LUCIANO DA COSTA E SILVA

SIMULAÇÃO DO TAMANHO DA POPULAÇÃO E DA SATURAÇÃO DO GENOMA PARA MAPEAMENTO GENÉTICO DE RILs

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 *Magister Scientiae*.

VIÇOSA
MINAS GERAIS – BRASIL
2005

LUCIANO DA COSTA E SILVA

SIMULAÇÃO DO TAMANHO DA POPULAÇÃO E DA SATURAÇÃO DO GENOMA PARA MAPEAMENTO GENÉTICO DE RILs

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 *Magister Scientiae*.

Aprovada: 03 de fevereiro de 2005

_____	_____
Prof. Ney Sussumu Sakiyama	Dr. Luiz Antônio dos Santos Dias
_____	_____
Prof. Cosme Damião Cruz (Conselheiro)	Prof. Maurílio Alves Moreira (Conselheiro)

_____
Prof. Everaldo Gonçalves de Barros
(Orientador)

Aos meus pais Amador e Maria de Lourdes.

Aos meus irmãos Admilson, Alisson e Leidiane.

À minha sobrinha Vitória.

Aos meus amigos José Manoel, Daniel, Wagner e Lauro.

AGRADECIMENTO

Aos meus pais, Amador e Maria de Lourdes, pela educação dada aos filhos.

Aos meus irmãos, Admilson, Alisson e Leidiane pela amizade e companheirismo.

À Universidade Federal de Viçosa (UFV), especialmente ao BIOGRO, pela oportunidade e excelentes condições de trabalho.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pelo auxílio financeiro, indispensável à condução deste trabalho.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), pelo auxílio financeiro, indispensável à condução deste trabalho.

Ao Professor Everaldo Gonçalves de Barros, pela orientação, amizade e atenção em todas as vezes que precisei de sua ajuda, pessoa exemplar em cordialidade, em dedicação ao ensino e à pesquisa.

Ao Professor Conselheiro, Cosme Damião Cruz, pela orientação, amizade e atenção. Exemplo de dedicação aos estudos, na busca incansável de novos conhecimentos científicos, os quais transmite de forma brilhante aos seus estudantes e colegas.

Ao Professor Conselheiro, Maurílio Alves Moreira, pela amizade, pelos ensinamentos e pelo exemplo de dedicação à pesquisa, especialmente na busca por recursos financeiros.

Ao Dr. Luis Antônio dos Santos Dias, pela participação na banca de defesa de tese, na qual propôs importantes modificações contribuindo para a melhoria do trabalho.

Especial ao amigo Professor Ney Sussumu Sakiyama, pelos primeiros ensinamento na pesquisa e convívio, o qual teve início no segundo semestre do curso de graduação.

Ao prezado amigo Antônio Alves Pereira (Tonico), pesquisador da EPAMIG na área de melhoramento do cafeeiro.

Aos amigos colegas de trabalho: Taís, Newton, Adésio, Márcia Flor, Klever, Vilmar, Demerson, Willian, Luiz Pessoni, Tatiana, Gerardo, Fábio, Wagner, Marcio pela amisade, ajuda e agradável convívio.

Aos colegas de república, Lauro (Pé-de-Pano), Edgar (Bão) e Rogério (Batatinha), Daniel (Porqueira), Wagner (Guinho), José Manoel (Içá), Hannuar (Cabeção).

Às secretárias do curso de pós-graduação em Genética e Melhoramento, Maria da Conceição L. Vieira e Rita de Cássia R. Cruz pela dedicação e apoio.

A todos os meus mestres, desde o pré-escolar até hoje.

A todos os brasileiros que, por meio de seus impostos, permitiram a realização deste trabalho e minha educação.

BIOGRAFIA

Luciano da Costa e Silva, filho de Amador Francisco da Silva e Maria de Lourdes da Costa e Silva, nasceu em Pitangui, MG, em 20 de janeiro de 1976.

Cursou o ensino fundamental nas escolas: Escola Estadual Cel. Pedro Lino e Escola Estadual Dr. José Gonçalves, em Martinho Campos, MG.

Em dezembro de 1995, concluiu o curso de Técnico em Agropecuária na Central de Ensino e Desenvolvimento Agrário de Florestal (CEDAF), em Florestal, MG.

De dezembro de 1995 a julho de 1997 desempenhou a função de Supervisor de Granja de Matrizes Pesadas, na empresa ASA ALIMENTOS, em Brasília, DF.

Em marcos de 1998 iniciou o curso de Engenharia Agronômica na Universidade Federal de Viçosa (UFV), colando grau em março de 2003. Na UFV desenvolveu atividades de monitoria nas disciplinas de Estatística Experimental e Genética Básica, também desenvolveu atividades de pesquisa nas áreas de Melhoramento Genético do Cafeeiro, Genética Molecular Aplicada ao Melhoramento do Cafeeiro e Soja.

Em março de 2003 iniciou o curso de Mestrado em Genética e Melhoramento, na área de Genética Molecular, na UFV, defendendo tese em 3 de fevereiro de 2005.

CONTEÚDO

RESUMO

COSTA e SILVA, Luciano, M.S., Universidade Federal de Viçosa, fevereiro de 2005. **Simulação do tamanho da população e da saturação do genoma para mapeamento genético de RILs**. Orientador: Everaldo Gonçalves de Barros. Conselheiros: Maurílio Alves Moreira e Cosme Damião Cruz.

Na construção de mapas genéticos são usados vários tamanhos de população e número de marcas. Contudo, não se sabe a *priori* qual é o número mínimo de indivíduos e marcas a ser utilizado para a obtenção de mapas confiáveis. Desta forma, este trabalho, por meio da simulação de dados em computador, teve como objetivos o estudo de população RIL (*Recombinant Inbred Line*), buscando determinar: 1) a influência do número de indivíduos na população segregante sobre o mapeamento; 2) o efeito da saturação do genoma por marcas sobre o mapeamento e; 3) o número adequado de indivíduos a ser utilizado no mapeamento. Foram gerados três genomas com níveis de saturação de 5, 10 e 20 cM, com 231, 121 e 66 marcas, respectivamente. Cada genoma foi composto por 11 grupos de ligação, 100 cM cada. Para cada saturação do genoma foram geradas populações com 50, 100, 154, 200, 300, 500 e 800 indivíduos, com 100 repetições cada. Portanto, foram geradas um total de 2.100 populações. Estas populações foram mapeadas utilizando um $LOD_{mín}$ de 3 e freqüência máxima de recombinação de 30%. Dos mapas obtidos foram extraídas as informações: número de grupos de ligação e de marcas por grupo, tamanho de grupo de ligação, distância entre marcas adjacentes, variância das distâncias entre marcas adjacentes, inversão de marcas (dada pela correlação de Spearman) e grau de concordância das distâncias nos mapas com o genoma original (dada pelo estresse). População com tamanho igual ou superior a 100, 154 e 500 indivíduos, saturação de 5, 10 e 20 cM, respectivamente, levaram a formação de 11 grupos de ligação, em todas as repetições. Inversão na ordem de marcas e presença de marcas não ligadas foi maior tanto quanto menores foram os tamanhos de populações estudadas. A precisão na estimativa de distância entre marcas adjacentes foi maior tanto quanto maiores foram os tamanhos de população, independente do nível de saturação do genoma. Valores de estresse e variância reduziram significativamente com o aumento do tamanho de população. Concluindo, obtenção de mapas confiáveis a partir de população RIL deve, necessariamente, levar em conta o tamanho da população e

número de marcas, uma vez que mapas com sérias distorções foram obtidos com utilização de populações de tamanhos insuficientes, mesmo com grande quantidade de marcas. Também, mapas com sérias distorções foram obtidos com o uso de pequena quantidade de marcas, mesmo com populações com grande número de indivíduos. Tamanhos mínimos de 100, 154 e 500 indivíduos foram necessários para a obtenção de mapas com o mesmo número de marcas por grupo de ligação do genoma original, nos casos de saturação de 5, 10 e 20 cM, respectivamente.

ABSTRACT

COSTA e SILVA, Luciano, M.S., Universidade Federal de Viçosa, February, 2005. **Simulation of the population size and genome saturation level for genetic mapping of RILs**. Advisor: Everaldo Gonçalves de Barros. Committee Members: Maurilio Alves Moreira and Cosme Damião Cruz.

During the construction of genetic maps different population sizes and number of markers have been used by different authors. However, the minimum numbers of individuals and markers to be used in order to obtain reliable maps are not known. This work used data simulation aiming to determine: 1) the influence of population size; 2) the effect of genome saturation degree by markers and; 3) the adequate number of individuals to be used in the mapping of recombinant inbred line populations. Three genomes were generated with saturation levels of 5, 10 and 20 cM, each with 231, 121 and 66 markers, respectively. The genomes consisted of 11 linkage groups, each with 100 cM in length. For each saturation level, populations containing 50, 100, 154, 200, 300, 500 and 800 individuals were generated and for each population size 100 replications were analyzed. Thus, a total of 2,100 populations were generated. These populations were mapped using a LOD minimum of 3 and a maximum recombination frequency of 30%. From these maps the following variables were determined: number of linkage groups and number of markers in each group, length of the linkage group, distance between adjacent markers, variance of the distance between adjacent markers, inversion of markers (given by the Spearman correlation) and level of agreement between map distances and original genome distances (given by the stress). Population sizes equal to or greater than 100, 154 and 500 individuals in the saturation levels of 5, 10 and 20 cM, respectively, led the formation of 11 linkage groups, in all replications. As the population size increased, the inversion of markers and the presence of non-linked markers decreased, and the precision of the distance estimates between markers increased, independently of the genome saturation level. Values of stress and variance decreased significantly with the increase on the population size. In conclusion, the construction of reliable maps from RIL populations necessarily needs to consider the population size and the number of markers used, as maps with serious distortions were obtained from small sized populations, even using a large number of markers. On the other hand, maps with

serious distortions were obtained with a small number of markers, even using populations with a large number of individuals. The minimum sizes of 100, 154 and 500 individuals were necessary for obtaining maps with the same number of markers per linkage group of the original genome, in the cases of saturation level of 5, 10 and 20 cM, respectively.

1 INTRODUÇÃO

No estudo genético de determinado caráter busca-se conhecer o número de genes/alelos envolvidos no controle da sua expressão, a localização e posição relativa desses genes nos cromossomos, assim como a sua relação com outros genes. Neste contexto, os mapas genéticos são de fundamental importância, uma vez que permitem a visualização, mesmo que de forma relativa, da organização dos genes nos cromossomos.

No melhoramento de plantas e animais, os mapas genéticos são de grande importância, pois possibilitam a cobertura completa e análise de genomas, a decomposição de caracteres complexos em seus componentes mendelianos simples, a localização de regiões do genoma responsáveis pelo controle da expressão de caracteres importantes, sejam eles qualitativos ou quantitativos (QTLs – *Quantitative Trait Loci*).

Com o desenvolvimento da biologia molecular surgiram ferramentas novas que possibilitam a construção de mapas genéticos mais fidedignos. Uma destas ferramentas são os marcadores moleculares do DNA, que têm permitido a construção de mapas genéticos para várias espécies vegetais e animais. Tais mapas podem atingir um alto grau de saturação devido à disponibilidade de grande número de marcas genéticas, que têm a vantagem de não ser influenciadas pelo ambiente, além de serem altamente polimórficas.

Para a construção de mapas genéticos é necessário a integração de técnicas que vão desde a obtenção e caracterização dos genitores, o desenvolvimento de populações segregantes, a mensuração de características quantitativas em delineamentos experimentais adequados, a identificação dos genótipos nos locos marcadores por meio de técnicas de biologia molecular e a utilização de diversas técnicas de análise estatística e computacional.

No mapeamento genético podem ser usadas, dependendo da espécie de interesse, vários tipos de populações segregantes, tais como, populações F_2,

progênies provenientes de retrocruzamentos, progênies F_1 (*pseudo-testcross*), duplo-haplóides e linhas endogâmicas recombinantes (RILs – *Recombinant Inbred Lines*), sendo que cada uma delas tem as suas vantagens e desvantagens.

A disponibilidade de mapas genéticos fidedignos depende de uma série de fatores, tais como tipo e número de marcadores moleculares utilizados, tipo e tamanho de população analisada. Além disso, aspectos de natureza metodológica devem ser considerados, tais como, padrão de segregação de locos individuais, análise de segregação conjunta, níveis de freqüência de recombinação e níveis máximos de LOD *score* (*Likelihood of odds*). O LOD *score* é o logaritmo da razão entre a probabilidade de dois locos estarem ligados e a probabilidade de não estarem ligados. O entendimento destas questões é fundamental, pois elas determinam a confiabilidade do mapa e sua aplicação nos programas de melhoramento.

Na construção de mapas genéticos e mapeamento de QTLs são usados modelos estatísticos para descrever sistemas genéticos e biológicos reais. Contudo, estes sistemas são complexos, impossibilitando a inclusão de todas as variáveis nos modelos utilizados. Portanto, é razoável que existam diversos modelos estatísticos para estudos de populações segregantes e mapeamento de QTLs. Alguns destes são muito complexos, enquanto outros são bastante simples. O conhecimento das propriedades dos estimadores destes modelos é de fundamental importância. Elas podem ser obtidas parametricamente, se a distribuição do estimador é conhecida e bem caracterizada. Contudo, na maioria dos modelos de mapeamento de QTLs, é difícil a obtenção das propriedades do estimador parametricamente. Portanto, a simulação em computadores tem sido utilizada para a obtenção de propriedades e verificação do desempenho de modelos, métodos, tamanhos e tipos de população. Não há maneira de examinar o desempenho de um modelo usando dados reais (experimentais), devido ao desconhecimento dos verdadeiros parâmetros. Na simulação de dados via computador, no entanto, os “verdadeiros” parâmetros são conhecidos e podem ser usados na comparação da eficiência dos diversos modelos.

Este trabalho, feito por meio de simulação de dados em computador, teve como objetivos, determinar, em populações RIL (*Recombinant Inbred Line*): 1) a influência do número de indivíduos da população segregante sobre o mapeamento; 2) o efeito da saturação do genoma por marcadores moleculares no mapeamento e; 3) o número adequado de indivíduos a ser utilizado no processo de mapeamento genético.

2 REVISÃO DE LITERATURA

2.1 Mapeamento genético

2.1.1 Marcadores genéticos

O marcador genético corresponde a uma característica do organismo que pode ser facilmente detectada a olho nu ou com a ajuda de algum aparato tecnológico, e que co-segrega com genes de interesse. Uma característica para ser útil como marcador deve evidenciar diferenças entre os indivíduos analisados e, além disso, ser reproduzida com precisão na prole.

Até meados da década de 1960, em estudos de genética e melhoramento, eram utilizados basicamente marcadores morfológicos, em geral fenótipos de fácil identificação visual, como cor do hipocótilo, de pétalas, de brotações, de sementes, morfologia foliar, entre outros. Estes marcadores contribuíram bastante com o desenvolvimento teórico da análise de ligação gênica e na construção das primeiras versões de mapas genéticos. Contudo, devido ao reduzido número de marcadores morfológicos polimórficos, a probabilidade de encontrar associações entre esses marcadores e características importantes era reduzida (Ferreira e Grattapaglia, 1995).

Este quadro começou a mudar com a utilização dos primeiros marcadores moleculares, as aloenzimas, que são variantes de uma mesma proteína detectados por eletroforese. Este tipo de marcador tem a vantagem de ser relativamente barato em estudos que necessitam de um grande número de indivíduos, porém, apresenta

um baixo número de variantes nas proteínas, portanto, limitando o desenvolvimento de mapas genéticos altamente saturados.

Quando os métodos de avaliação de variações do DNA tornaram-se amplamente disponíveis, durante a metade dos anos 80, os marcadores baseados no DNA substituíram largamente as aloenzimas em estudos de mapeamento. O DNA é o material genético dos organismos e, portanto, diferenças genéticas entre indivíduos são reflexo direto de diferenças nas seqüências de nucleotídeos do DNA. Uma ampla variedade de técnicas pode ser utilizada para permitir a detecção de variações no DNA. O método mais avançado é o seqüenciamento, porém, existem métodos mais simples e mais baratos, suficientes para a maioria dos propósitos. Estes métodos incluem a análise de polimorfismo de comprimento de fragmentos de restrição de DNA (RFLP - *Restriction Fragment Length Polymorphism*), os baseados na reação de polimerização em cadeia (PCR – *Polymerase Chain Reaction*) (Mullis e Fallona, 1987), dentre estes: o polimorfismo de DNA amplificado ao acaso (RAPD - *Randomly Amplified Polymorphic DNA*) (Williams *et al.*, 1990), microssatélites (SSR – *Simple Sequence Repeats*) (Litt e Luty, 1989), polimorfismo de comprimento de fragmentos amplificados (AFLP – *Amplified Fragment Length Polymorphism*) (Zabeau, 1993), sítios marcados por seqüência (STS - *Sequence Tagged Site* ou EST - *Expressed Sequence Tags*) (Olson *et al.*, 1989) e regiões amplificadas caracterizadas por seqüência (SCAR - *Sequence Characterized Amplified Regions*) (Paran e Michelmore, 1993). Recentemente, têm sido utilizado também os marcadores baseados em polimorfismo de base única (SNPs - *Single Nucleotide Polymorphism*).

Os marcadores RFLP foram muito utilizados durante um certo tempo, para detectar polimorfismos na molécula de DNA. O método consiste na digestão do DNA com uma variedade de enzimas de restrição e análise do DNA digerido por eletroforese. Após a separação dos fragmentos, bandas individuais são detectadas pela utilização de sondas de DNA marcadas, que tem bases complementares à determinada região do genoma, geralmente, as sondas correspondem a locos únicos no genoma. Marcadores RFLP possuem expressão co-dominante, assim, possibilitando a identificação de genótipos homozigotos e heterozigotos. Este tipo de marcador tem sido utilizado na construção de mapas genéticos de diversas espécies, incluindo plantas de interesse agronômico como arroz (He *et al.*, 2001; Chen *et al.*, 2003; Uga *et al.*, 2003), milho (Séne *et al.*, 2000; Groh *et al.*, 1998; Cardinal *et al.* 2001), soja (Njiti *et al.*, 2002), feijão (Miklas *et al.*, 2001) e trigo (Nachit *et al.*, 2001).

A PCR é uma técnica que utiliza a enzima DNA polimerase para a síntese de fragmentos de DNA *in vitro*. A reação de polimerização do DNA baseia-se no pareamento de um par de oligonucleotídeos, pequenas moléculas de DNA fita simples, utilizados como iniciadores ou *primers* e que delimitam a seqüência de DNA de fita dupla, alvo da amplificação. Esta técnica é utilizada na obtenção de marcadores RAPD, SSR, AFLP, SCAR e STS. No caso da técnica de RAPD, fragmentos de DNA são amplificados no genoma com a utilização de *primers* curtos, aproximadamente 10 nucleotídeos, de seqüência arbitrária. Os marcadores RAPD têm a vantagem de requerer uma pequena quantidade de DNA para análise, além de um simples *primer* poder revelar muitos locos de uma só vez, cada loco correspondendo a uma região diferente do genoma analisado. Contudo, estes marcadores são dominantes, portanto, não sendo possível a distinção do genótipo homozigoto dominante do heterozigoto. Vários pesquisadores têm utilizado os marcadores RAPD no mapeamento genético de vários organismos, como feijão (Faleiro, 2000; Miklas *et al.*, 2001), café (Teixeira, 2001), arroz (Uga *et al.*, 2003; Cai e Morishima, 2002) e soja (Hnetkovsky *et al.*, 1996; Chang *et al.*, 1997; Njiti *et al.*, 2002; Ferreira *et al.*, 2000).

