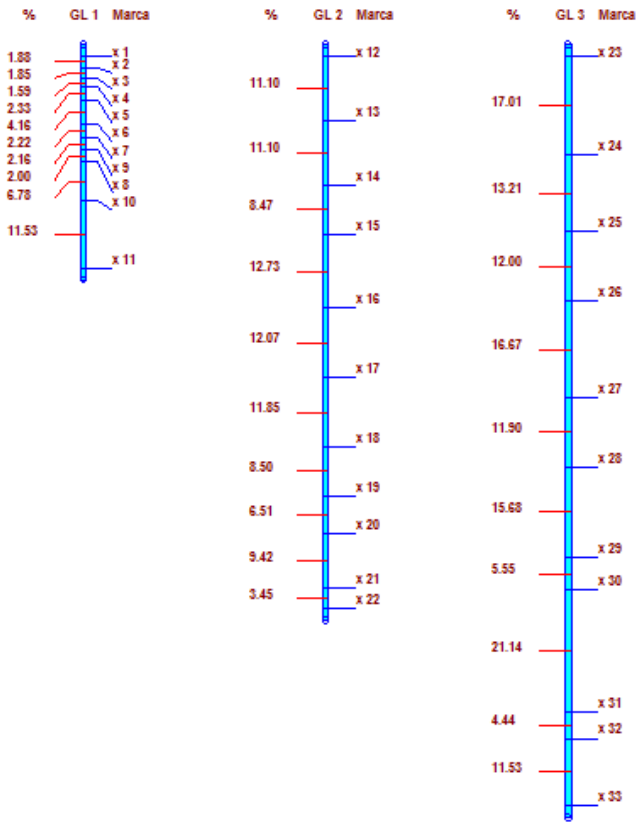


Figura 15. Mapa de Ligação recuperado para tamanho da população (N =200) e nível de informatividade (s = 2).

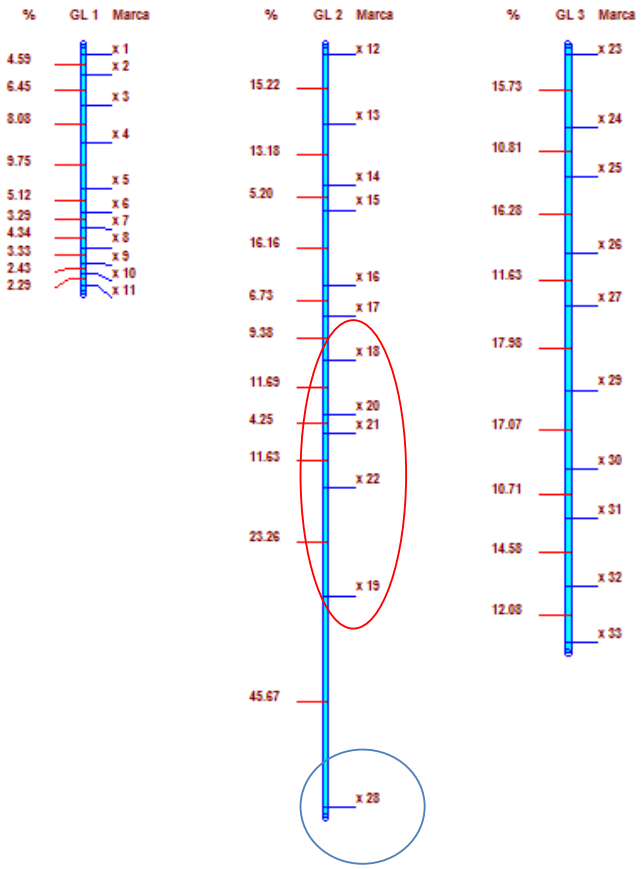


Figura 16. Mapa de Ligação recuperado para tamanho da população (N =200) e nível de informatividade (s = 3). A elipse vermelha indica um caso de inversão do ordenamento de marcas. A elipse azul indica um caso de junção da marca X 28 (=C36) no grupo de ligação 2 que pertence ao grupo de ligação 3 do genoma de referência.