Microssatélites consistem de pequenas seqüências com 1 a 4 nucleotídeos de comprimento, repetidas em tandem no genoma. Os marcadores microssatélites são obtidos pela amplificação destas seqüências de nucleotídeos repetidas por meio da PCR, utilizando para isto um par de *primers* específicos contendo de 20 a 30 bases de comprimento, complementares a seqüências únicas que flanqueiam o microssatélite. Devido à expressão codominante destes marcadores e ao seu multialelismo, eles podem ser aplicados a todos os tipos de populações segregantes empregadas no mapeamento genético e estudos de ligação (Ferreira e Grattapaglia, 1995). Trabalhos, como os de He *et al.* (2001), Xing *et al.* (2002), Price *et al.* (2002) e Chen *et al.* (2003) em arroz, Chang *et al.* (1997), Meksem *et al.* (2001), Iqbal *et al.* (2001), Yuan *et al.* (2002) e Njiti *et al.* (2002) em soja, Nachit *et al.* (2001) em trigo e Cardinal *et al.* (2001) em milho, empregaram marcadores microssatélites.

Os AFLPs são outro tipo de marcador amplamente utilizado na construção de mapas genéticos. Desde o seu desenvolvimento e divulgação, esta técnica tem sido utilizada para a obtenção de um grande número de marcadores distribuídos em genomas de procariotos e eucariotos. O aspecto relativo à obtenção de grande

número de polimorfismo tem levado a sua utilização na obtenção de mapas altamente saturados (Ballvora *et al.*, 1995, Meksem *et al.*, 1995).

Os marcadores AFLP e RAPD podem ser clonados e convertidos em marcadores específicos denominados, respectivamente, STS e SCAR. Estes marcadores, geralmente, são desenvolvidos a partir de seqüências de genes ou de marcadores associados a características de interesse, como os marcadores SCARF10 e SCARFBA8 ligados a genes de resistência à ferrugem do feijoeiro (Faleiro, 2000).

Os marcadores moleculares apresentam algumas vantagens em relação aos marcadores morfológicos, como: o nível de polimorfismo é geralmente alto para cada loco estudado, facilitando o desenvolvimento de mapas genéticos a partir de populações segregantes de cruzamentos específicos; a neutralidade em relação aos efeitos de ambiente, com pouco ou nenhum efeito de epistasia ou pleiotropismo; são, em geral, codominantes, o que permite a obtenção de maiores informações dos locos estudados (Ferreira e Grattapaglia, 1995).

2.1.2 Populações segregantes para mapeamento genético

Para o mapeamento genético, diferentes tipos de populações segregantes podem ser empregadas. Tradicionalmente, são utilizadas populações derivadas do cruzamento de linhas puras, o que origina uma geração F_1 a qual é autofecundada ou retrocruzada com um dos pais para a produção de uma geração F_2 ou RC_1, respectivamente (Ferreira e Grattapaglia, 1995). Alternativamente, são utilizadas populações F_n (n= 3, 4,...,∞), duplo-haplóides (DH) e linhagens endogâmicas recombinantes (RILs) (Burr *et al.*, 1988). A escolha da população deve levar em conta os objetivos do pesquisador, além do tempo e recursos disponíveis para a execução do trabalho. Nas plantas F_1 o desequilíbrio de ligação é máximo, e os estudos das populações derivadas a partir destas plantas F_1 procuram explorar este desequilíbrio (Schuster e Cruz, 2004).

Nas populações F_2 derivadas de cruzamentos de genitores homozigotos, os marcadores codominantes segregam na proporção 1:2:1 e os dominantes, na proporção 3:1. Como vantagens apresenta maior facilidade e rapidez para a sua obtenção e maior precisão no mapeamento de QTLs, devido à disponibilidade de informações dos três genótipos (AA, Aa e aa). Como desvantagens, há perda de

precisão na mensuração de características quantitativas e a impossibilidade do mapeamento de genes de resistência a doenças em que seja necessária a inoculação da população segregante com diferentes raças fisiológicas ou diferentes patógenos, uma vez que não é possível a replicação dos indivíduos da população. Estas desvantagens podem ser contornadas por meio da propagação vegetativa ou com a utilização de populações $F_{2:3}$. Neste caso, os genótipos das plantas são determinados na geração F_2 e as características fenotípicas são medidas em plantas F_3, com repetições. Os dados das médias das famílias F_3 são geralmente utilizados como o valor fenotípico das plantas F_2 que originaram estas famílias. Porém, isto só é verdadeiro na ausência de dominância. Para a utilização de dados de famílias F_3 na análise de plantas F_2 é necessário levar em consideração a composição genotípica das plantas na geração F_3 (Schuster e Cruz, 2004). População F_2 foi utilizada para o mapeamento genético de espécies, como o girassol (Yu *et al.*, 2003), a soja (Tasma e Shoemaker, 2003) e a lentilha (Rubeena *et al.*, 2003).

Nas populações derivadas por retrocruzamento, apenas dois genótipos são obtidos, heterozigotos (Aa) e homozigotos (aa) ou (AA). Apresentam como vantagens a rapidez na obtenção da população e o fato de que a segregação obtida representa a composição dos grãos de pólen do progenitor doador. Apresentam como desvantagens: o grande número de cruzamentos necessários para obter a população, o que pode inviabilizar sua utilização em espécies em que a hibridação é difícil ou cujo ciclo de vida é demasiadamente longo; não é possível estimar componentes genéticos associados aos desvios de dominância, uma vez que apenas dois genótipos são obtidos; quando são utilizados marcadores dominantes e o genitor recorrente é dominante para um dado loco, há perda de informação, pois não é possível a distinção dos dois genótipos (AA e Aa) na população segregante; não é possível obter dados com repetição, o que diminui a precisão das estimativas de parâmetros no mapeamento de QTLs. Neste tipo de população também é possível a utilização das mesmas alternativas aplicadas nas populações F_2 para contornar o problema da replicação. Na população de retrocruzamento há segregação apenas para alelos derivados do progenitor não-recorrente, de modo que haverá locos em heterozigose e locos idênticos aos do progenitor recorrente, o que leva a maior número de associações entre alelos marcadores do progenitor recorrente. Como exemplo de mapa construído com este tipo de população pode ser citado o de Ky *et al.* (2000) em café.

Nos programas de melhoramento genético de plantas é comum o uso de populações derivadas por autofecundação a partir da geração F_2, tais como $F_{2:3}$, $F_{2:4}$, etc. Se um mesmo número de plantas F_3 é obtido de cada planta F_2, e um mesmo número de plantas F_4 é obtido de cada planta F_3, e assim por diante, estas populações terão estrutura genética previsível e, portanto, podem ser utilizadas no desenvolvimento de mapas genéticos. Estas populações são genericamente tratadas como populações F_n. Tar'an (2002) utilizou uma população $F_{2:4}$ para a construção de um mapa genético e análise de QTLs em feijoeiro comum. A vantagem da sua utilização é que não há a necessidade de produzir uma população especialmente para o mapeamento genético, uma vez que estas populações geralmente estão prontas e disponíveis para uso nos programas de melhoramento. Porém, este tipo de população tem aplicação mais apropriada na identificação de marcadores moleculares ligados a genes ou QTLs de interesse específico, do que para a construção de um mapa completo de ligação de uma espécie, além disto, apresentam as mesmas limitações das populações F_2 (Schuster e Cruz, 2004).

As populações de RILs são derivadas por sucessivas autofecundações a partir de uma população F_2 pelo método da descendência de uma única semente (SSD) ou por cruzamentos entre irmãos, até atingirem um elevado grau de homozigose, de maneira que ao final do processo todos os locos gênicos terão uma segregação de 1:1 (AA:aa). Na formação das RILs por meio de autofecundações, cada planta F_2 gera uma planta F_3, que por sua vez gera uma planta F_4 e assim por diante. Quando a população atinge a geração F_6 ou F_7, abrem-se linhas, que são as RILs. Nelas, cada planta F_2 é representada por uma linha endogâmica homozigota. Assim, toda a variabilidade existente na população F_2 original estará representada pelas RILs, desde que um tamanho de população adequado seja utilizado. As desvantagens da utilização de RILs são o tempo requerido para a obtenção da população, de 6 a 8 gerações de autofecundações e, a impossibilidade da estimação de efeitos devido à dominância, uma vez que apresentam apenas dois genótipos na população segregante. Entretanto, por serem compostas apenas por indivíduos homozigotos, as RILs podem ser perpetuadas. A disponibilidade de grande quantidade de sementes de cada RIL permite o cultivo em vários locais, possibilitando a obtenção de estimativas mais precisas dos parâmetros dos caracteres quantitativos a serem mapeados. Além disso, é possível mapear locos associados à interação genótipo x ambiente destas características. A necessidade de muitos ciclos de meiose para atingir a homozigose

resulta em oportunidades adicionais para a recombinação. A freqüência de recombinação nas RILs (r') aproxima-se do valor r'=2r/(1+2r) para linhas autofecundadas e r'=4r/(1+6r) para linhas formadas a partir de cruzamentos entre irmãos, sendo r a freqüência de recombinação na geração F_2, que expressa a distância entre os locos (Burr e Burr, 1991). Contudo, cada RIL tem dois cromossomos homólogos idênticos, enquanto que indivíduos F_2 têm dois homólogos distintos, os quais contribuem com diferentes informações.

As populações de duplo-haplóides são obtidas pela duplicação do número de cromossomos de grãos de pólen da geração F_1. Conseqüentemente, todos os indivíduos na população segregante são homozigotos e representam a variabilidade genética encontrada no indivíduo parental que produziu os grãos de pólen. Assim como na população de RIL, este tipo de população apresenta razão de segregação de 1:1 (AA:aa), pode ser replicada e não permite a estimação de componentes devido à dominância. Porém, o tempo necessário para a produção da população é bem menor.

Além das considerações feitas anteriormente a respeito de cada tipo de população, outra informação muito importante, sob o ponto de vista da eficácia do mapeamento, é a variância das estimativas da porcentagem de recombinação obtidas nos diferentes tipos de populações. No Quadro 1 estão apresentadas, para os diversos tipos de populações e marcadores, as estimativas de variância da porcentagem de recombinação quando da utilização de população segregante de tamanho igual a 200 indivíduos e estimativas de freqüência de recombinação (r) iguais a 0,05, 0,10 e 0,20. A população de F_2 com marcadores co-dominantes é a que proporciona menores estimativas de variância para as três freqüências de recombinação consideradas, portanto, é a melhor população para o propósito de mapeamento genético. Considerando apenas as estimativas de recombinação de 0,05 e 0,10, a segunda melhor população para mapeamento é RIL. Porém, para estimativas de freqüência de recombinação de 0,20, esta população não é a segunda melhor, pois apresenta estimativas de variância superior às obtidas nas populações de retrocruzamento, F_2 com marcador co-dominante, duplo-haplóide e F_2 com marcadores co-dominantes e dominantes. Tal fato já permite concluir que populações de RILs devem ser estruturadas para o mapeamento com boa saturação, caso contrário seu uso será ineficaz.

Quadro 1 – Valores de variâncias[1] das estimativas da porcentagem de recombinação, para vários tipos de populações com tamanho igual a 200 indivíduos

População e marcador	Freqüência de recombinação (r)		
	0,05	0,10	0,20
Retrocruzamento	2,37	4,50	8,00
F_2 com marcador co-dominante	1,25	2,53	5,23
F_2 com marcador dominante em fase de acoplamento	2,52	5,09	10,42
F_2 marcador dominante em fase de repulsão	49,69	48,77	45,33
F_2 com marcadores co-dominante e dominante	2,47	4,91	9,69
Duplo-haplóide	2,37	4,50	8,00
RILs derivadas de autofecundações	1,51	3,60	9,80

[1] Valores multiplicados por 10^4.

Fonte: Schuster e Cruz (2004)

2.1.3 Análise de segregação de marcas

Na construção de um mapa genético, o primeiro procedimento a ser realizado é o estudo individual dos locos, os quais devem apresentar razão de segregação de acordo com o tipo de população estudada. Nas populações de retrocruzamento, duplo-haplóides e RILs espera-se uma razão de segregação de 1:1 e, nas populações F_2 uma razão de segregação de 1:2:1 ou 3:1, para locos codominantes ou dominantes, respectivamente. Para verificar se um loco está segregando na razão esperada, procede-se a comparação do número de indivíduos observados em cada classe com o esperado de acordo com a razão de segregação. Um teste que tem se mostrado bastante eficiente e útil para este fim, é o teste de qui-quadrado (χ^2), pois, além de levar em consideração os desvios ocorridos entres os valores esperados e observados, também é sensível ao tamanho da amostra.

A segregação dos locos gênicos pode não ocorrer de acordo com o esperado, isto tem sido relatado freqüentemente em cruzamentos interespecíficos, independentemente do tipo de marcador utilizado. A presença de locos apresentando distorção na razão de segregação é altamente variável, dependendo da espécie e tipo

de marcador utilizado. No mapeamento de *Cryptomeria japonica* (Nikaido *et al*., 1999), *Prumus* (Foolad *et al*., 1995), *Helianthus* (Quillet *et al*., 1995), *Oryza* (Xu *et al*., 1997), *Lens sp*. (Eujayl *et al*. 1998) e *Hevea spp*. (Lespinasse *et al*., 2000), as percentagens de locos com razão de segregação distorcida foram de 41%, 40,7%, 23%, 14,5%, 8,4%, e 1,4%, respectivamente. Ky *et al.* (2000) encontraram, em *Coffea* sp., locos com razão de segregação distorcida na freqüência de 30% e Faleiro (2000) encontrou uma freqüência de 14,3% de locos com distorção na razão de segregação em *Phaseolus vulgaris* L.

A distorção na razão de segregação esperada pode ser atribuída a diversas causas, tais como: i) processos de seleção no estádio de gameta ou zigoto (Zamir e Tadmor, 1986); ii) locos gênicos que apresentam seleção natural, locos próximos a genes que levam a menor viabilidade de gametas, como observado em arroz por He *et al*. (2001) e; iii) conversão gênica. Quando uma célula diplóide sofre meiose são produzidas quatro células haplóides, exatamente metade dos alelos nestas células deveriam ser de origem materna (alelos que a célula diplóide recebeu de sua mãe) e a outra metade paterna (alelos que a célula diplóide recebeu de seu pai). Porém, estudos em alguns organismos, por exemplo, fungos, têm revelado raros casos nos quais as regras padrões da genética tem sido violadas. Ocasionalmente, as meioses produzem três copias de alelos maternais e somente uma cópia do alelo paternal. Este fenômeno é conhecido com conversão gênica. A conversão gênica sempre ocorre em associação com eventos de recombinação genética de homólogos, e acredita-se ser uma conseqüência direta de mecanismos gerais de recombinação e reparo do DNA (Alberts *et al*., 2002)

Ky *et al*. (2000) obtiveram, em população de retrocruzamento derivada do cruzamento interespecífico de *Coffea sp*., razão de segregação de 3:1 e 1:3, em favor de um parental ou outro. Os autores relatam que as distorções são, na maioria das vezes, devidas à seleção de gametas, zigoto ou pós-zigoto, mas que estes motivos não são suficientes para explicar as os resultados de 3:1 e 1:3. Eles relatam que a conversão gênica pode ser a causadora de tais distorções, conforme já observado por outros autores, além disto, sugerem que a melhor explicação para os seus resultados seria a ocorrência de alta taxa de conversão gênica e baixa taxa de segregação pós-meiótica, reflexo do alto nível de formação de heteroduplex e eficiente reparo dos pareamentos errados. A alta taxa de formação de heteroduplex e pareamento errôneo do DNA são esperados durante o pareamento de cromossomos homólogos, em

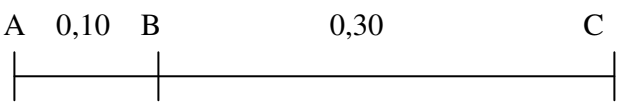
híbridos interespecíficos, sendo que a conversão gênica parece ser característica de cruzamentos interespecíficos, não dependendo do tipo de marcador.

2.1.4 Ligação fatorial e recombinação

Ligação fatorial é definida como a associação entre genes localizados num mesmo cromossomo, sendo que para estes genes as propriedades genotípicas da 2ª Lei de Mendel (Lei da Segregação Independente) não se aplicam, salvo quando eles estão separados por uma grande distância. Portanto, genes ligados tendem a ser herdados em conjunto. O fenômeno foi descoberto em 1906 por Bateson e Punnett (Griffiths et *al*., 1996), que verificaram a falta de independência de dois genes em ervilhas. Quando os genes estão muito próximos entre si, diz-se que ocorre ligação (*linkage*) completa e quando eles estão no mesmo grupo de ligação mas suficientemente separados, diz-se que a ligação é parcial.

Sturtevant, no início do século XX, estudando a mosca das frutas (*Drosophila melanogaster*), na tentativa de elucidar o processo de ligação fatorial, propôs como base para a quantificação da distância de dois genes num mesmo cromossomo, a freqüência de recombinação entre estes genes. Quanto maior for a distância entre dois genes, maior será a probabilidade de ocorrer um quiasma naquela região e, por conseqüência, maior será a probabilidade de recombinação. A partir desta descoberta, ficou estabelecida uma maneira de determinar a distância entre dois genes (percentagem de recombinantes) e, que até hoje é o princípio utilizado na construção dos mais diversos tipos de mapas de ligação.

Entretanto, quando se analisam 3 locos gênicos deve-se considerar, na estimativa da distância entre os locos extremos, a ocorrência do fenômeno da interferência. Considerando-se o fragmento cromossômico abaixo, com três genes (A, B e C), cujas freqüências de recombinação, ou seja, as distâncias entre eles sejam r_{AB}= 0,10 e r_{BC} = 0,30:



a probabilidade de recombinação entre A e C, admitindo ausência de interferência, portanto, coincidência igual à unidade (c = 1), isto é, que os *crossing overs* (permuta) sejam independentes, será:

$$r_{AC} = r_{AB} + r_{BC} - 2cr_{AB}\, r_{BC} \qquad \text{Eq.(1)}$$

Substituindo os valores de r_{AB}, r_{BC} e c, tem-se que: r_{AC}= 0,10 + 0,30 – 2(1)(0,10)(0,30) = 0,34, valor este diferente da soma 0,10 + 0,30 = 0,40, demonstrando que a escala de freqüência de recombinação, não é aditiva, exceção, quando a interferência é total, ou seja, coincidência igual a zero (c = 0).

Portanto, a utilização da freqüência de recombinação para determinar a distância entre genes apresenta o problema da falta de aditividade, o qual posteriormente foi contornado com o desenvolvimento das funções de mapeamento de Haldane (1919) e de Kosambi (1944).