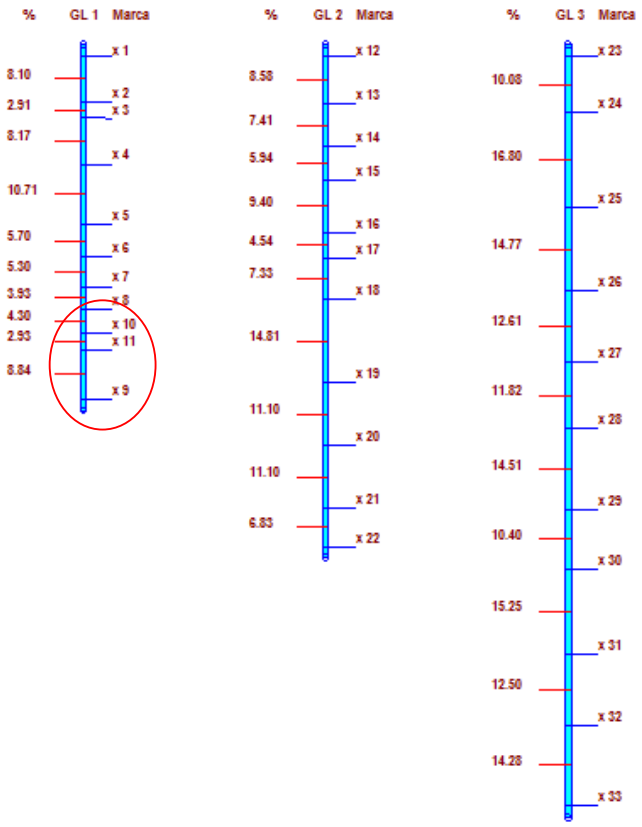


Figura 17. Mapa de Ligação recuperado para tamanho da população (N =200) e nível de informatividade (s = 4). A elipse vermelha indica um caso de inversão do ordenamento de marcas.

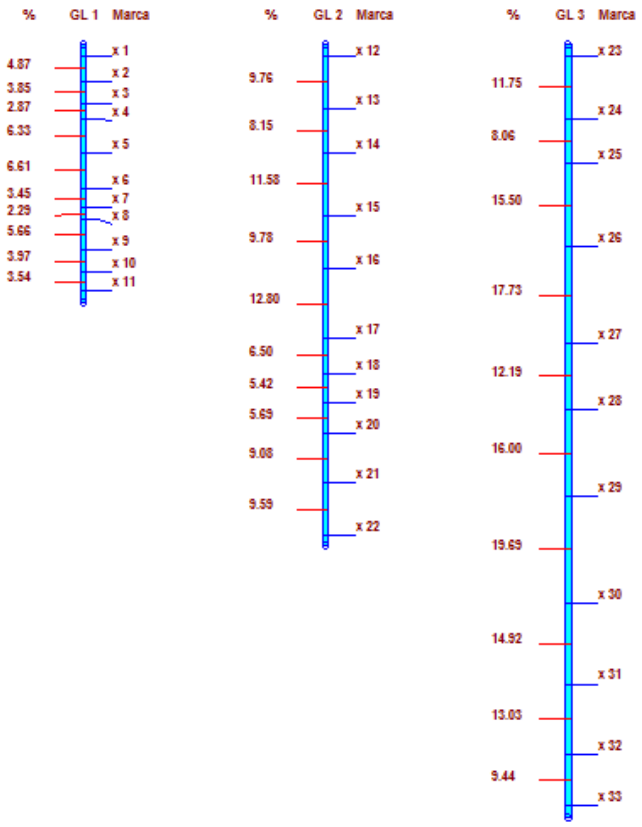


Figura 18. Mapa de Ligação recuperado para tamanho da população (N =200) e nível de informatividade (s = 5).