2.1.5 Funções de mapeamento

As funções de mapeamento são utilizadas para a conversão da distância entre genes, dada pela freqüência de recombinação, para uma unidade de distância que apresente a propriedade de aditividade. As funções de mapeamento propostas por Haldane e Kosambi são as mais utilizadas.

Para obter a função de mapeamento de Haldane, são admitidas as seguintes pressuposições:

a) As ocorrências de permutas gênicas são independentes, isto é, não ocorre interferência;

b) As permutas gênicas ocorrem ao acaso ao longo do cromossomo.

Assim, o número ímpar de *crossing overs* (k) num intervalo definido por dois locos gênicos pode ser modelado segundo a distribuição de Poisson com média θ, isto é:

$$P(r) = r = \sum_{k} \frac{\theta^k e^{-\theta}}{k!} = e^{-\theta}\left(\frac{\theta}{1!} + \frac{\theta^3}{3!} + \ldots\right) = \frac{e^{-\theta}\left(e^{\theta} - e^{-\theta}\right)}{2} = \frac{1}{2}\left(1 - e^{-2\theta}\right) \qquad \text{Eq.(2)}$$

onde, P(r) é a probabilidade de ocorrência de recombinação e θ é o número de unidades de mapa, em Morgans, entre dois locos gênicos. Após a resolução da equação acima para θ têm-se a função de mapeamento de Haldane:

$$\theta_H = -\frac{1}{2}\ln(1-2r) \qquad \text{Eq.(3)}$$

onde,

θ_H é a distância entre dois locos gênicos expressa em Haldane (cM);

r é a freqüência de recombinação entre dois locos gênicos ($0 \leq r \leq 0{,}5$).

Se r for igual a 0, o valor de θ também será 0, isto é, os genes estão completamente ligados. Se r for igual a 0,5, o valor de θ tende ao infinito (∞) e, isto significa que os genes não estão ligados. A causa disto poderia ser pelo fato dos genes estarem localizados em cromossomos diferentes ou estarem num mesmo cromossomo, mas separados por uma grande distância.

Liu (1998) discute a presença de desvios dos valores obtidos pela função de Haldane em relação à verdadeira distância entre dois genes ($r_{AC} = r_{AB} + r_{BC} - 2cr_{AB}r_{BC}$). Estes desvios são atribuídos ao fenômeno da interferência, que é caracterizada pelo fato da ocorrência de um *crossing over* numa região geralmente diminuir a probabilidade de ocorrência de *crossing over* numa região adjacente. Na maioria dos casos, o valor da interferência varia de zero a um. Se os *crossing overs* forem independentes, o valor de interferência será zero. Quando a ocorrência de *crossing over* numa dada região impedir a ocorrência de outro numa região adjacente, a interferência será completa, e o seu valor será um.

Das várias funções de mapeamento que consideram o fenômeno da interferência, a mais utilizada no desenvolvimento de mapas genéticos é a de Kosambi, a qual considera a coincidência igual a 2r, portanto interferência igual a 1 – 2r, e que é dada por:

$$\theta_K = \frac{1}{4}\ln\left(\frac{1+2r}{1-2r}\right) \qquad \text{Eq.(4)}$$

onde,

θ_K é a distância entre dois locos gênicos expressa em Kosambi (cM);

r é a freqüência de recombinação entre dois locos gênicos ($0 \leq r \leq 0{,}5$).

Na Figura 1 está representada a relação entre a freqüência de recombinação (r) e a distância de mapa em centiMorgans, expressa pelas funções de mapeamento de Haldane e Kosambi. Para pequenos valores de freqüência de recombinação (inferior a 20%), ambas as funções de mapeamento têm valores similares à freqüência de recombinação. Contudo, à medida em que a distância entre dois dados locos aumenta, ou seja, com uma maior freqüência de recombinação, mais discrepantes tornam-se os valores obtidos pela função de Kosambi e Haldane, entre si e, em relação à freqüência de recombinação. Com a função de Haldane sempre apresentando os maiores valores (adaptado de Wang, 2000).

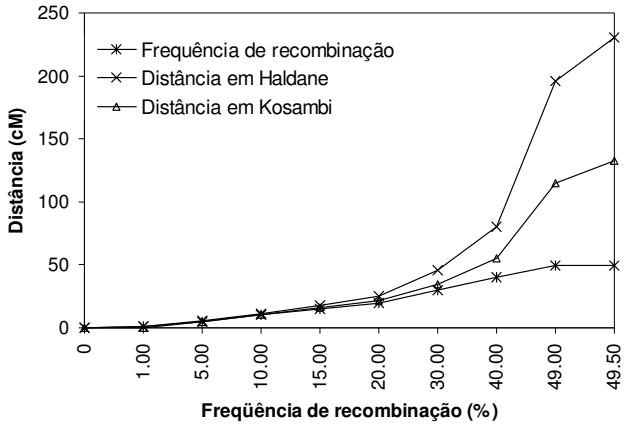


Figura 1 – Relação entre freqüência de recombinação (%) e distância de mapa (centiMorgan) expressa por Haldane e Kosambi.

2.1.6 Estimação de freqüência de recombinação

No mapeamento é necessário estimar a probabilidade de recombinação entre cada par de marcas. No caso de populações RILs, o processo do seu desenvolvimento envolve muitas gerações de recombinação, sendo que os genótipos homozigotos são normalmente fixados após uma série de autofecundações. Assim, iniciando com um indivíduo F_1 e considerando apenas dois genes (AaBb), apenas quatro genótipos são possíveis nas RILs resultantes (AABB, AAbb, aaBB e aabb) e, a

freqüência de recombinação aproxima-se do valor r'=2r/(1+2r) para linhas autofecundadas e r'=4r/(1+6r) para linhas formadas a partir de cruzamentos entre irmãos, sendo r' a freqüência de recombinação observada nas RILs, e r a freqüência de recombinação na geração F_2, que expressa a distância entre os marcadores (Quadro 2) (Burr e Burr, 1991; Schuster e Cruz, 2004).

Quadro 2 - Freqüências genotípicas esperadas numa população de RILs obtida por dois processos, linhas autofecundadas e cruzamento entre irmãos[1]

Genótipos (classes)	Observado (n_i)	Freqüências ou probabilidades (p_i)		
		Linhas autofecundadas	Cruzamento entre irmãos	Geral [2]
AABB	n_1	$\frac{1}{2(1+2r)}$	$\frac{1+2r}{2(1+6r)}$	$\frac{1-r'}{2}$
aabb	n_2	$\frac{1}{2(1+2r)}$	$\frac{1+2r}{2(1+6r)}$	$\frac{1-r'}{2}$
AAbb	n_3	$\frac{r}{1+2r}$	$\frac{2r}{1+6r}$	$\frac{r'}{2}$
aaBB	n_4	$\frac{r}{1+2r}$	$\frac{2r}{1+6r}$	$\frac{r'}{2}$

[1] Adaptado de Schuster e Cruz (2004).

[2] r'=2r/(1+2r) para linhas autofecundadas e r'=4r/(1+6r) para linhas formadas por cruzamentos entre irmãos.

Nos trabalhos de mapeamento tem-se utilizado o método da máxima verossimilhança para estimação das freqüências de recombinação, conforme descrito a seguir, de acordo com Schuster e Cruz (2004). Para a utilização do método da máxima verossimilhança é necessário o conhecimento da função densidade de probabilidade (fdp) da distribuição de uma variável aleatória x. Para populações de RILs, esta função segue uma distribuição multinomial. Portanto, a função de máxima verossimilhança para uma população de tamanho N, com n possíveis classes, assumindo distribuição multinomial é dada por:

$$L(p_i \mid n_i) = \lambda p_1^{n1} p_2^{n2} p_3^{n3} ... p_n^{nn} \qquad \text{Eq.(5)}$$

sendo

$$\lambda = \frac{N!}{n_1! n_2! \ldots n_n!}$$

$$N = \sum_{i=1}^{n} n_i$$

onde

n_i é o número de ocorrência do evento x_i com probabilidade p_i.

A partir da Eq. (5) é obtida a função suporte, aplicando-se o logaritmo natural (ln) em ambos os termos da expressão anterior, como dado a seguir:

$$\ell(p_i \mid n_i) = \ln(\lambda p_1^{n1} p_2^{n2} \cdots p_n^{nn}) \qquad \text{Eq.(6)}$$

ou ainda

$$\ell(p_i \mid n_i) = \ln\lambda + n_1 \ln p_1 + n_2 \ln p_2 + \cdots + n_n \ln p_n \qquad \text{Eq.(7)}$$

O último passo do processo de estimação pelo método da máxima verossimilhança (EMV) é a obtenção da função escore (f') , que é dada pela primeira derivada da função suporte [Eq. (7)] em relação ao parâmetro a que se deseja estimar, como dado a seguir:

$$f'\left[\ell(p_i \mid n_i)\right] = \frac{\partial \ell(p \mid n)}{\partial p} \qquad \text{Eq.(8)}$$

Como neste caso o único parâmetro a ser estimado é a freqüência de recombinação entre duas marcas, o estimador de máxima verossimilhança é obtido igualando-se a Eq. (8) a zero, como segue:

$$f'\left[\ell(p_i \mid n_i)\right] = \frac{\partial \ell(p \mid n)}{\partial p} = 0 \qquad \text{Eq.(9)}$$

Tendo demonstrado o processo de obtenção dos estimadores de máxima verossimilhança, a estimação *per se* fica dependente apenas do conhecimento dos valores de n_i (número de ocorrência de cada classe genotípica) e da probabilidade p_i associada a cada uma dessas classes. A obtenção destes valores está apresentada no Quadro 2.

Uma vez obtidos os valores de n_i e as probabilidades p_i associadas, basta realizar as devidas substituições na equação de máxima verossimilhança [Eq. (5)], de maneira que se obtém:

$$L(r'|n_i) = \lambda\left[\frac{1}{2}(1-r')\right]^{n1}\left[\frac{1}{2}(1-r')\right]^{n2}\left(\frac{1}{2}r'\right)^{n3}\left(\frac{1}{2}r'\right)^{n4} \qquad \text{Eq.(10)}$$

em que:

$$\lambda = \frac{N!}{n_1!n_2!n_3!n_4!}$$

$$N = n_1 + n_2 + n_3 + n_4$$

Aplicando-se o logaritmo natural em ambos os termos de Eq. (10), obtém-se a função suporte, como a seguir:

$$\ell(r'|n_i) = \ln(\lambda) + N\ln(1/2) + (n_1 + n_2)\ln(1-r') + (n_3 + n_4)\ln(r') \qquad \text{Eq.(11)}$$

Assim, derivando-se a função suporte em relação ao parâmetro r' obtém-se a função escore, como dado:

$$\frac{\partial\ell(r'|n_i)}{\partial r'} = \frac{(n_1 + n_2)}{1-r'}(-1) + \frac{(n_3 + n_4)}{r'} = f'(r') \qquad \text{Eq.(12)}$$

ou ainda

$$f'(r') = \frac{(n_1 + n_2)}{1 - r'}(-1) + \frac{(n_3 + n_4)}{r'} = \frac{-Nr' + n_3 + n_4}{(1 - r')r'} \qquad \text{Eq.(13)}$$

fazendo-se $f'(r') = 0$, tem-se:

$$\hat{r}' = \frac{n_3 + n_4}{N} \qquad \text{Eq.(14)}$$

onde $\hat{r}'$ é o estimador de máxima verossimilhança da freqüência de recombinação na população RIL.

Uma estatística de grande utilidade para inferir a respeito da qualidade da estimativa de freqüência de recombinação é a estimativa de sua variância. A variância de uma estimativa de freqüência de recombinação é dada pelo inverso do conteúdo de informação I(r) de Fisher, conforme descrito por Schuster e Cruz (2004), e dada a seguir:

$$V(r) = \frac{1}{I(r)} \qquad \text{Eq.(15)}$$

onde, $I(r)$ é definido como o negativo da derivada segunda da função suporte (ou a derivada primeira da função escore):

$$I(r) \cong -\left(\frac{\partial^2 \ln L(r \mid n_i)}{\partial r} \right) = -f''(r) \qquad \text{Eq.(16)}$$

em que, para uma população de RILs derivada por autofecundações, a expressão de máxima verossimilhança em função de r é dada como segue:

$$L(r \mid n_i) = \lambda \left[0{,}5 \left(\frac{1}{1 + 2r} \right) \right]^{n1} \left[0{,}5 \left(\frac{1}{1 + 2r} \right) \right]^{n2} \left(\frac{r}{1 + 2r} \right)^{n3} \left(\frac{r}{1 + 2r} \right)^{n4}$$

Aplicando-se o logaritmo natural na expressão acima e derivando-se em relação a r, obtém-se a função escore, como a seguir:

$$f'(r) = \frac{-2Nr + (n_3 + n_4)(1+2r)}{(1+2r)r}$$

Ainda, considerando que a função escore pode ser estabelecida pela razão de duas funções de r, tem-se:

$$f'(r) = \frac{G(r)}{H(r)} \qquad \text{Eq.(17)}$$

e, sabendo que o I(r) é dado pelo negativo da derivada de $f'(r)$, tem-se:

$$I(r) = -f''(r) = -\left\{ G(r)\left[\frac{1}{H(r)}\right]' + [G(r)]'\left[\frac{1}{H(r)}\right] \right\} \qquad \text{Eq.(18)}$$

e que, para obtenção da estimativa de r, tem-se que:

$$G(\hat{r}) = 0 \qquad \text{Eq.(19)}$$

então,

$$I(\hat{r}) = -f''(\hat{r}) = -[G(\hat{r})]'\left[\frac{1}{H(\hat{r})}\right] = \frac{-[G(\hat{r})]'}{H(\hat{r})} \qquad \text{Eq.(20)}$$

Onde, para população de RIL obtida por autofecundações, o estimador do conteúdo de informação de Fisher é dado por:

$$I(\hat{r}) = \frac{2(n_1 + n_2)}{r(1+2r)} \qquad \text{Eq.(21)}$$

mas, como apresentado na Quadro 2 os valores esperados de n_1 e n_2 para uma população de tamanho N podem ser obtidos por:

$$n_1 = n_2 = \left[\frac{1}{2(1+2r)}\right]N \qquad \text{Eq.(22)}$$

e então, substituindo os valores de n_1 e n_2 da equação (22) na equação (21), tem-se:

$$I(\hat{r}) = \frac{2N}{r(1+2r)^2} \qquad \text{Eq.(23)}$$

Sendo, portanto, a variância dada por:

$$V(\hat{r}) = \frac{r(1+2r)^2}{2N} \qquad \text{Eq.(24)}$$

2.1.7 Análise de ordens de marcas

O passo seguinte à determinação da freqüência de recombinação entre pares de marcas, seria a formação dos grupos de ligação. O ordenamento inicial dos locos é feito com base nas informações sobre a porcentagem de recombinação entre cada par de marcas e os respectivos valores de LOD escore. Neste processo, o ordenamento é feito inicialmente considerando o par de marcas mais próximo entre si, atendendo a dois critérios (porcentagem máxima de recombinação e LOD mínimo) e, então, identificam-se as marcas mais próximas em relação às marcas já consideradas. Se estas atenderem aos dois critérios adotados, são incorporadas ao grupo de ligação. Assim, o processo continua, investigando a porcentagem de recombinação e LOD entre marcas e possíveis vizinhos a serem incorporados às extremidades do grupo de ligação em construção.

Após a formação preliminar do grupo de ligação, o próximo passo é verificar se os *n* marcadores encontram-se na melhor ordem possível. Um dos processos é gerar todas as ordens possíveis (*n*!/2) e adotar como critério a identificação da melhor ordem como sendo aquela que proporciona menor soma de distâncias. Contudo, na medida em que o número de locos marcadores aumenta, aumenta-se grandemente o número de ordens possíveis, consumindo, portanto, um grande tempo de processamento. Outro processo utilizado é o método da soma das frações de recombinação adjacentes (SARF - *Sum of Adjacent Recombination Fractions*), em

que a melhor ordem é aquela que apresentar 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 de menor soma de distâncias adjacentes (Schuster e Cruz, 2004).

Após a obtenção da ordem final dos marcadores, determina-se facilmente a distância de mapa entre marcadores utilizando-se as freqüências de recombinação e a função de mapeamento a qual se deseja apresentar o mapa de ligação.

Como mencionado anteriormente, no processo de análise de ligação fatorial além da freqüência de recombinação é utilizado um segundo critério para inferir sobre a presença ou ausência de ligação. Este critério, denominado de LOD escore, é um teste de razão de verossimilhança que utiliza o logaritmo de base 10 ($\log_{10}$) da razão entre a hipótese de ligação com um determinado valor de freqüência de ligação diferente de 0,5 e a hipótese de ausência de ligação, ou seja, freqüência de recombinação igual a 0,5. A expressão de LOD escore é apresentada a seguir:

$$LOD = \log_{10}\left[\frac{L(r \mid n_i)}{L(r = 0{,}5 \mid n_i)}\right] \qquad \text{Eq.(25)}$$

Para as populações de RIL a expressão de cálculo do valor de LOD escore para qualquer estimativa de freqüência de recombinação (r) e tamanho de população é dada por:

$$LOD = \log_{10}\left\{\frac{\lambda\left[\frac{1}{2}(1-r')\right]^{n1}\left[\frac{1}{2}(1-r')\right]^{n2}\left(\frac{1}{2}r'\right)^{n3}\left(\frac{1}{2}r'\right)^{n4}}{\lambda\left(\frac{1}{4}\right)^{N}}\right\} \qquad \text{Eq.(26)}$$

Trabalhando-se a expressão acima se chega a:

$$LOD = N\log_{10}\left(\frac{1}{2}\right) + (n1+n2)\log_{10}(1-r') + (n3+n4)\log_{10}(r') - N\log_{10}\left(\frac{1}{4}\right) \qquad \text{Eq.(27)}$$

Onde, N é o tamanho da amostra e as demais variáveis são definidas de acordo com o apresentado anteriormente no Quadro 2.

O LOD escore é considerado um teste de significância usado para testar a hipótese de ligação de dois locos gênicos. O resultado deste teste pode ser interpretado da seguinte forma: quando $L(r \mid n_i) > L(r = 0{,}5 \mid n_i)$, o valor de LOD escore é positivo. Conclui-se que os locos estão ligados geralmente quando o valor de LOD escore é maior do que 3. O valor de LOD escore igual a 3 implica que há uma probabilidade 1.000 vezes maior da presença de ligação do que da ausência de ligação entre os locos gênicos estudados.

2.2 Tamanho de população RIL e número de marcadores

O tamanho da população utilizada em mapeamento, assim como o número mínimo de marcas para representar os cromossomos nos grupos de ligação ainda não são bem definidos, não havendo ainda um consenso quando da análise de trabalhos de mapeamento disponíveis na literatura. Para ilustrar esta deficiência foi realizada uma revisão de trabalhos em que foram utilizadas populações de RIL para o mapeamento genético e/ou identificação de QTLs (Quadro 3).

Quadro 3 - Tamanho de população RIL, número e tipos de marcadores, além de outras características de mapeamento observados em diversos trabalhos disponíveis na literatura, os quais objetivaram a obtenção de mapas genéticos e/ou identificação de QTLs

Referência	Cultura ou espécie	Tamanho de População	Marcadores	Função de mapeamento	Comprimento do mapa (cM)	Parâmetros	
						$LOD_{mín}$	$r_{máx}$
Burr *et al.* (1988)	milho	48	134 marcadores (isoenzimas e RFLP)	-	-	-	-
Hnetkovsky *et al.*(1996)	soja	100	70 marcadores dos quais 57 ligados (22 RFLP e 48 RAPD) e 13 marcadores não ligados	-	1065	-	-
Chang *et al.* (1997)	soja	100	131 marcadores (90 RAPD, 3 SSR, 27 RFLP, 10 DAF – DNA Amplification Fingerprint, 1 morfológico) dos quais 74 foram mapeados em 23 grupos de ligação	Haldane	(18)	2	30 cM
Groh *et al.* (1998)	milho	143	146 RFLP	Haldane	2117 (14,4)	3	-
Groh *et al.* (1998)	milho	187	136 RFLP	Haldane	1564 (11,5)	3	-
Séne *et al.* (2000)	milho	145	148 RFLP e 31 Isoenzimas	-	-	-	-
Flowers *et al.* (2000)	arroz	150	-	-	-	-	-

Continua na próxima página

Quadro 3 - Continuação

Referência	Cultura ou espécie	Tamanho de População	Marcadores	Função de mapeamento	Comprimento do mapa (cM)	Parâmetros	
Price *et al.* (2000)	arroz	205	232 marcadores distribuídos da seguinte forma: 101 RFLP mais 34 AFLP ligados corretamente; 5 RFLP e 75 AFLP colocados de forma não confiável nos grupos; 16 AFLP e 1 RFLP colocados em dois grupos de ligação ou não ligados	Haldane	1680	3	40%
Flowers *et al.* (2000)	arroz	150	-	-	-	-	-
Ferreira *et al.* (2000)	soja	330	250 RFLP e 106 RAPD	Kosambi	3275	3	30 cM
Zhang *et al.* (2001)	arroz	150	103 RFLP e 104 AFLP	Kosambi	2419,5 (11,7)	3	-
Nachit *et al.* (2001)	trigo	110	444 marcadores dos quais 306 ligados (138 RFLP, 26 SSR, 134 AFLP, 5 proteína de semente e 3 genes conhecidos)	Haldane	3598 (11,8)	3	40 %
Miklas *et al.* (2001)	feijão	67	245 marcadores (222 AFLP, 10 RFLP, 3 RAPD, 2 SCARS, 4 fenotípicos, 3 isoenzimas e 1 proteína de semente	Kosambi	853	3	30 cM
Meksem *et al.* (2001)	soja	100	133 marcadores SSR	Haldane	2823 (26)	2	30 cM
He *et al.* (2001)	arroz	107	154 marcadores (RFLP e SSR)	Kosambi	1465	3	-

Continua na próxima página

Quadro – 3 Continuação

Referência	Cultura ou espécie	Tamanho de População	Marcadores	Função de mapeamento	Comprimento do mapa (cM)	Parâmetros	
Iqbal *et al.* (2001)	soja	100	133 SSR dos quais 107 ligados	Haldane	2823,1 (26,4)	2	30 cM
Cardinal *et al.* (2001)	milho	183	185 marcadores (120 RFLP e 65 SSR) dos quais 161 compuseram o mapa de ligação	Haldane	1667	3	40 cM
Yuan *et al.* (2002)	soja	100	135 SSR dos quais 107 ligados	-	2823,1 (26,4)	2	30 cM
Xu *et al.* (2002)	arroz	160	182 RFLP	-	-	-	-
Wang *et al.* (2002)	arroz	150	207	Kosambi	2419,5	3	-
Xing *et al.* (2002)	arroz	240	213 marcadores (168 RFLP, 45 SSR e 1 característica morfológica)	-	1796 (8,7)	3	-
Price *et al.* (2002)	arroz	140	142 marcadores (102 RFLP, 34 AFLP, 6 SSR)	-	1779	-	-
Njiti *et al.* (2002)	soja	90	112 marcadores (8 RFLP, 34 RAPD, 55 SSR, 13 AFLP, 2 morfológicos)	Haldane	1662	2	30 cM
Cho *et al.* (2002)	soja	70	1 marcador SSR a cada 25 cM do mapa consenso	-	-	2	30
Cai e Morishima (2002)	arroz	125	147 marcadores (121 RFLP, 17 isoenzimas, 2 proteínas marcadoras, 1 RAPD e 6 características qualitativas)	-	1192	-	-

Continua na próxima página

Quadro – 3 Continuação

Referência	Cultura ou espécie	Tamanho de População	Marcadores	Função de mapeamento	Comprimento do mapa (cM)	Parâmetros	
Borevitz *et al.* (2002)	*Arabidopsis thaliana*	160	163	-	-	-	-
Zheng *et al.* (2003)	arroz	96	249	Kosambi	-	-	-
Burnham *et al.* (2003)	soja	66	59	-	-	3	-
Burnham *et al.* (2003)	soja	64	75	-	-	3	-
Burnham *et al.* (2003)	soja	79	65	-	-	3	-
Média geral		129	171				

Observações: a) O sinal de traço (-) significa dado não disponível no artigo;

b) Os números entre parênteses na coluna “Comprimento do mapa (cM)” são referentes às distâncias médias dos marcadores nos grupos de ligação.

Apesar da ausência de consenso na escolha do tamanho da população RIL de mapeamento, Wu *et al*. (2003), por meio de simulação de dados em computador, estudaram a eficiência na identificação de QTLs com a utilização de três métodos de obtenção de populações RIL em três diferentes tamanhos de população. Neste estudo foi simulado, por meio do procedimento de Monte Carlo, um genoma com 4 cromossomos e 64 marcadores (16 marcadores por cromossomo não eqüidistantes) com distância média de marcadores adjacentes de 10 cM. Quatro QTLs com efeitos diferentes foram colocados nos cromossomos, um em cada. Além disto, foram testados dois níveis de herdabilidade, 20% e 50%. Para populações RIL derivadas pelo método SSD, geralmente os efeitos dos QTLs e suas posições foram mais próximas dos valores pré-fixados para populações maiores do que para as menores, sendo, que as populações com 150 indivíduos apresentaram desejáveis propriedades no mapeamento de QTLs.

3 MATERIAL E MÉTODOS

3.1 Simulação de dados

Para a geração dos dados de populações de RILs (*Recombinant inbred line*) de uma espécie vegetal diplóide, foi utilizado o módulo de simulação do aplicativo computacional GQMOL (Cruz, 2004), o qual permite gerar informações sobre genomas, genótipos de genitores, indivíduos de diferentes tipos de populações e dados de características quantitativas.

3.1.1 Simulação do genoma

Com base em informações da literatura tomando por base a espécie *Phaseolus vulgaris* (2n= 2x= 22), cujo comprimento total do genoma é estimado em 1.200 cM, foram gerados três genomas com vários níveis de saturação por marcas. Cada genoma foi composto por 11 grupos de ligação, com 100 cM em cada grupo e, portanto, com um comprimento total de 1.100 cM.

3.1.1.1 Níveis de saturação do genoma e tipos de marcas

Foram gerados genomas com três níveis de saturação: saturado, medianamente saturado, pouco saturado, com intervalos entre marcas adjacentes de 5, 10 e 20 cM, respectivamente (Figuras 2, 3 e 4). O número de marcas nos três genomas foi de 231, 121 e 66, respectivamente.

Para as populações do tipo RIL, o uso de marcas dominantes, codominantes ou dominantes e codominantes levam aos mesmos resultados, pois, a proporção genotípica esperada na população é de 1:1 (AA:aa).

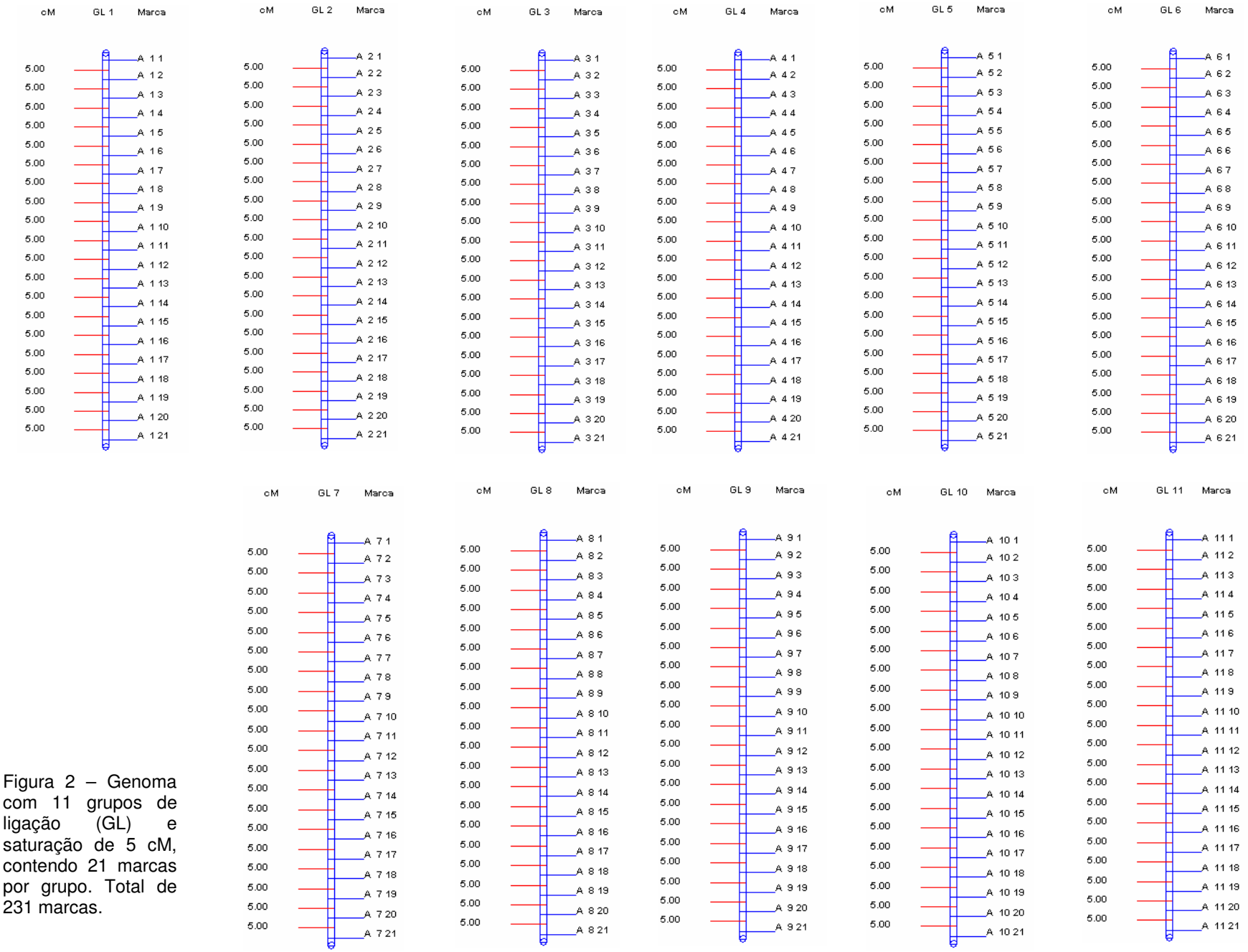


Figura 2 – Genoma com 11 grupos de ligação (GL) e saturação de 5 cM, contendo 21 marcas por grupo. Total de 231 marcas.

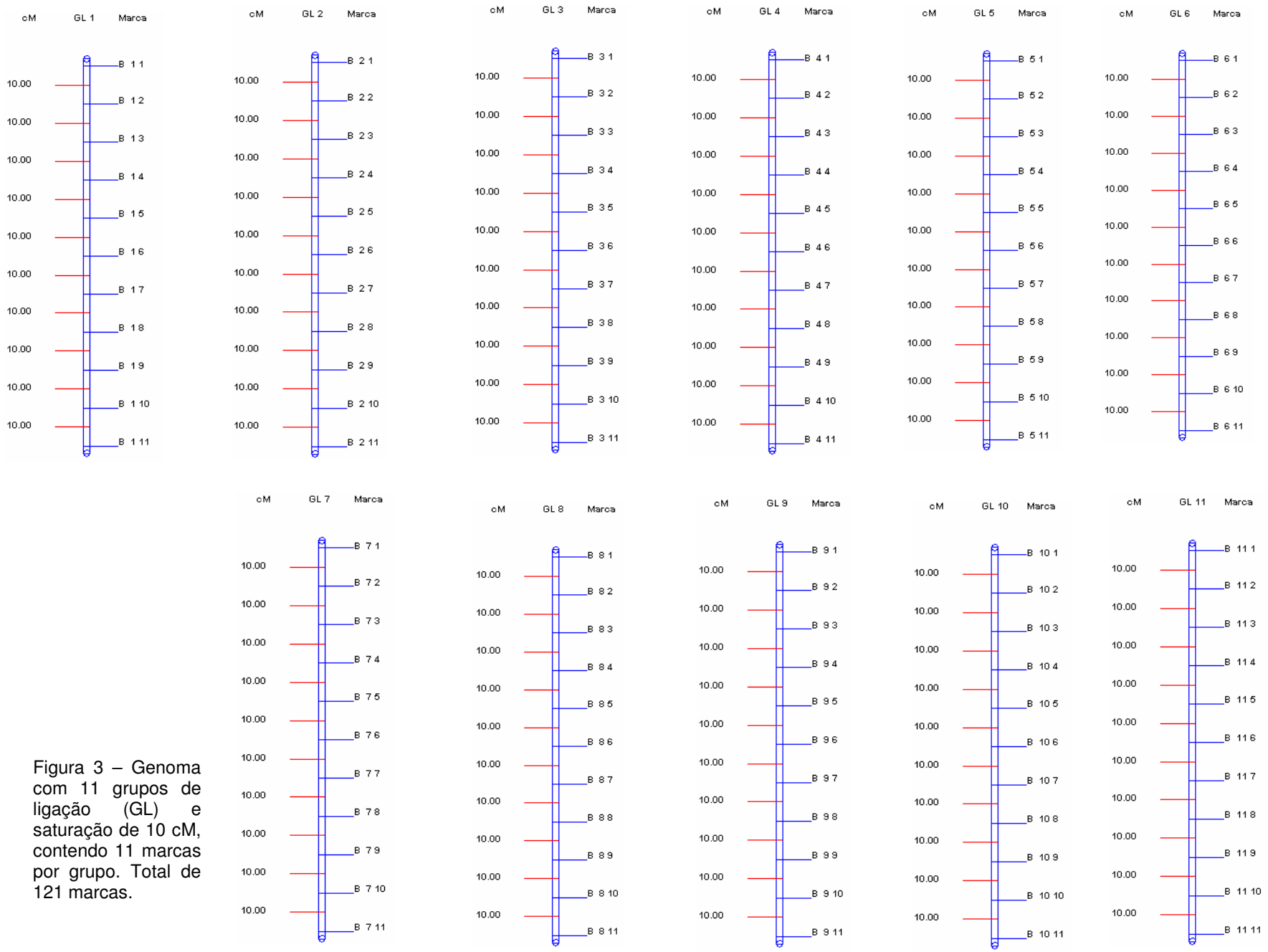


Figura 3 – Genoma com 11 grupos de ligação (GL) e saturação de 10 cM, contendo 11 marcas por grupo. Total de 121 marcas.

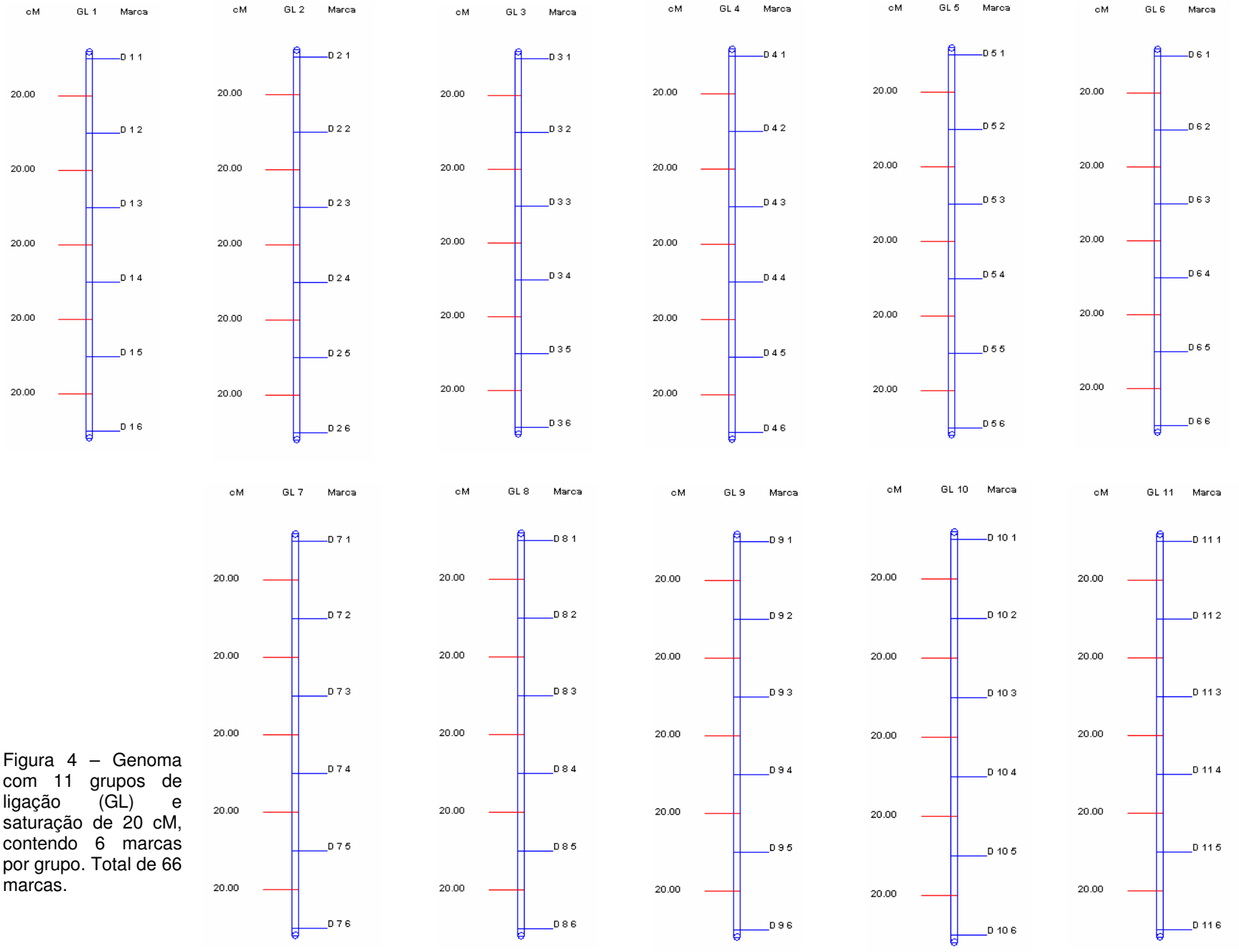


Figura 4 – Genoma com 11 grupos de ligação (GL) e saturação de 20 cM, contendo 6 marcas por grupo. Total de 66 marcas.