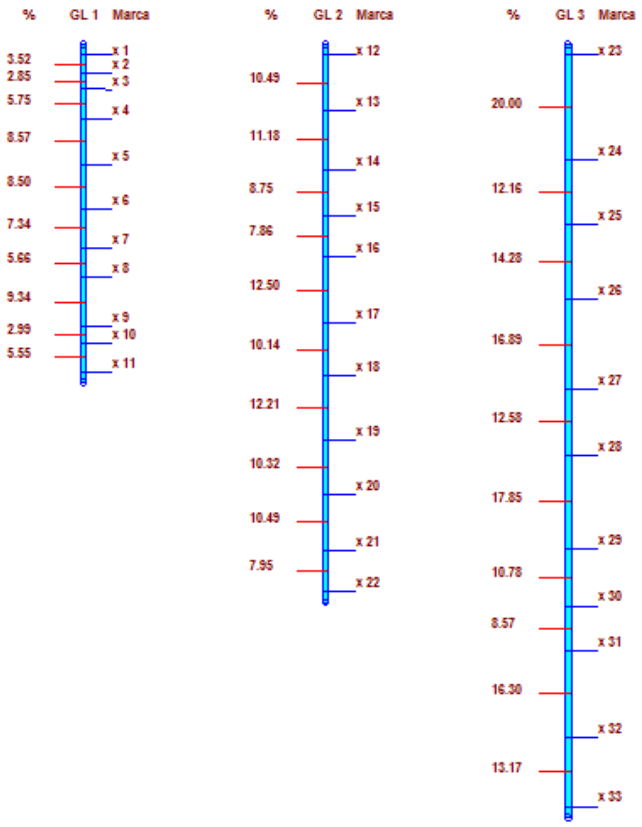


Figura 19. Mapa de Ligação recuperado para tamanho da população (N =200) e nível de informatividade (s =6).

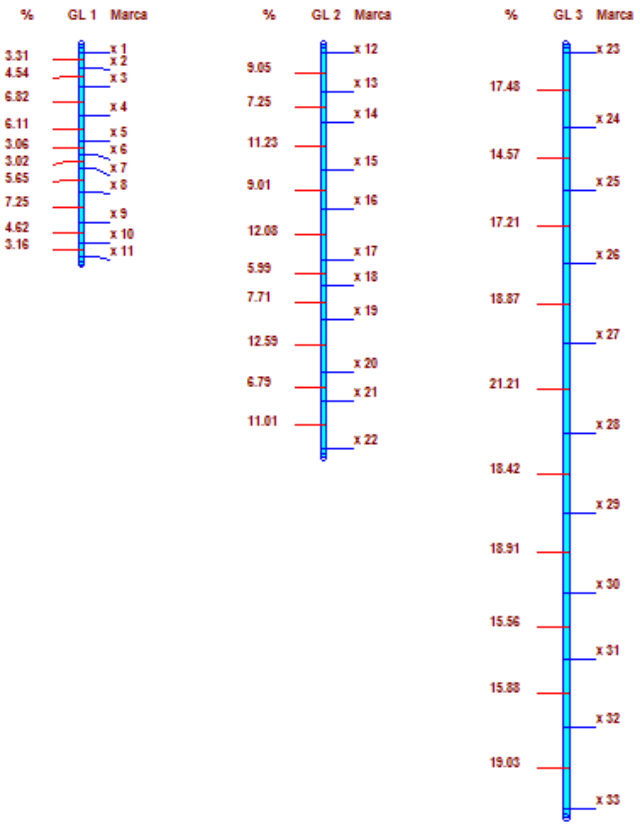


Figura 20. Mapa de Ligação recuperado para tamanho da população (N =1000) e nível de informatividade (s =2).

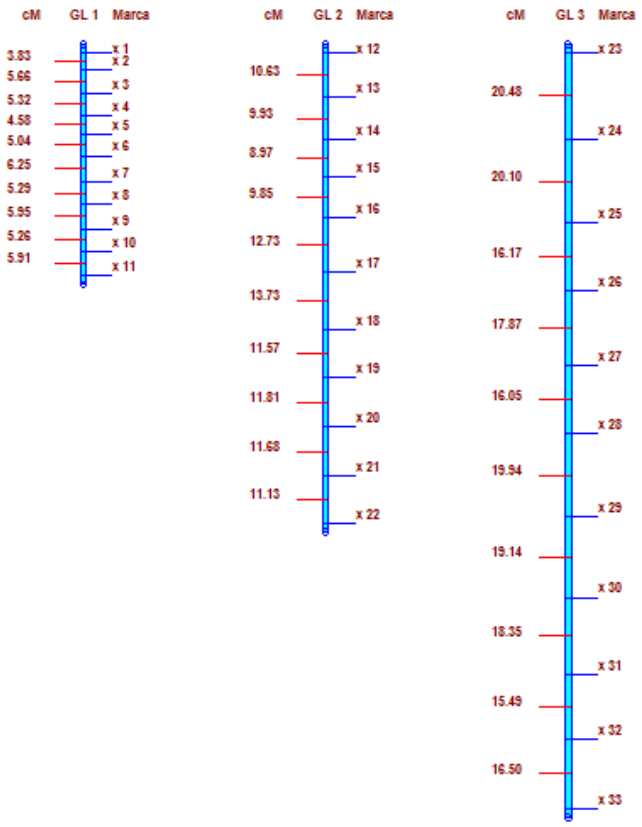


Figura 21. Mapa de Ligação recuperado para tamanho da população (N =1000) e nível de informatividade (s =3).

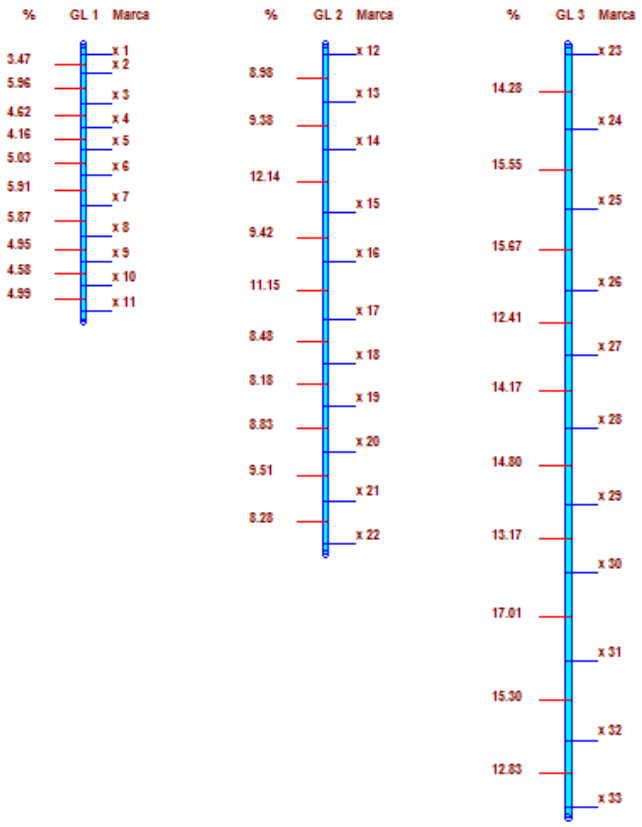


Figura 22. Mapa de Ligação recuperado para tamanho da população (N =1000) e nível de informatividade (s =4)