3.1.2 Simulação de genitores

Para cada saturação do genoma estudado foi simulado apenas um tipo de arranjo de genes para a geração F_1, qual seja, um genitor homozigoto dominante e o outro homozigoto recessivo (Figura 5). Situação esta que produziu uma geração F_1 com todos os locos em fase de acoplamento.

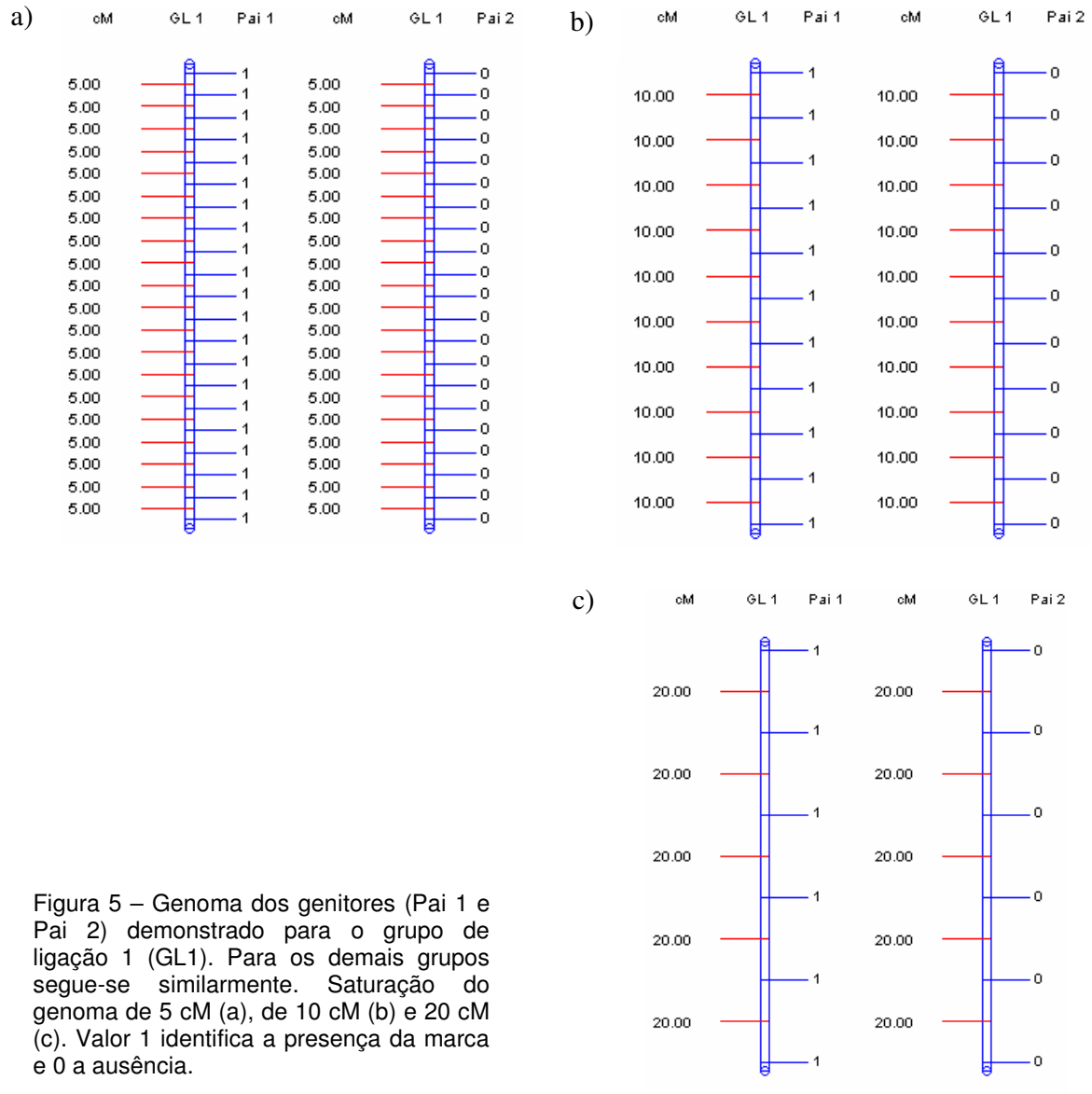


Figura 5 – Genoma dos genitores (Pai 1 e Pai 2) demonstrado para o grupo de ligação 1 (GL1). Para os demais grupos segue-se similarmente. Saturação do genoma de 5 cM (a), de 10 cM (b) e 20 cM (c). Valor 1 identifica a presença da marca e 0 a ausência.

3.1.3 Tamanho de população

Para cada uma das situações estudadas foram geradas 100 populações com 50, 100, 154, 200, 300, 500 e 800 indivíduos. Foram, portanto, gerados 3 “genomas simulados” e 2.100 “genomas analisados” (3 níveis de saturação x 7 tamanhos de população x 100 repetições).

Uma população de RIL com 154 indivíduos foi utilizada por Faleiro (2000) para o mapeamento genético e identificação de QTLs em *Phaseolus vulgaris*. Para verificar a sua qualidade para fins de mapeamento, este tamanho de população foi utilizado neste trabalho de simulação.

3.1.4 Procedimento de simulação dos indivíduos da população

Para cada indivíduo podem ser produzidos os dados genotípicos referentes aos marcadores de acordo com a informação do genoma. A estratégia básica de simulação é baseada em permutas nos intervalos entre marcas adjacentes em cada cromossomo, de acordo com as distâncias dos marcadores, conforme descrito abaixo.

O processo de simulação seguiu os seguintes passos: i) a partir do genoma simulado foram construídos os genótipos parentais homozigotos e contrastantes para os marcadores, de tal forma que a geração F_1 estava em fase de acoplamento para todos os pares de marcadores; ii) a partir do genótipo da geração F_1 foram gerados os gametas para a formação dos indivíduos das populações de RILs. A produção de gametas foi feita simulando-se o pareamento dos homólogos e realizando-se permutas ao longo dos cromossomos nas regiões delimitadas por dois marcadores adjacentes. A probabilidade de ocorrência de recombinação numa região entre marcadores adjacentes foi dada de acordo com a distância destes marcadores no genoma simulado. Por exemplo, se a distância entre os dois primeiros marcadores num cromossomo era de 10 cM no genoma simulado, a qual em população de RIL obtida por autofecundações equivale a 16,667 [r’=2r/(1+2r)], a probabilidade de recombinação nesta região foi de 16,667%. Depois de decidir se havia ocorrido ou não a recombinação nesta região, passava-se para a próxima região delimitada pelo segundo e terceiro marcadores. O procedimento continuou até que todas as regiões entre marcadores adjacentes no cromossomo tivessem sido alcançadas. Para a formação de cada indivíduo nas populações foram simulados vários gametas, sendo

sorteado apenas um gameta para formação de cada indivíduo, sendo, que o genótipo do indivíduo obtido pela duplicação da informação contida do gameta.

O esquema apresentado na Figura 6 e Quadro 4 abaixo ilustra o processo de obtenção dos 50 indivíduos de uma população segregante originada de uma F_1 produzida pelos genitores com nível de saturação do genoma de 20 cM.

Nas Figuras 6 e 7 e Quadro 4 é ilustrado o processo de formação dos gametas e indivíduos das populações simuladas neste trabalho.

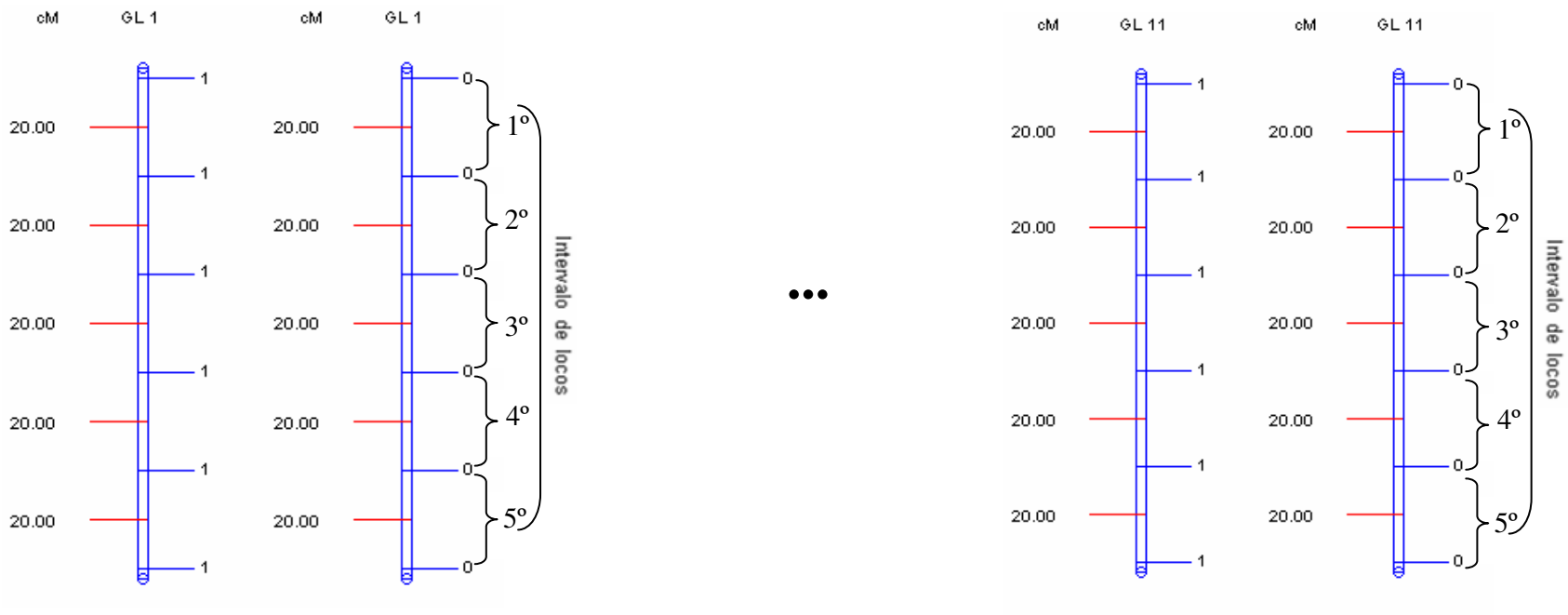


Figura 6 – Ilustração do processo de formação de gametas para a geração dos indivíduos de uma população segregante RIL com a utilização de uma geração F_1 cujo genótipo apresenta nível de saturação de 20 cM. Para simplificação são apresentados apenas os homólogos referentes aos grupos de ligação 1 e 11. Os intervalos entre locos estão apresentados para os grupos de ligação.

Quadro 4 – Complementação ao apresentado na Figura 6 ilustrando a formação dos indivíduos de uma população segregante de RIL com 50 indivíduos oriundo de uma geração F_1 cujo genótipo apresenta nível de saturação de 20 cM. É apresentado um exemplo hipotético da formação dos gametas, assim como dos genótipos de três indivíduos formados (1, 2 e 50)

Grupos de ligação	Intervalo de locos	Número sorteado[1]	Gametas		Indivíduo 1	Número sorteado[1]	Gametas		Individuo 2	...	Número sorteado[1]	Gametas		Indivíduo 50
			1	2			1	2				1	2	
GL 1			1	0	1		1	0	1	...		1	0	1
	1º	200	0	1	0	900	0	1	0	...	4.780	1	0	1
	2º	500	1	0	1	760	1	0	1	...	290	0	1	0
	3º	7.000	1	0	1	5.500	1	0	1	...	6.150	0	1	0
	4º	300	0	1	0	9.110	1	0	1	...	1.110	1	0	1
	5º	8.000	0	1	0	4.560	1	0	1	...	8.010	1	0	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
GL 11			1	0	1		1	0	1	...		1	0	1
	1º	400	0	1	0	20	0	1	0	...	1.290	0	1	0
	2º	500	1	0	1	6.500	0	1	0	...	9.310	0	1	0
	3º	5.000	1	0	1	7.500	0	1	0	...	6.740	0	1	0
	4º	150	0	1	0	1.700	1	0	1	...	190	1	0	1
	5º	7.320	0	1	0	5.900	1	0	1	...	5.400	1	0	1

[1]/ Número sorteado foi um número gerado aleatoriamente de 1 a 10.000 pelo computador. Para a sua geração foi utilizada uma função que levou em consideração o tempo, evitando assim a geração de números viciados;

Notas:

a) A distância entre marcas é de 20 cM (Figura 6), o que leva a uma freqüência de recombinação observada nas RIL de 38,46 % [r'=2r/(1+2r)]. Para evitar erros em função das casas decimais, as freqüências de recombinação das RIL foram multiplicadas por 100. Portanto, como explicado anteriormente, quando o número sorteado foi inferior a 3.846 houve *crossing*-over na região delimitada pelos locos em questão;

b) Para a formação do indivíduo foi sorteado um gameta (para cada um dos indivíduos 1, 2 e 50 demonstrados foi sorteado o gameta 1 em cada caso). Neste exemplo, foi sorteado um dos dois gametas formados em uma única meiose. Porém, no processo de simulação das populações estudadas neste trabalho foram gerados milhares de gametas e sorteado um para formação de cada indivíduo.

c) A ilustração da formação dos gametas 1 e 2 mostrados na quarta e quinta coluna do quadro acima é apresentada na Figura 7.

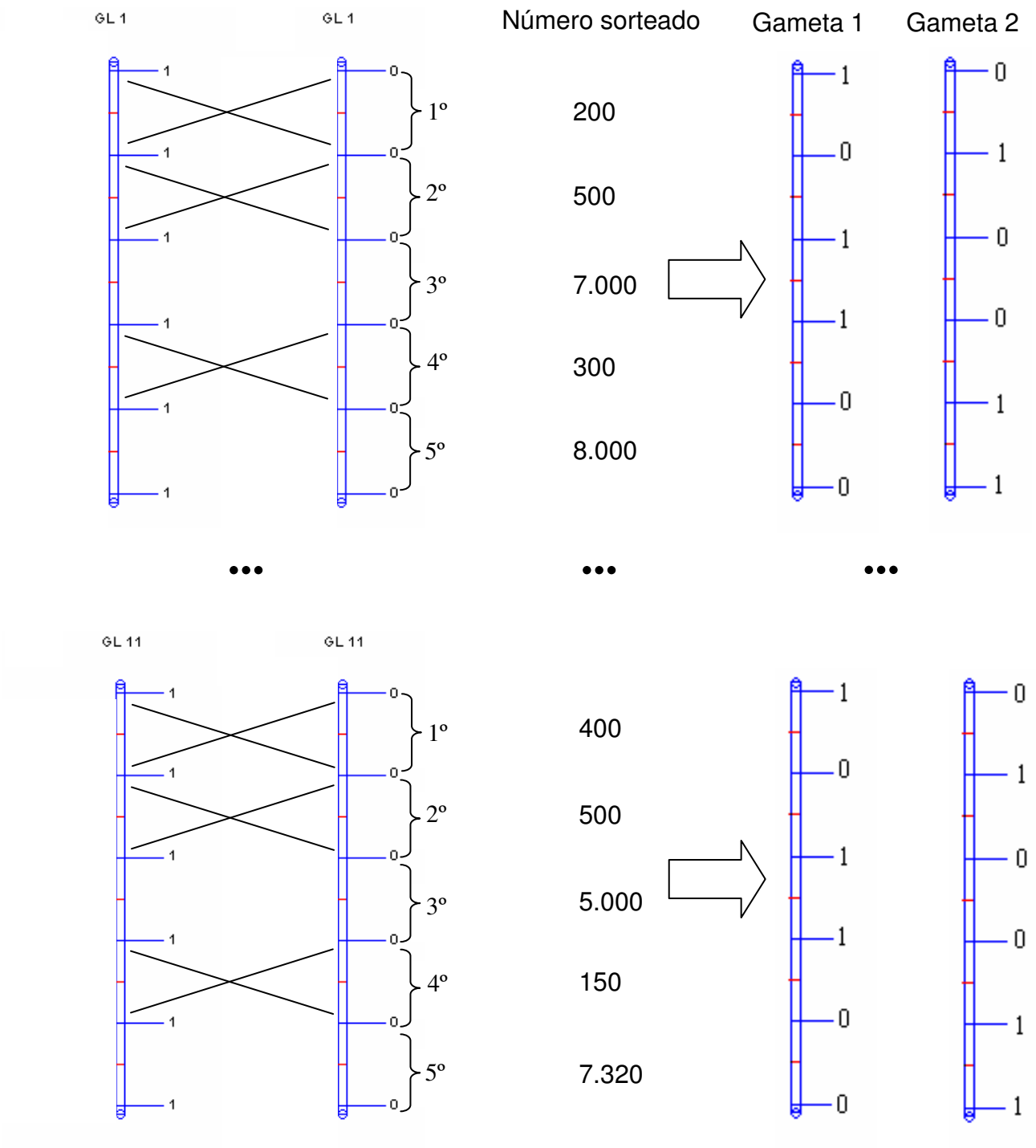


Figura 7 – Ilustração da formação dos gametas 1 e 2 mostrados na quarta e quinta coluna do Quadro 4. Quando o número sorteado foi menor que 3.846 houve *crossing over*, representado pelas linhas cruzadas entre os homólogos.

3.2 Análise genômica – Mapeamento

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

3.2.1 Análise de segregação de locos individuais

Foram aplicados testes de qui-quadrado (χ^2) para verificar a razão de segregação de cada marca em todas as populações geradas. No processo de mapeamento foram utilizadas todas as marcas, mesmo aquelas que não segregaram de acordo com a proporção esperada de 1:1 (AA:aa).

A estatística qui-quadrado é dada por:

$$\chi^2 = \sum_{i=1}^{n} \left[\frac{(Obs_i - Esp_i)^2}{Esp_i} \right] \qquad \text{Eq.(28)}$$

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 (i= 1, 2,..., n), respectivamente.

A hipótese (H_0) de segregação do loco de 1:1 (AA : aa) foi testada a 5% de probabilidade (erro tipo I). Se o valor de probabilidade calculado foi inferior ao pré-estabelecido, a hipótese H_0 foi rejeitada, ou seja, a segregação não ocorreu de acordo com o esperado.

3.2.2 Análise de pares de marcas – estimação da percentagem de recombinação

Após a aplicação dos testes de segregação, seguiu-se a etapa da estimação da percentagem de recombinação entre pares de marcas. Para esta estimação foi utilizado o método da máxima verossimilhança, conforme descrito anteriormente (Item 2.1.6 da Revisão de Literatura).

3.2.3 Determinação dos grupos de ligação

O passo seguinte à estimação das freqüências de recombinação entre pares de marcas consistiu na determinação dos grupos de ligação. Na formação do grupo de ligação utilizou-se a propriedade transitiva, ou seja, se o loco A está ligado ao loco B, e o loco B está ligado ao loco C, logo o loco A está ligado ao loco C, independente da freqüência de recombinação estimada entre A e C e, portanto, A, B e C pertencem ao mesmo grupo de ligação. Os critérios utilizados no agrupamento foram a freqüência máxima de recombinação (r_{max}) e o LOD mínimo (LOD_{min}), para inferir se dois locos estavam ligados. Neste trabalho foram utilizados os valores de 30% e 3, respectivamente, para r_{max} e LOD_{min}.