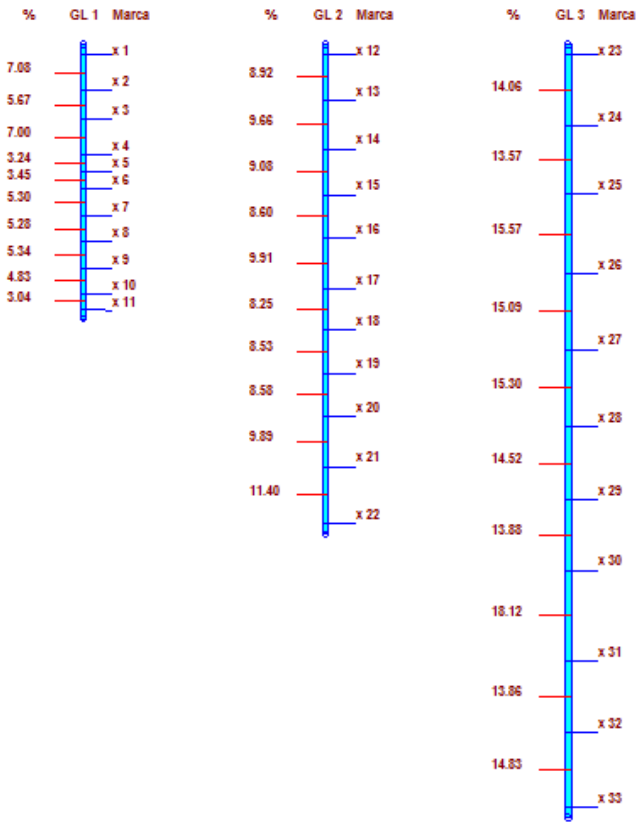


Figura 23. Mapa de Ligação recuperado para tamanho da população (N =1000) e nível de informatividade (s =5)

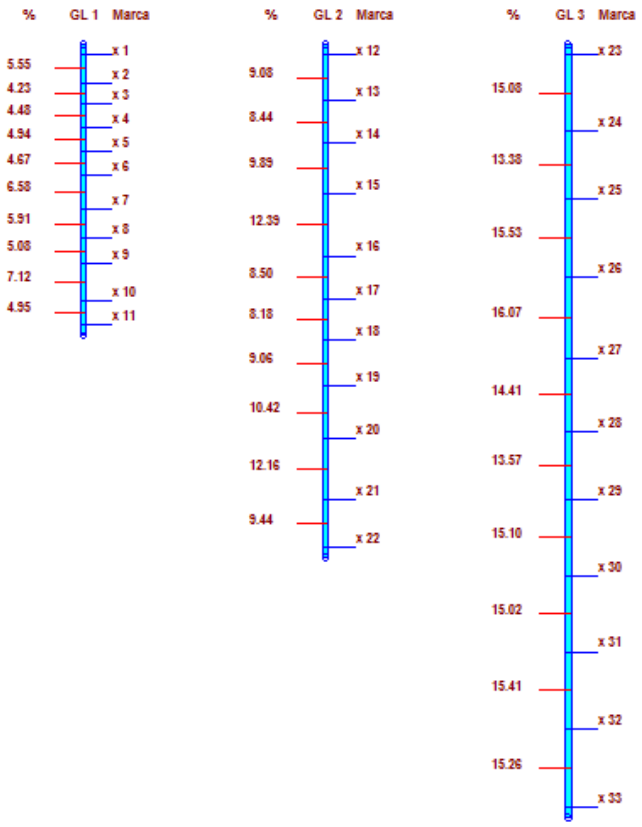


Figura 24. Mapa de Ligação recuperado para tamanho da população (N =1000) e nível de informatividade (s =6)

5. CONCLUSÕES

Além do tamanho da população de mapeamento e do grau de saturação como fatores fundamentais de influência na acurácia de mapas genéticos, em cruzamentos exogâmicos também deve ser levado em conta o polimorfismo na população base. Em um estudo recente, BHERING & CRUZ (2008) verificaram em progênies de irmãos completos que a informatividade dos genitores determina de forma marcante a acurácia de mapas genéticos. No presente trabalho as seguintes conclusões podem ser feitas com base nos resultados encontrados:

(1) Populações de Meios-Irmãos com N = 50 e 100 indivíduos são mais restritivas mesmo para um alto nível de polimorfismo;

(2) Populações de Meios-Irmãos com N = 200 indivíduos podem ser utilizadas para o mapeamento genético desde que o nível de polimorfismo amostrado seja superior a um valor de PIC = 0,5926, que corresponde ao número de 3 alelos com freqüências iguais na população base. Uma vez que o valor de PIC atinge o máximo para um dado número de alelos, quando estes são equi-freqüentes, deve ser investigada com cautela a possibilidade do uso de populações com N = 200 para três alelos segregando na progênie.

(3) Quando o nível de polimorfismo na população é mínimo, ou seja, apenas dois alelos são amostrados, as condições de mapeamento tornam-se bastante restritivas, sendo praticamente inviável o estabelecimento de mapas com N = 50, 100 e 200 indivíduos. Mesmo com tamanhos populacionais com N > 300, as condições de mapeamento são restritivas, principalmente para mapas com baixo grau de saturação, quando ocorre um aumento relativo de inversões de marcas no mapa de ligação;

(4) O grau de saturação de marcas teve influência fundamental na obtenção de mapas acurados. Mapas menos saturados apresentaram menor recuperação de genomas, maiores estimativas de correlação de Spearman e maiores variâncias das distâncias entre marcas adjacentes, principalmente para baixa informatividade (s = 2) nos tamanhos de N = 50, 100 e 200. Para

alta informatividade (s = 6) a recuperação de genomas foi baixa para N = 50 em qualquer grau de saturação.

Como conclusão geral, pode-se dizer que não é recomendável o uso de populações com N = 50 e 100 mesmo com alto nível de informatividade. Para N = 200 é possível obter mapas com certa fidelidade desde que o número de alelos segregando na população base seja igual ou maior do que 4, à semelhança do que ocorre em FIC para acasalamentos entre genitores completamente informativos.

6. REFERÊNCIAS BIBLIOGRÁFICAS

AMOS, C.I. & ELSTON, R. C. Robust methods for detection of genetic linkage for quantitative data from pedigrees. **Genet. Epidemiol.** 6:349-360. 1989.

BARROS, W.S. **Genotipagem seletiva e outras estratégias de amostragem no mapeamento genético e na detecção de QTL em populações F_2 simuladas**. 2007. 158p. Tese (Doutorado em Genética e Melhoramento) Universidade Federal de Viçosa, Viçosa.

BEARZOTI E. Mapeamento de QTL. In: PINHEIRO, J.B.; CARNEIRO, I.F. **Análises de QTL no melhoramento de plantas: segunda jornada em genética e melhoramento de plantas**. Goiânia: FUNAPE. 63-209. 2000.

BHERING, L.L. & CRUZ, C.D. Tamanho de população ideal para mapeamento genético em famílias de irmãos completos. **Pesq. Agrop. Bras.** 43(3):379-385. 2008.

BHERING, L.L.; CRUZ, C.D. e GOOD GOD, P.I.V. Estimativa de freqüência de recombinação no mapeamento genético de família de irmãos completos. **Pesq. Agropec. Bras.** 43(3):363-369. 2008.

BOTSTEIN, D.; WHITE, R.L.; SKOLNICK, M and DAVIS, R.W. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. **American Journal of Human Genetics**. 32(3):314-331. 1980.

CHURCHILL G.A. & DOERGE, R.W. Empirical Threshold Values for Quantitative Trait Mapping. **Genetics.** 138: 963-971. 1994.

CHURCHILL G.A. & DOERGE R.W. **Mapping quantitative trait loci in experimental populations**. In: Paterson, A.H. Molecular dissection of complex traits. CRC Press: New York. p.31-41. 1998.

CRUZ, C.D. **Programa Genes/Versão Windows: Aplicativo computacional em genética e estatística.** 2. ed. Viçosa, MG: Imprensa Universitária. 648p.2001.