3.2.4 Ordenamento das marcas no grupo de ligação

Quando há apenas duas marcas em um grupo de ligação, apenas uma ordem é possível, uma vez que a orientação do grupo pode ser ignorada. Porém, para três marcadores são possíveis 3 ordens, e para n marcadores, são possíveis n!/2 ordens. Portanto, após a formação dos grupos de ligação, é necessário determinar a melhor ordem das marcas nos grupos. O método utilizado para isso foi o da SARF (*Sum of Adjacent Recombination Fractions*), conforme descrito anteriormente (Item 2.1.7 da Revisão de Literatura).

Realizados todos os passos descritos até então, os grupos de ligação para as populações simuladas foram formados utilizando-se como medida de distância a porcentagem de freqüência de recombinação. O próximo passo foi a comparação dos grupos de ligação formados para todas as populações simuladas com aqueles grupos de ligação estabelecidos no genoma simulado, como descrito a seguir.

3.3 Comparação de genomas

Para facilitar a compreensão, o termo “genoma simulado” refere-se, genericamente, aos genomas simulados conforme descrito nos itens 3.1.1 e 3.1.1.1 dessa seção, o qual deve ser referenciado como o verdadeiro. O termo “genoma analisado” refere-se, genericamente, aos genomas construídos a partir das populações simuladas, o qual apresentará distorção em razão do processo de estimação, amostragem, critérios de agrupamento, dentre outros.

Foram estudados nos genomas analisados, em cada situação, os números de grupos de ligação obtidos, o número de marcas por grupo, 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 se houve ou não inversão da ordem dos marcadores, verificada pela correlação de Spearman. Todas estas comparações foram realizadas com utilização do módulo "Comparação de genomas" do aplicativo computacional GQMOL (Cruz, 2004).

Nas análises apresentadas a seguir foram utilizadas apenas as repetições (populações) em que houve a formação de onze grupos de ligação no mapeamento genético. Exceção feita somente para o "número de grupos de ligação" em que foram utilizadas todas as 100 repetições independentemente do número de grupos de ligação formado.

3.3.1 Número de grupos de ligação e marcas por grupo

Para todos os genomas analisados foi feita uma contagem do número de grupos de ligação e número de marcas por grupo de ligação obtidos do mapeamento das populações simuladas.

3.3.2 Tamanho do grupo de ligação

É o somatório das distâncias entre marcas adjacentes no grupo de ligação do genoma analisado, como segue:

$$L = \sum_{k=1}^{m-1} d_k \qquad \text{Eq.(29)}$$

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

3.3.3 Distância média entre marcadores adjacentes no grupo de ligação

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

$$\bar{d} = \frac{L}{I} \qquad \text{Eq. (30)}$$

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

3.3.4 Variâncias das distâncias entre marcas adjacentes

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

$$\hat{\sigma}^2 = \frac{\sum_{k=1}^{m-1}\left(d_k - \bar{d}\right)^2}{I-1} \qquad \text{Eq. (31)}$$

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

3.3.5 Correlação de Spearman

Também conhecida como correlação de *rank*, é utilizada quando não é possível mensurar variação contínua das variáveis *x* e *y* nos *n* membros de uma população. Porém, é possível mensurar um *x* e um *y*, em forma de nota (*rank*), onde cada nota pode ser colocada em ordem para os *n* membros. Esta correlação expressa o grau de concordância nas notas das duas variáveis, desta forma estamos propondo a sua utilização na análise de genomas conforme descrito a seguir.

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

$$r_s = 1 - \frac{6\sum_{k=1}^{m} \Delta_k^2}{m\left(m^2 - 1\right)} \qquad \text{Eq. (32)}$$

em que: r_s é o valor estimado da correlação de Spearman para um grupo de ligação do genoma analisado $\left(-1 \leq r_s \geq 1\right)$, Δ_k é a diferença da nota do marcador m_k (k = 1,..., m) na posição k do grupo de ligação do genoma simulado e a nota do marcador m_k na posição k do grupo de ligação do genoma analisado.

Onde, m é o número de marcadores no grupo de ligação do genoma simulado e a nota do marcador m_k, tanto no grupo de ligação do genoma simulado quanto no grupo de ligação do analisado, é o valor do índice k do referido marcador. Por exemplo, se no grupo de ligação do genoma simulado a ordem dos marcadores for: m_1-m_2-m_3-...-m_m, então, a nota do marcador m_1 é 1, do m_2 é 2 do m_3 é 3 e do m_m e´ m. E, se no grupo de ligação do genoma analisado a ordem dos respectivos marcadores for m_2-m_1-m_3-...-m_m, então, a nota do marcador m_2 é 2, do m_1 é 1, do m_3 é 3 e do m_m é m. Portanto, os valores de Δ_k serão: Δ_1 = (1-2)= -1, Δ_2 =(2-1)= 1, Δ_3 = (3-3)= 0 e Δ_m = (m-m)= 0.

Sua aplicação na análise de genoma é ilustrada utilizando-se o esquema apresentado na Figura 8, que mostra um cromossomo simulado e outro analisado. No cromossomo simulado (a) tem-se os marcadores de M_1 a M_7, em ordem crescente da esquerda para a direita; no cromossomo analisado (b), tem-se, também, os mesmos marcadores, porém, os marcadores M_1 e M_2 estão em posições trocadas em relação ao cromossomo simulado (a).

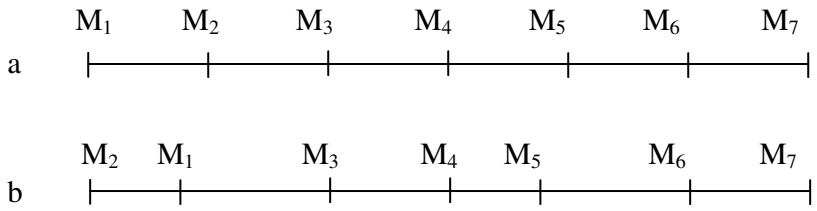


Figura 8- Cromossomo simulado (a) e analisado (b). Ambos com os mesmos 7 marcadores.

Para o cálculo da correlação entre os marcadores para estes dois cromossomos devem-se atribuir notas a cada marcador. Atribuindo-se o número do

índice de cada marcador como sua nota (Quadro 5), a correlação de Sperman é obtida como se segue.

Quadro 5 – Nota (*rank*) para os marcadores nos cromossomos simulado e analisado e obtenção dos valores de Δ_k para cada par de marcadores

Rank dos marcadores no cromossomo *a*	1	2	3	4	5	6	7
Rank dos marcadores no cromossomo *b*	2	1	3	4	5	6	7
Valor de Δ_k	-1	1	0	0	0	0	0

Assim, o valor estimado da correlação é:

$$r_s = 1 - \frac{6\left[(-1)^2 + (1)^2 + (0)^2 + (0)^2 + (0^2) + (0) + (0)^2\right]}{7(7^2 - 1)} = 0{,}9643$$

Portanto, se todos os marcadores mapeados no genoma analisado mantiverem as ordens originais existentes no genoma simulado, a correlação de Spearman será igual à unidade.

Esta correlação não é afetada pelas distâncias entre marcadores. Para estimar os efeitos das mudanças nas distâncias entre os marcadores no genoma analisado em relação ao genoma original (simulado), foi utilizado o coeficiente de estresse, descrito a seguir.

3.3.6 Estresse

O coeficiente de estresse (S) é utilizado como medida de adequação da representação gráfica de medidas de dissimilaridade convertidas em escores relativos às variáveis X e Y em estudos de divergência genética (Cruz e Carneiro, 2003). Estamos propondo a sua utilização na análise de genomas conforme demonstrado a seguir:

$$S = 100.\sqrt{\frac{\sum_{k=1}^{m-1}(d_{ok} - d_k)^2}{\sum_{k=1}^{m-1} d^2_{ok}}} \qquad \text{Eq.(32)}$$

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

A aplicação do estresse na análise de genoma é ilustrada utilizando-se o esquema apresentado na Figura 9, na qual tem-se um cromossomo simulado e outro analisado. No cromossomo simulado (a), tem-se os marcadores M_1 a M_7, em ordem crescente da esquerda para a direita e eqüidistantes; no cromossomo analisado (b), tem-se, também, os mesmos marcadores, porém, não eqüidistantes.

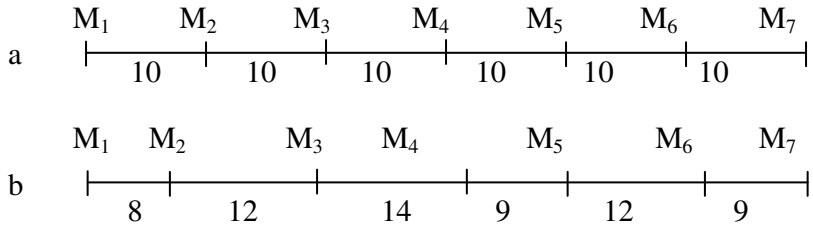


Figura 9- Cromossomo simulado (a) e analisado (b). Ambos com os mesmos 7 marcadores. As distâncias (cM) entre marcas estão representadas abaixo dos cromossomo.

Deste modo, para o cálculo do coeficiente de estresse basta obter os valores dos desvios, $d_{ok} - d_k$, como ilustrado no Quadro 6.

Quadro 6 - Distância em centiMorgan (cM) entre marcadores no genoma simulado (d_{ok}), no genoma analisado (d_k) e obtenção dos valores desvios, $d_{ok} - d_k$

	Pares de marcadores					
	M_1-M_2	M_2-M_3	M_3-M_4	M_4-M_5	M_5-M_6	M_6-M_7
d_{ok}	10	10	10	10	10	10
d_k	8	12	14	9	12	9
$d_{ok} - d_k$	2	-2	-4	1	-2	1

Assim, o valor estimado do coeficiente de estresse será:

$$S = 100.\sqrt{\frac{(2)^2 + (-2)^2 + (-4)^2 + (1)^2 + (-2)^2 + (1)^2}{10^2 + 10^2 + 10^2 + 10^2 + 10^2 + 10^2}} = 22{,}36\%$$

Portanto, se as distâncias entre os marcadores no genoma analisado mantiverem-se as mesmas com relação ao genoma simulado, o valor estimado do estresse será zero.

3.4 Testes de comparação múltipla de médias

As médias das variáveis, tamanho de grupo de ligação, distância média de marcas adjacentes, variância e estresse para cada grupo de ligação obtidas para vários tamanhos de população, dentro de cada nível de saturação do genoma, foram comparadas pelo teste de comparação múltipla de médias de Tukey a 1% de probabilidade (erro tipo I), com o auxílio do icativo computacional GENES (Cruz, 1999). Também foram comparadas as médias gerais (médias de todos os grupos de ligação), para cada tamanho de população, dentro de cada nível de saturação do genoma.

3.5 Fluxograma ilustrativo

Para facilitar o entendimento da logística utilizada no trabalho de simulação é apresentado o fluxograma abaixo (Figura 10). Os passos seguidos no processo de simulação foram: 1º) simulação dos genomas em três níveis de saturação por marcas (5, 10 e 20 cM); 2º) a partir dos genomas simulados foram construídos os genótipos dos genitores; 3º) após construídos os genótipos do genitores foi simulada a geração F_1; 4º) utilizando a geração F_1 foram simuladas as populações segregantes com o número de indivíduos desejado; 5º) as populações segregantes foram mapeadas; 6º) os mapas obtidos foram comparados com o genoma simulado. Para a comparação foram utilizados os critérios apresentados no quadro mostrado no interior do fluxograma.

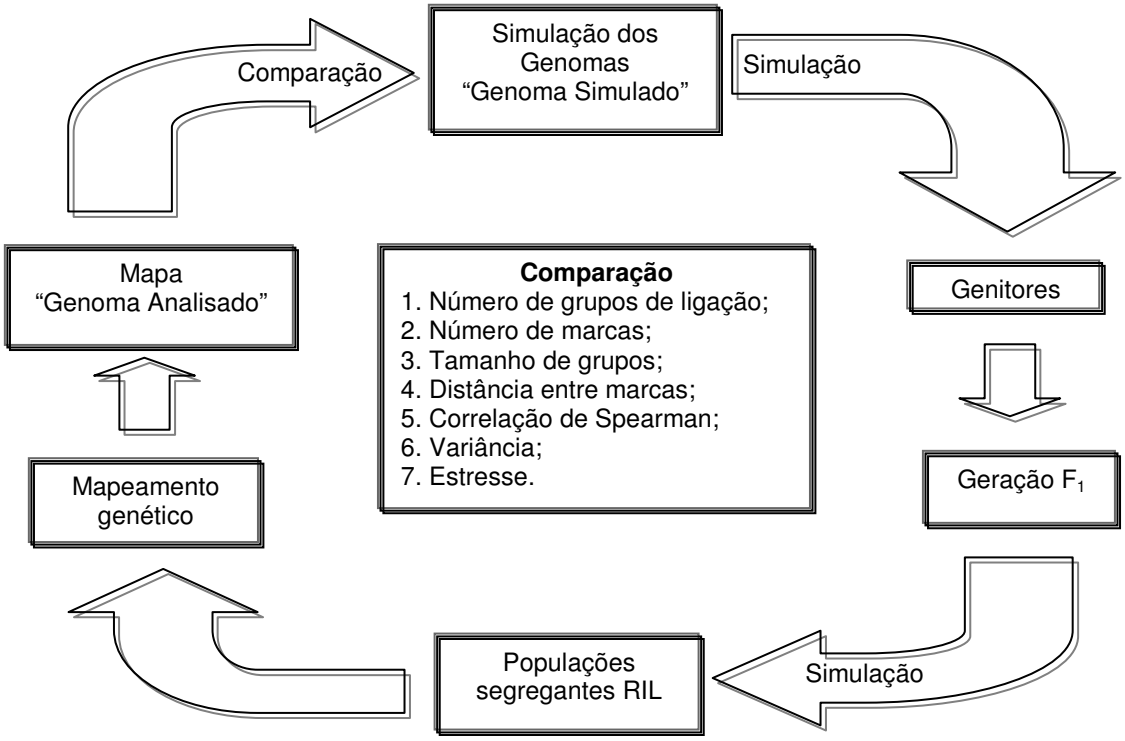


Figura 10 – Fluxograma ilustrativo do processo de simulação utilizado neste trabalho.

4 RESULTADOS E DISCUSSÃO

A espécie *Phaseolus vulgaris* (2n= 2x= 22), cujo comprimento total do genoma é estimado em 1.200 cM, foi utilizada como modelo para a simulação de genomas com três níveis de saturação por marcas. Os três níveis de saturação foram: saturado, medianamente saturado, pouco saturado, com distâncias entre marcadores adjacentes de 5 cM, 10 cM e 20 cM, respectivamente. Cada genoma com um dado nível de saturação foi composto por 11 grupos de ligação, 100 cM em cada grupo e, portanto, com um comprimento total de 1.100 cM. Os três genomas construídos foram utilizados para a geração de diferentes populações simuladas segregantes do tipo RIL, com 50, 100, 154, 200, 300, 500 e 800 indivíduos em cada um dos três níveis de saturação. Cem réplicas de cada uma destas sete populações foram utilizadas para a construção de mapas genéticos, que, posteriormente, foram comparados com o genoma a partir do qual cada população foi gerada, utilizando para a comparação, os seguintes critérios: 1) número de grupos de ligação obtidos no mapeamento; 2) número de marcas obtidas por grupo de ligação no mapeamento; 3) correlação de Spearman; 4) tamanho de grupos de ligação; 5) distância média entre marcadores adjacentes; 6) variância das distâncias entre marcas adjacentes e; 7) estresse.

4.1 Número de grupos de ligação obtidos no processo de mapeamento

Previamente à realização do mapeamento genético de cada população segregante simulada, foram feitos testes de qui-quadrado (χ^2) para verificar a razão de segregação de locos individuais. Observou-se raras ocorrências de marcadores com razão de segregação diferente do esperado de 1:1 ($P< 0,05$) (dados não apresentados). Estes resultados serviram como um indicativo da boa qualidade dos dados gerados no processo de simulação.

O número de grupos de ligação esperado no processo de mapeamento das populações simuladas era igual a 11, independente do tamanho de população e nível

de saturação do genoma utilizado para geração das populações. Contudo, esse número variou de acordo com o nível de saturação do genoma o qual foi utilizado para geração de cada população simulada e de acordo com o tamanho da população segregante, conforme apresentado nas Figuras 11, 12 e 13 e Quadro 7, e discutido a seguir.

4.1.1 Efeito do tamanho de população na formação dos grupos de ligação

4.1.1.1 Populações segregantes simuladas a partir do genoma com saturação de 5 cM

Para as populações geradas a partir do genoma com nível de saturação de 5 cM, apenas aquelas com número de indivíduos igual a 50 levaram à formação de mais de 11 grupos de ligação (Figura 11 e Quadro 7). Além disso, também houve a presença de marcas não ligadas, embora de, no máximo, uma marca não ligada em 231 marcas (Quadro 8). Para esse tamanho de população, do total de 100 repetições, 70 apresentaram 11 grupos de ligação e as outras 30 apresentaram de 12 a 14 grupos (Figura 11).

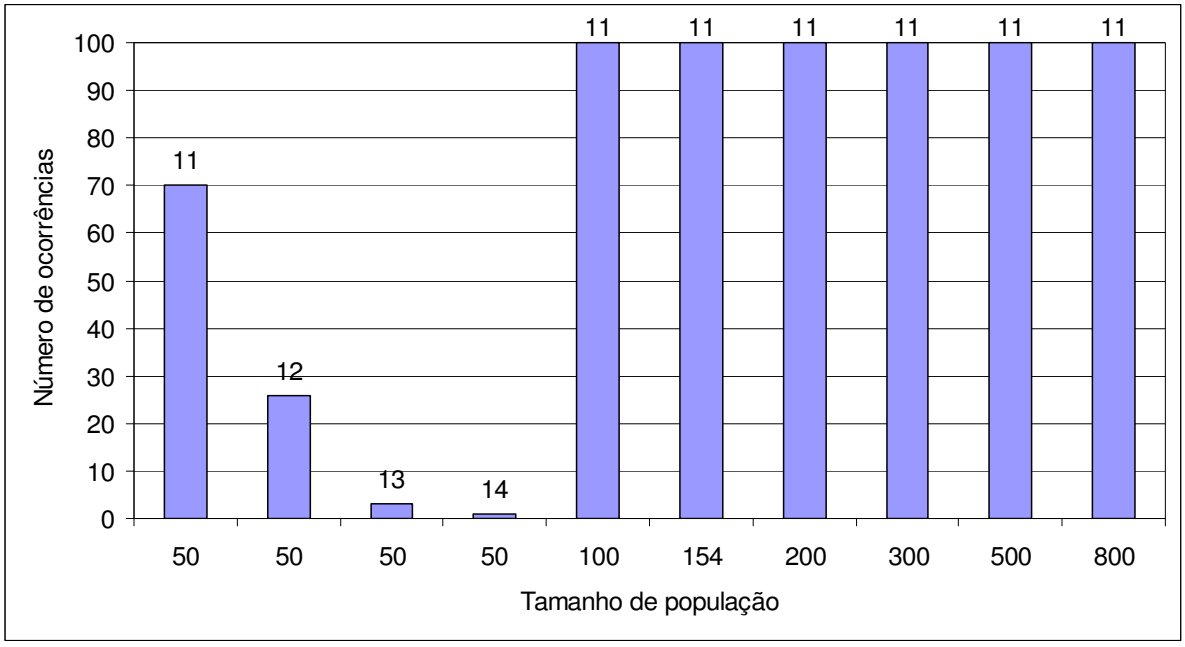


Figura 11 - Distribuição do número de grupos de ligação (indicados no topo de cada barra) obtidos no processo de mapeamento das populações simuladas em função do tamanho de população segregante. Avaliação feita em 100 repetições para cada tamanho de população com 50, 100, 154, 200, 300, 500 e 800 indivíduos, simuladas a partir do genoma com nível de saturação de 5 cM

Quadro 7 - Número de repetições em que houve a formação de 11 grupos de ligação no mapeamento das populações simuladas. O número total de repetições avaliadas foi de 100 para cada tamanho de população

Saturação do genoma	Tamanho de população	Número de repetições
5 cM	50	70 (30) [1]
	100	100 (0)
	154	100 (0)
	200	100 (0)
	300	100 (0)
	500	100 (0)
	800	100 (0)
10 cM	50	0 (100)
	100	98 (2)
	154	100 (0)
	200	100 (0)
	300	100 (0)
	500	100 (0)
	800	100 (0)
20 cM	50	12 (88) *
	100	0 (100)
	154	43 (57)
	200	86 (14)
	300	98 (2)
	500	100 (0)
	800	100 (0)

[1]\ O número de repetições em que o total de grupos de ligação foi diferente de 11 está entre parênteses;
*\ Dentre as 100 repetições para o tamanho de população de 50 indivíduos e com saturação do genoma de 20 cM houve 12 repetições com formação de 11 grupos de ligação no processo de mapeamento, mas estes grupos continham poucas marcas ligadas, conseqüentemente, estas repetições não foram utilizadas nas demais análises de comparação de genoma apresentadas neste trabalho.

Quadro 8 - Número máximo e mínimo de marcas não ligadas obtidos no mapeamento das populações simuladas. Avaliação feita nas 100 repetições, independentemente do número de grupos de ligação formado no mapeamento

Saturação do genoma	Tamanho de população	Número mínimo de marcas não ligadas	Número máximo de marcas não ligadas	Número total de marcas utilizadas no mapeamento
5 cM	50	0	1	231
	100	0	0	231
	154	0	0	231
	200	0	0	231
	300	0	0	231
	500	0	0	231
	800	0	0	231
10 cM	50	0	9	121
	100	0	1	121
	154	0	0	121
	200	0	0	121
	300	0	0	121
	500	0	0	121
	800	0	0	121
20 cM	50	24	47	66
	100	2	19	66
	154	0	3	66
	200	0	1	66
	300	0	1	66
	500	0	0	66
	800	0	0	66

Nota: A análise correta deste quadro é feita da seguinte forma, por exemplo, para o tamanho de população de 50 indivíduos e saturação de 20 cM tem-se que as repetições que apresentaram menos marcas não ligadas foi de 24 marcas e as repetições que apresentaram maiores números de marcas não ligadas foi de 47 marcas.

O número de grupos de ligação observado acima do esperado foi devido à divisão de grupos de ligação em conseqüência da não detecção de ligação entre marcas sabidamente ligadas. Visto que, no processo de mapeamento são utilizados dois parâmetros, freqüência de recombinação máxima e LOD escore mínimo, para a

inferência a respeito da presença de ligação gênica entre duas marcas, faz-se necessária uma discussão sobre as causas de não ligação gênica enfocando estes dois fatores, para verificar se apenas um ou ambos estão afetando o processo de mapeamento.

Sobre a freqüência de recombinação, sabe-se que nas populações segregantes com pequeno número de indivíduos não há uma boa representatividade da diversidade de gametas produzidos pelos parentais. E, como a determinação da distância de dois locos para o mapeamento genético é feita pela genotipagem dos indivíduos na população segregante, e contagem daqueles que são recombinantes para os locos em estudo, ou seja, aqueles originados de gametas recombinantes, uma pobre amostragem gamética certamente leva a estimativas viesadas de distância. Portanto, o tamanho inadequado de 50 indivíduos nas populações simuladas a partir do genoma com saturação de 5 cM proporcionou a não detecção de ligação genética onde sabidamente ela existia.

O tamanho de população e número de marcas utilizadas no mapeamento genético de populações RIL tem grande importância na qualidade das estimativas de freqüência de recombinação, isto é observado em termos dos parâmetros: conteúdo médio de informação, variância e desvio padrão da freqüência de recombinação. No Quadro 9, abaixo, são apresentados valores dos parâmetros acima para três estimativas de freqüências de recombinação, 0,05, 0,10 e 0,20 (ou 5, 10 e 20 cM) para vários tamanhos de população RIL.

Quadro 9 - Valores de conteúdo de informação, variância e desvio padrão das estimativas de freqüência de recombinação para vários tamanhos de população segregante RIL

Freqüência de Recombinação (r)	Tamanho de população	Conteúdo de informação[1]	Variância[2]	Desvio padrão
0,05	50	1652,89	$6,050x10^{-4}$	0,0245
	100	3305,78	$3,030x10^{-4}$	0,0173
	154	5090,90	$1,960x10^{-4}$	0,0140
	200	6611,57	$1,510x10^{-4}$	0,0122
	300	9917,35	$1,010x10^{-4}$	0,0100
	500	16528,92	$0,605x10^{-4}$	0,0077
	800	26446,28	$0,378x10^{-4}$	0,0061
0,10	50	694,44	$14,4x10^{-4}$	0,0379
	100	1388,88	$7,20x10^{-4}$	0,0268
	154	2138,88	$4,60x10^{-4}$	0,0216
	200	2777,77	$3,60x10^{-4}$	0,0189
	300	4166,66	$2,40x10^{-4}$	0,0154
	500	6944,44	$1,40x10^{-4}$	0,0120
	800	11111,11	$0,90x10^{-4}$	0,0094
0,20	50	255,10	$3,92x10^{-3}$	0,0626
	100	510,20	$1,96x10^{-3}$	0,0442
	154	785,71	$1,27x10^{-3}$	0,0356
	200	1020,40	$0,98x10^{-3}$	0,0313
	300	1530,61	$0,65x10^{-3}$	0,0255
	500	2551,02	$0,39x10^{-3}$	0,0197
	800	4081,63	$0,24x10^{-3}$	0,0156

[1]/ Valores obtidos de acordo com a Eq. (23); [2]/ Valores obtidos de acordo com a Eq. (24).

Analisando-se, primeiramente, somente o efeito do tamanho de população independentemente da estimativa da freqüência de recombinação, verifica-se que o conteúdo médio de informação é maior tanto quanto for maior o tamanho da população, e, como a variância é inversamente relacionada com o conteúdo médio de informação, o seu valor é menor tanto quanto for maior o tamanho da população. O mesmo comportamento é observado para o desvio padrão. Para melhor entendimento, será analisado o caso em que a freqüência de recombinação é 0,05 (ou 5 cM), onde, pode-se verificar que o valor do conteúdo médio é de 1652,89 e 26446,28, o valor da variância é de $6,05x10^{-4}$ e $3,78x10^{-5}$ e o valor de desvio padrão é de 0,0245 (2,45 cM) e 0,0061 (0,61 cM), para as populações com 50 e 800 indivíduos, respectivamente. Portanto, populações de tamanho reduzido proporcionam

estimativas de freqüência de recombinação menos confiáveis do que populações de tamanhos maiores.

Analisando-se somente o efeito da magnitude da estimativa da freqüência de recombinação, nota-se que quanto menor a estimativa da freqüência de recombinação maior será o valor do conteúdo médio de informação e menores os valores de variância e desvio padrão, para um mesmo tamanho de população. Isto pode ser facilmente entendido, quando se analisa, por exemplo, o tamanho de população de 50 indivíduos com estimativas de freqüência de recombinação de 0,05, 0,10 e 0,20, onde, os valores de conteúdo médio de informação são, respectivamente, 1652,89, 694,44, e 255,10, os de variância são $6,050x10^{-4}$, $14,4x10^{-4}$ e $3,92x10^{-3}$, e os de desvio padrão são 0,0245, 0,0379 e 0,0626.

Pode-se concluir, portanto, que quanto mais indivíduos forem genotipados para a construção de mapas genéticos, maior será a confiabilidade dos resultados encontrados. E, também se conclui que quanto maior for a perspectiva de saturação do mapa por marcas, maior será a confiabilidade das estimativas de freqüência de recombinação obtidas, e conseqüentemente, dos mapas construídos a partir destas estimativas.

O valor de LOD escore é função do tamanho da amostra (N) e da freqüência de recombinação (r), conforme pode ser visto na Equação 26 da Revisão de Literatura (item 2.1.7). Utilizando-se uma combinação de vários tamanhos de população e diferentes distâncias entre marcas moleculares, foram gerados os dados apresentados no Quadro 10.

Quadro 10 – Valores de LOD escore para algumas estimativas teóricas de freqüência de recombinação e tamanhos de população de RIL

Freqüência de Recombinação (r)	Tamanho de população	Valores de LOD escore[1]
0,05 (0,0909)[2]	50	8,43
	73	12,31
	100	16,87
	154	25,98
	200	33,74
	300	50,61
	500	84,36
	800	134,98
0,10 (0,1666)	50	5,26
	73	7,69
	100	10,53
	154	16,22
	200	21,07
	300	31,60
	500	52,67
	800	84,28
0,20 (0,2857)	50	2,06
	73	3,00
	100	4,12
	154	6,34
	200	8,24
	300	12,36
	500	20,60
	800	32,96

[1]\ Para a estimativa dos valores de LOD escore considerou-se que $n_1 = n_2$ e $n_3 = n_4$ [(Eq. (27)];

[2]\ Os valores entre parênteses são os valores de r', ou seja, a freqüência de recombinação observada nas RIL. E, os valores de 0,05, 0,10 e 0,20 são de r, ou seja, a freqüência de recombinação na geração F_2, que expressa a distância entre os marcadores [$r' = 2r/(1+2r)$].

Analisando-se, primeiramente, o efeito da variável N (Quadro 10), independentemente da estimativa de r, verifica-se que o valor de LOD escore é maior tanto quanto maior for N. Por exemplo, para um valor de r de 0,05 (ou 5 cM), tem-se para as populações com 50 e 800 indivíduos, valores de LOD escore de 8,43 e 134,98, respectivamente. Por outro lado, analisando-se somente o efeito da magnitude de r, mantendo-se constante o valor de N, verifica-se que quanto menor a estimativa r, maior será o valor do LOD escore. Isto pode ser facilmente entendido quando é analisado, por exemplo, o tamanho de população de 50 indivíduos com estimativas de r de 0,05, 0,10 e 0,20, para as quais os valores de LOD escore são respectivamente, 8,43, 5,26 e 2,06. Pode-se concluir, portanto, que quanto maior for o

número de indivíduos genotipados para a construção de mapas genéticos, maior poderá ser o valor de LOD escore mínimo utilizado para inferência de ligação entre marcadores moleculares, proporcionando uma maior confiabilidade dos mapas construídos. E, também, conclui-se que quanto maior for a perspectiva de saturação do mapa por marcadores, maior poderá ser o limite inferior de LOD escore necessário para inferência de ligação entre marcadores moleculares.

Para populações com 50 indivíduos e estimativa de r de 0,05, o valor de LOD escore está limitado a 8,43 (Quadro 10). Portanto, se no processo de mapeamento é utilizado um valor de LOD escore mínimo superior ao imposto pelo tamanho da população, marcas sabidamente ligadas serão inferidas como não ligadas, levando à formação de um número de grupos de ligação maior do que o real. Neste trabalho, foi utilizado um valor de LOD escore mínimo igual a 3 para inferência sobre a presença de ligação gênica, valor este inferior ao valor de LOD escore imposto pelo tamanho da população, que neste caso é de 8,43. Portanto, nas populações com 50 indivíduos geradas a partir do genoma com saturação de 5 cM, o fator LOD escore não foi um causador da não detecção de ligação gênica onde esta realmente existia.

As populações de tamanho de 100, 154, 200, 300, 500 e 800 indivíduos geradas a partir do genoma com saturação de 5 cM levaram, em todas repetições, à 11 grupos de ligação (Figura 11 e Quadro 7), sem marcas não ligadas (Quadro 8).

4.1.1.2 Populações segregantes simuladas a partir do genoma com saturação de 10 cM

Nas populações segregantes geradas a partir do genoma com nível de saturação de 10 cM, apenas aquelas constituídas por 50 e 100 indivíduos levaram à formação de mais de 11 grupos de ligação (Figura 12 e Quadro 7).

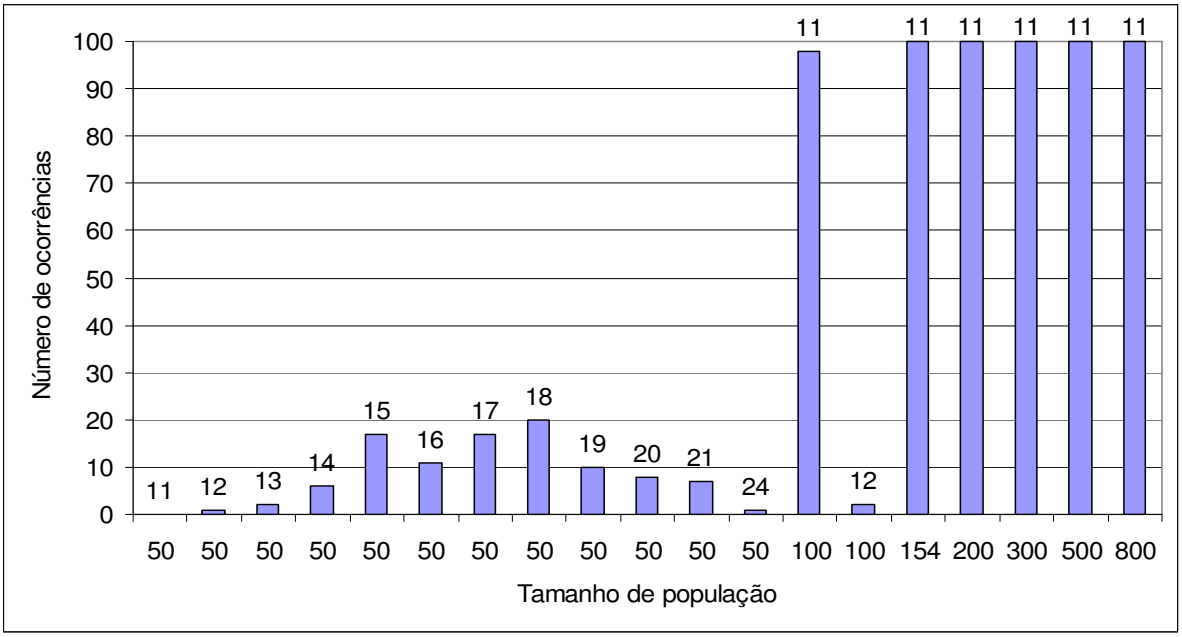


Figura 12 - Distribuição do número de grupos de ligação (indicados no topo de cada barra) obtidos no processo de mapeamento das populações simuladas em função do tamanho de população segregante. Avaliação feita em 100 repetições para cada tamanho de população com 50, 100, 154, 200, 300, 500 e 800 indivíduos, simuladas a partir do genoma com nível de saturação de 10 cM

Para as populações com 50 indivíduos, do total de 100 repetições, não houve nenhuma com formação de 11 grupos de ligação. Sendo, portando, observado em todas as repetições a formação de número de grupos de ligação variando de 12 a 24, não consecutivamente (Figura 12). Por outro lado, para as populações com 100 indivíduos, do total de 100 repetições, 98 formaram 11 grupos de ligação e as outras duas repetições formaram 12 grupos, cada. Além, da formação de maior número de grupos de ligação do que o esperado, também houve, para estes dois tamanhos de população, a presença de marcas não ligadas, embora em pequeno número (Quadro 8). Tanto para as populações com 50 indivíduos como para as com 100 indivíduos, o número de grupo de ligação observado acima do esperado foi devido à divisão de grupos de ligação em conseqüência da não detecção de ligação entre marcas sabidamente ligadas. Analogamente ao apresentado anteriormente para as populações com 50 indivíduos geradas a partir do genoma com saturação de 5 cM, a discussão sobre as causas de não detecção da ligação gênica onde sabidamente ela existia dar-se-á sobre a freqüência de recombinação e o LOD escore.

Sobre a freqüência de recombinação, sabe-se que nas populações segregantes com pequeno número de indivíduos não há uma boa representatividade

da diversidade de gametas produzido pelos parentais, levando a problemas na estimativa da freqüência de recombinação, conforme já discutido anteriormente. Portanto, os tamanhos inadequados de população com 50 e 100 indivíduos proporcionaram a não detecção de ligação genética onde sabidamente ela existia. Sendo, que as populações de 50 indivíduos apresentaram-se muito inferiores para a estimativa de freqüência de recombinação quando comparadas às populações com 100 indivíduos.

O efeito do tamanho da população segregante na confiabilidade das estimativas de freqüência de recombinação pode ser observado por meio da estimativa do conteúdo de informação, da variância ou do desvio padrão da estimativa de freqüência de recombinação (Quadro 9). Para uma estimativa de freqüência de recombinação de 0,10 (ou 10 cM) observa-se uma discrepância dos valores de desvio padrão de 0,0094 (0,94 cM) a 0,0379 (3,79 cM), para os tamanhos de população de 800 a 50 indivíduos, respectivamente. Portanto, populações de tamanho reduzido proporcionam estimativas de freqüência de recombinação menos confiáveis do que populações de tamanhos maiores.

Como apresentado anteriormente, o número de indivíduos utilizados para a construção do mapa genético é um dos fatores limitantes da estimativa de LOD escore (Quadro 10). Para populações com 50 e 100 indivíduos e estimativa de freqüência de recombinação de 0,10, o valor de LOD escore está limitado a 5,26 e 10,53, respectivamente. Neste trabalho utilizou-se no processo de mapeamento um valor de LOD escore mínimo igual a 3 para inferência sobre a presença de ligação gênica, valor este, inferior aos valores de LOD escore de 5,26 e 10,53, impostos pelos tamanhos de população de 50 e 100 indivíduos, respectivamente. Portanto, nas populações com 50 e 100 indivíduos geradas a partir do genoma com saturação de 10 cM o fator LOD escore não foi um causador da não detecção de ligação gênica onde esta realmente existia.

As populações de tamanho de 154, 200, 300, 500 e 800 indivíduos geradas a partir do genoma com saturação de 10 cM apresentaram, em todas as repetições, formação de 11 grupos de ligação (Figura 12 e Quadro 7), sem marcas não ligadas (Quadro 8).

4.1.1.3 Populações segregantes simuladas a partir do genoma com saturação de 20 cM

Nas populações segregantes geradas a partir do genoma com nível de saturação de 20 cM, as constituídas por 50, 100, 154, 200 e 300 indivíduos levaram à formação de mais de 11 grupos de ligação (Figura 13 e Quadro 7).