CRUZ, C.D. **Programa para análise de dados moleculares e quantitativos** – GQMOL. Viçosa: UFV, 2004.

DA, Y. & LEWIN, H.A. Linkage information content and efficiency of full-sib and half-sib designs for gene mapping. **Theor. Appl. Genet**. 90:699-706.1995.

DARVASI, A.; WEINREB, A.; MINKE, V.; WELLER, J.I. and SOLLER, M. Detecting marker-QTL linkage and estimating QTL gene effect and map location using a saturated genetic map. **Genetics**, 134:943-951. 1993.

DOERGE, R.W. Constructing genetic maps by rapid chain delineation. **Journal of Quantitative Trait Loci**, 2:121-132, 1996.

DOERGE, R.W.; WEIR, B.S. and ZENG, Z-B. Statistical issues in the search for genes affecting quantitative traits in experimental populations. **Statist. Sci.** 12 (3): 195-219. 1997.

DOERGE, R.W. Mapping and analysis of quantitative trait loci in experimental populations. **Nat. Rev. Genetics.** 3: 43-52. 2002.

ERICKSON D.L.; FENSTER, C.B.; STENØIEN, H.K. and PRICE, D. Quantitative trait locus analyses and the study of evolutionary process. **Molecular Ecology.** 13: 2505-2522. 2004.

FALCONER, D.S. & MACKAY, T.F.C. **Introduction to quantitative genetics.** 4th ed. Harlow, UK: Longman. 464p. 1996.

FERREIRA, M.E. & GRATTAPAGLIA, D. **Introdução ao uso de marcadores RAPD e RFLP em análise genética.** Brasília: EMBRAPA-CENARGEN. 220p. 1995.

FERREIRA, A.; SILVA, M.F.; SILVA, L.C. and CRUZ, C.D. Estimating the effects of population size and type on the accuracy of genetic maps. **Genetics and Molecular Biology**. 29:187-192. 2006.

FULKER, D.W. & CARDON, L.R. A sib-pair approach to interval mapping of quantitative trait loci. **Am. J. Hum. Genet.** 54:1092-1103. 1994.

GOLDGAR, D.E. Multipoint analysis of human quantitative genetic variation. **Am. J. Hum. Genet.** 47: 957-967. 1990.

GRATTAPAGLIA, D. & 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.

GUO, X. & ELSTON, R.C. Linkage information content of polymorphic genetic markers. **Human Heredity**. 49:112-118. 1999.

GUO, X.; OLSON, J.M.; ELSTON, R.C. and NIU, TIANHUA. The linkage information content value of polymorphism genetic markers in model-free linkage analysis. **Human heredity**. 53:45-48. 2002.

HALDANE, J.B.S. The combination of linkage values, and the calculation of distance between 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.

JANSEN, R.C. A general mixture model for mapping quantitative trait loci by using molecular markers. **Theor. Appl. Genet.** 85: 252-260. 1992.

KAO, CH.; ZENG, Z.B. and TEASDALE, R.D. Multiple interval mapping for quantitative trait loci. **Genetics**. 152(3):1203-1216. 1999.

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

LANDER, E.S. & BOTSTEIN, D. Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. **Genetics**. 121(1):185-199.1989.

LANDER, E. & KRUGLYAK, L. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. **Nature Genet.** 11: 241-246. 1995.

LANZA, M.A.; GUIMARÃES, C.T. and SCHUSTER, I. Aplicações de marcadores moleculares no melhoramento genético. **Informe Agropecuário**, Belo Horizonte. 21(204):97-108. 2000.

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

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

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

MARTINEZ, M.L.; VUKASINOVIC, N. and FREMAN, G. Random model approach for QTL mapping in half-sib families. **Genet. Sel. Evol.** 31:319-340. 1999.

MARTINEZ, M.L. & VUKASINOVIC, N. Algoritmo para cálculo da proporção de genes idênticos por descendência, para mapear QTL em famílias de meio-irmãos. **Revista Brasileira de Zootecnia.** 29:443-451. 2000.

NONES, K.; LEDUR, M.C.; RUY, D.C.; BARON, E.E.; MELO, C.M.R.; MOURA, A.S.A.M.T.; ZANELLA, E.L.; BURT, D.W. and COUTINHO, L.L. Mapping QTLs on chicken chromosome 1 for performance and carcass traits in a broiler x layer cross. **Animal Genetics.** 37:95-100. 2006.

PATERSON, A.H. **Of blending, beans and bristles: the foundations of QTL mapping.** In: PATERSON, A.H. Molecular dissection of complex traits. New York: CRC Press. p.1-10. 1998.

PAYNE, F. The effect of artificial selection on bristle number in *Drosophila ampelophila* and its interpretation. **Proc. Natl. Acad. Sci. USA.** 4: 55-58. 1918.

RIJSDIJK, F.V. & SHAW, P.C. Estimation of sib-pair IBD sharing and multipoint polymorphism information content by linear regression. **Behavior Genetics**. 32(3):211-220. 2002.

ROCHA, R.B.; CRUZ, C.D.; BARROS, W.S.; FERREIRA, F.M. and ARAÚJO, E.F. Comparisons of segregating populations for genetic mapping. **Crop Breeding and Applied Biotechnology.** 4:408-415. 2004.

SAX, K. The association of size differences with seed-coat pattern and pigmentation in *Phaseolus vulgaris*. **Genetics.** 8:552-560. 1923.

SILVA, M.V.G.B.; MARTINEZ, M.L.; TORRES, R.A.; LOPES, P.S.; EUCLYDES, R.F.; MACHADO, M.A. and ARBEX, W. Mapeamento de QTL em famílias de irmãos completes por meio de modelos aleatórios. **Arq. Bras. Med. Vet. Zootec.** 56(2):232-241. 2004.

SILVA, L.C. **Simulação do tamanho da população e da saturação do genoma para mapeamento genético de RILs**. 2005. 120p. Dissertação (Mestrado em Genética e Melhoramento) Universidade Federal de Viçosa, Viçosa.

SILVA, L.C.; CRUZ, C.D.; MOREIRA, M.A. and BARROS, E.G. Simulation of population size and genome saturation level for genetic mapping of recombinant inbred lines (RILs). **Genetics and Molecular Biology**. 30(4):1101-1108. 2007.

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

VAN DER BEEK, S.; VAN ARENDOK, J.A.M. and GRÖEN, A.F. Power of two- and three-generation QTL mapping experiments in a outbreed population containing full-sib or half-sib families. **Theor. Appl. Genet.** 91:115-1124. 1995.

VAN OOIJEN, J.W. Accuracy of mapping quantitative trait lociin autogamous species. **Theory and Applied Genetics** 84:803-811. 1992.

VISSCHER, P.M.; HALEY, C.S. and THOMPSON, R. Marker assisted introgression in backcrossbreeding programs. **Genetics.** 144:1923-1932. 1996.

VISSCHER, P.M. & HOPPER, J.L. Power of regression and maximum likelihood methods to map QTL from sib pair data and DZ twin data. **Ann. Hum. Genet.** 65:583-601. 2001.

WEIR, B.S. **Genetic data analysis II.** Sinsuer Associates, Sunderland, 445p. 1996.

WELLER, J.I. **Quantitative trait loci analysis in animals.** 1st ed. New York, NY: CABI Publishing. 287p. 2001.

XU, S. & ATCHLEY, W.R. A random model approach to interval mapping of quantitative trait loci. **Genetics.** 141: 1189-1197. 1995.

XU, S. Theoretical basis of the Beavis effect. **Genetics.** 165: 2259–2268. 2003.

ZENG, Z-B. Theoretical basis of separation of multiple linked gene effects on mapping quantitative trait loci. **Proc. Natl. Acad. Sci. USA.** 90:10972-10976. 1993.