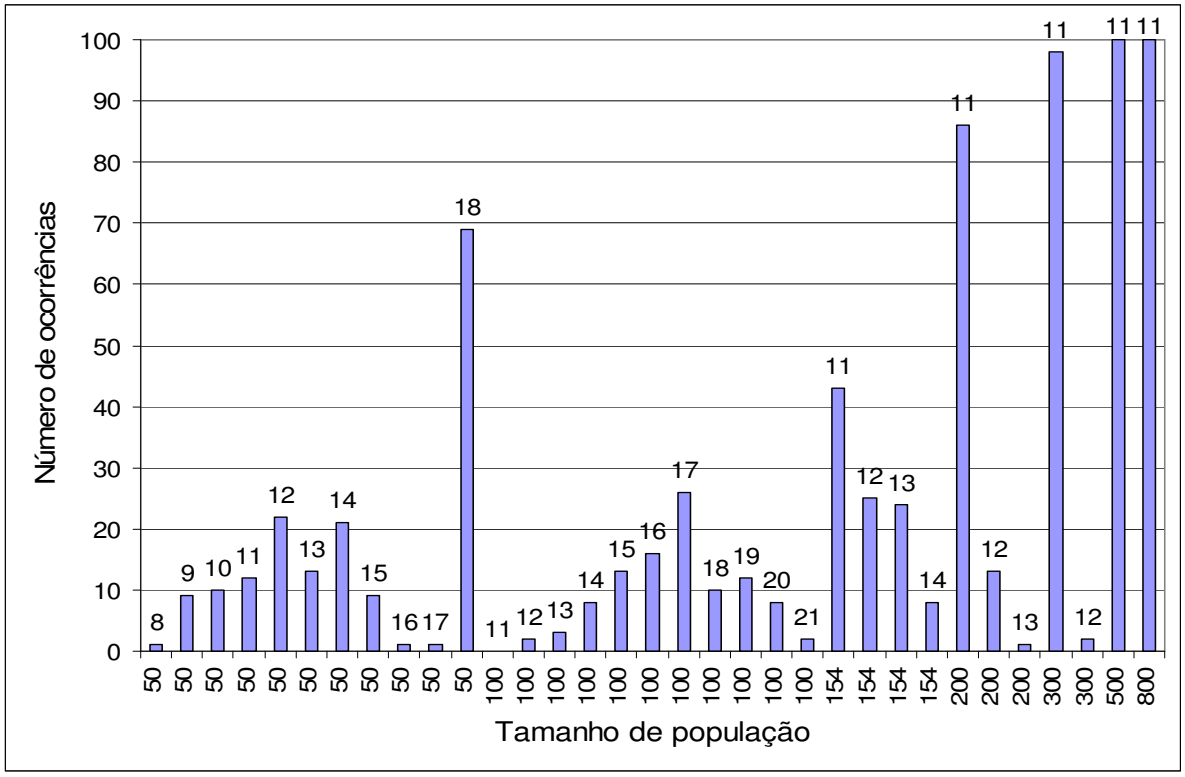


Figura 13 - Distribuição do número de grupos de ligação (indicados no topo de cada barra) obtidos no processo de mapeamento das populações simuladas em função do tamanho de população segregante. Avaliação feita em 100 repetições para cada tamanho de população com 50, 100, 154, 200, 300, 500 e 800 indivíduos, simuladas a partir do genoma com nível de saturação de 20 cM

Para as populações com 50 indivíduos, do total de 100 repetições, houve 12 repetições que levaram à formação de 11 grupos de ligação (Quadro 7), e as outras 88 repetições apresentaram de 8 a 18 grupos de ligação (Figura 13). As 12 repetições em que houve a formação de 11 grupos de ligação não foram utilizadas para as demais análises de comparação de genoma apresentadas neste trabalho, pois, apesar de terem levado à formação de onze grupos de ligação, estes foram constituídos por poucas marcas ligadas, sendo observado que a maioria das marcas apresentou-se não ligadas (Quadro 8).

Para as populações com 100 indivíduos, do total de 100 repetições, nenhuma levou a formação de 11 grupos de ligação. Sendo, portando, observado em todas as repetições a formação de grupos de ligação variando de 12 a 21 (Figura 13).

À primeira vista, poder-se-ia incorrer no erro de afirmar que o tamanho de população de 50 indivíduos seria melhor do que o tamanho de população de 100 indivíduos, pois, para o tamanho de população de 50 indivíduos algumas repetições levaram a formação de 11 grupos de ligação, enquanto que para o tamanho de população de 100 indivíduos não houve nenhuma repetição que levou à formação de 11 grupos de ligação. Além disto, o número máximo de grupos de ligação formado nas populações com 50 indivíduos foi menor do que o número máximo observado nas populações com 100 indivíduos, 18 e 21, respectivamente (Figura 13). Contudo, a aparente vantagem do tamanho de população de 50 indivíduos sobre o tamanho de população de 100 indivíduos é refutada, pois, o número de marcas não ligadas para as populações com 50 indivíduos apresentou-se muito superior ao observado para o tamanho de população de 100 indivíduos (Quadro 8). Nas populações com 50 indivíduos o número mínimo e máximo de marcas não ligadas foi de 24 e 47, respectivamente. Enquanto nas populações com 100 indivíduos o número mínimo e máximo de marcas não ligadas foi de 2 e 19, respectivamente.

Para as populações com 154 indivíduos, do total de 100 repetições, 43 levaram à formação de 11 grupos de ligação e as outras 57 repetições levaram à formação de 12 a 14 grupos de ligação (Quadro 7 e Figura 13).

Para as populações com 300 indivíduos, do total de 100 repetições, 98 levaram a formação de 11 grupos de ligação e as outras duas repetições levaram a formação de 12 grupos de ligação, cada (Quadro 7 e Figura 13).

O número de grupos de ligação observado acima do esperado, para os tamanhos de populações citados anteriormente foi devido à divisão de grupos de ligação em conseqüência da não detecção de ligação entre marcas sabidamente ligadas. Novamente, na discussão, será enfocado o aspecto da freqüência de recombinação e do LOD escore.

Conforme exposto anteriormente, o efeito do tamanho da população segregante na confiabilidade das estimativas de freqüência de recombinação pode ser observado através da estimativa do conteúdo de informação, da variância ou do desvio padrão da estimativa de freqüência de recombinação (Quadro 9). Para uma estimativa de freqüência de recombinação de 0,20 (ou 20 cM) observa-se

discrepância dos valores de desvio padrão de 0,0156 (1,56 cM) a 0,0626 (6,26 cM), para os tamanhos de população de 800 a 50 indivíduos, respectivamente. Portanto, populações de tamanho reduzido proporcionam estimativas de freqüência de recombinação menos confiáveis do que populações de tamanhos maiores.

Conforme já descrito, o número de indivíduos utilizados para a construção de mapas genéticos é um dos fatores limitantes da estimativa de LOD escore. Para populações com 50 indivíduos e estimativa de freqüência de recombinação de 0,20, o valor de LOD escore está limitado a 2,06, 4,12, 6,34, 8,24 e 12,36, para populações com 50, 100, 154, 200 e 300 indivíduos, respectivamente (Quadro 10). Neste trabalho utilizou-se no processo de mapeamento um valor de LOD escore mínimo igual a 3 para inferência sobre a presença de ligação gênica. Para o tamanho de população de 50 indivíduos o valor de LOD escore mínimo igual a 3 foi superior ao LOD dado em função do tamanho da população (LOD escore de 2,06). Portanto, devido ao valor de LOD mínimo ser superior ao valor de LOD dado em função do tamanho de população de 50 indivíduos alguns marcadores sabidamente ligados foram inferidos como não ligados. A inferência de não ligação de marcas quando estas estão verdadeiramente ligadas incorre no denominado erro tipo II, ou seja, não se rejeitar uma hipótese sabidamente falsa. Para as populações de tamanho de 100, 154, 200 e 300, o valor de LOD escore mínimo igual a 3 foi inferior ao LOD dado em função do tamanho da população. Por outro lado, o fator LOD escore não foi um causador da não detecção de ligação gênica onde esta realmente existia nas populações de tamanho iguais a 100, 154, 200 e 300 indivíduos.

Os tamanhos inadequados de população com 50, 100, 154, 200, e 300 indivíduos gerados a partir do genoma com nível de saturação de 20 cM proporcionaram a não detecção de ligação genética onde sabidamente ela existia. Sendo, que as populações de 50 e 100 indivíduos apresentaram-se muito inferiores para a estimativa de freqüência de recombinação quando comparadas às demais populações.

As populações de tamanho de 500 e 800 indivíduos geradas a partir do genoma com saturação de 20 cM apresentaram, em todas as repetições, formação de 11 grupos de ligação (Quadro 7 e Figura 13) e, além disto, não apresentaram, em nenhum caso, marcas não ligadas (Quadro 8).

Toda a discussão até então apresentada foi somente do efeito de tamanho de população segregante na formação dos grupos de ligação mantendo-se constante a

saturação do genoma. A discussão do efeito da saturação do genoma mantendo-se constante o tamanho de população segregante é apresentada a seguir.

4.1.2 Efeito da saturação do genoma na formação dos grupos de ligação

O tamanho de população utilizado para o mapeamento genético é um fator importantíssimo na formação dos grupos de ligação. Mas, o processo de formação de grupos de ligação é também grandemente afetado pelo nível de saturação do genoma por marcadores.

4.1.2.1 Populações com 50 indivíduos

Para as populações de tamanho de 50 indivíduos, obteve-se número de grupos de ligação variando de 11 a 14, 12 a 24 e 8 a 18, para os níveis de saturação de genoma de 5 cM, 10 cM e 20 cM, respectivamente (Figuras 11, 12 e 13). Portanto, para o nível de saturação do genoma de 5 cM foi obtido a menor amplitude de número de grupos de ligação no processo de mapeamento. Além da menor amplitude, o nível de saturação de 5 cM foi o único que proporcionou em 70 repetições, do total de 100 repetições, a formação de 11 grupos de ligação (Quadro 7) e, com a maioria das marcas ligadas (Quadro 8). Para o nível de saturação de 10 cM e tamanho de população de 50 indivíduos nenhuma repetição levou a formação de 11 grupos de ligação (Quadro 7). Para o nível de saturação de 20 cM, como discutido anteriormente, apesar de 12 repetições do total de 100 repetições terem proporcionado a formação de 11 grupos de ligação (Quadro 7), estes apresentaram poucas marcas ligadas, ou seja, a maioria das marcas estava não ligadas (Quadro 8).

4.1.2.2 Populações com 100 indivíduos

Para as populações de tamanho de 100 indivíduos, obteve-se número de grupos de ligação igual a 11, variando de 11 a 12 e 12 a 21, para os níveis de saturação do genoma de 5, 10 e 20 cM, respectivamente (Figuras 11, 12 e 13). O número de repetições, de um total de 100, em que houve a formação de 11 grupos de ligação foi de 100, 98 e 0, para os níveis de saturação do genoma de 5, 10 e 20 cM, respectivamente (Quadro 7). Verifica-se novamente a superioridade de qualidade com a utilização de genomas mais saturados no mapeamento, uma vez que o nível de

saturação do genoma de 5 cM foi o que apresentou melhor resultado em termos de formação de grupos de ligação nas populações com 100 indivíduos, levando à formação de 11 grupos de ligação em todas as repetições (Quadro 7) e, não apresentando nenhuma marca não ligada (Quadro 8). Por outro lado, o nível de saturação do genoma de 20 cM foi o que proporcionou os piores resultados nas populações com 100 indivíduos, não propiciando a formação de 11 grupos de ligação em nenhuma das repetições (Quadro 7) e, apresentando um grande número de marcas não ligadas (Quadro 8).

4.1.2.3 Populações com 154 indivíduos

Para as populações com 154 indivíduos a variação do número de grupos de ligação para o nível de saturação do genoma de 20 cM (66 marcas) foi de 11 a 14 (Figura 13). Além disto, houve, para esta saturação de genoma, a presença de marcas não ligadas em algumas das repetições (Quadro 8). Por outro lado, para os níveis de saturação de 5 (231 marcas) e 10 cM (121 marcas) com não houve variação do número de grupos de ligação nas populações com 154 indivíduos, sendo observado para todas as repetições a formação de 11 grupos de ligação (Figuras 11 e 12 e Quadro 7) e, não houve a presença de marcas não ligadas em nenhuma das repetições (Quadro 8).

Faleiro (2000), utilizou uma população de RIL com 154 indivíduos e 53 locos (10 características de resistência a doenças, o hábito de crescimento e 42 marcadores moleculares) para o mapeamento da espécie *Phaseolus vulgaris* (2n=2x=22). Adotando $LOD_{mín}$ de 4,0 e $r_{máx}$ de 40%, foram mapeados 43 dos 53 locos em nove grupos de ligação perfazendo uma distância de recombinação total de 247,8 cM, com distância média entre locos de 7,3 cM. A obtenção de número de grupos de ligação inferior ao esperado para a espécie (11 grupos de ligação) e a presença de marcas não ligadas provavelmente foi devido ao reduzido número de locos utilizados. Uma vez que, no trabalho de simulação, população com 154 indivíduos levou a recuperação do número de grupos de ligação do genoma simulado quando foram utilizadas 231 e 121 marcas, mas não recuperou corretamente os grupos de ligação quando foram utilizadas 66 marcas, havendo, inclusive marcas não ligadas.

4.1.2.4 Populações com 200 indivíduos

Para as populações com 200 indivíduos o número de grupos de ligação no nível de saturação de genoma de 20 cM foi de 11, 12 ou 13 (Figura 13). Além disto, houve, para esta saturação de genoma, a presença de marcas não ligadas em algumas das repetições (Quadro 8), que embora tenham sido de poucas marcas, evidencia que este nível de saturação e número de indivíduos é ineficiente na reconstrução do genoma verdadeiro. Por outro lado, para os níveis de saturação de 5 e 10 cM e tamanho de população de 200 indivíduos não houve variação do número de grupos de ligação, sendo observado para todas as repetições a formação de 11 grupos de ligação (Figuras 11 e 12), sem marcas não ligadas, em todas as repetições (Quadro 8).

4.1.2.5 Populações com 300 indivíduos

Para as populações com 300 indivíduos o número de grupos de ligação no nível de saturação de genoma de 20 cM foi de 11 e 12 (Figura 13). Além disto, houve, para esta saturação de genoma, a presença de marcas não ligadas em algumas das repetições (Quadro 8), que embora tenham sido de poucas marcas, evidencia que este nível de saturação e número de indivíduos é ineficiente na reconstrução do genoma verdadeiro. Por outro lado, para os níveis de saturação de 5 e 10 cM e tamanho de população de 300 indivíduos não houve variação do número de grupos de ligação, sendo observado para todas as repetições a formação de 11 grupos de ligação (Figuras 11 e 12), sem marcas não ligadas, em todas as repetições (Quadro 8).

4.1.2.6 Populações com 500 e 800 indivíduos

As populações com 500 e 800 indivíduos, nos três níveis de saturação do genoma, 5, 10 e 20 cM, apresentaram em todas repetições formação de 11 grupos de ligação (Figuras 11, 12 e 13 e Quadro 7), sem marcas não ligadas (Quadro 8).

4.1.3 Considerações finais sobre a formação dos grupos de ligação

Uma análise sumarizada pode ser feita observando-se os resultados apresentados no Quadro 7. Nota-se que a utilização de genoma mais saturado levou a melhores resultados em termos de formação de grupos de ligação, como descrito abaixo.

Para as populações geradas a partir do genoma com nível de saturação de 5 cM, para todos os tamanhos de populações estudados neste trabalho, 50, 100, 154, 200, 300, 500 e 800 indivíduos, obteve-se na maioria ou em todas as repetições a formação de 11 grupos de ligação. Sendo, que populações com 100 ou mais indivíduos levaram a formação de 11 grupos de ligação em todas as repetições.

Para as populações geradas a partir do genoma com nível de saturação de 10 cM nem todos os tamanhos de população levaram a formação de 11 grupos de ligação, uma vez que as populações com 50 indivíduos não apresentaram em nenhuma das 100 repetições a formação de 11 grupos de ligação. E, somente, a partir de populações com 154 é que foram obtidos 11 grupos de ligação para todas as repetições.

A pior situação observada foi para as populações geradas a partir do genoma com nível de saturação de 20 cM. No qual, somente a partir de populações com 154 indivíduos é que houve a formação de 11 grupos de ligação, mas não em todas as repetições. Cabe ressaltar novamente que apesar da presença de 12 repetições com 11 grupos de ligação para a população de 50 indivíduos, estes grupos foram totalmente descacterizados pelo excesso de marcas não ligadas. Finalmente, a formação de 11 grupos de ligação em todas as repetições só foi observada nas populações com 500 e 800 indivíduos.

De acordo com o exposto anteriormente, fica evidenciado que o nível de saturação do genoma por marcadores moleculares é um fator importante na determinação da qualidade dos mapas obtidos. Pois, por exemplo, populações com 50 indivíduos levaram à formação de 11 grupos de ligação, somente quando estas populações foram geradas a partir do genoma com nível de saturação de 5 cM. Se trabalhamos com um genoma mais saturado, há uma maior probabilidade de detecção dos eventos de recombinação ocorridos na meiose e, conseqüentemente as estimativas de freqüência de recombinação são mais precisas.

4.2 Número de marcas obtidas em cada grupo de ligação no mapeamento

O número esperado de marcas por grupo de ligação no mapeamento das populações simuladas era de 21, 11 ou 6, para as populações simuladas a partir do genoma com nível de saturação de 5, 10 ou 20 cM, respectivamente, independente do tamanho da população. Contudo, para alguns tamanhos de população foram obtidos grupos de ligação com número de marcas inferior ao esperado, conforme apresentado no Quadro 11 e discutido a seguir.

O número de marcas por grupo de ligação foi analisado apenas nas repetições em que houve a formação de 11 grupos de ligação no mapeamento (vide Quadro 7). Ressalva-se, novamente, que apesar de terem sido formados 11 grupos de ligação em 12 repetições para o tamanho de população com 50 indivíduos e genoma com nível de saturação de 20 cM, estas repetições não foram utilizadas nas análises apresentadas a seguir em função do pequeno número de marcas ligadas nos grupos formados.

Repetições com presença de marcas não ligadas para pelo menos um dos 11 grupos de ligação foram obtidas apenas para as populações de tamanho de 50 e 100 indivíduos, nos níveis de saturação de genoma de 5 e 10 cM, respectivamente. Para o nível de saturação de 20 cM foram obtidas repetições com marcas não ligadas para os tamanhos de população de 154, 200 e 300 indivíduos (Quadro 11).

Nas populações com 50 indivíduos e nível de saturação do genoma de 5 cM, somente o grupo de ligação 9 apresentou repetição com menos de 21 marcas. Para as populações com 100 indivíduos e nível de saturação do genoma de 10 cM, somente os grupos de ligação 3, 5 e 8 apresentaram repetição com menos de 11 marcas. Para as populações com 154 indivíduos e nível de saturação do genoma de 20 cM, todos os grupos de ligação apresentaram repetições com menos de 6 marcas. Para este tamanho de população, o número máximo de repetições com menos de 6 marcas foi de 7, obtido para o grupo de ligação 10. Nas populações com 200 indivíduos e nível de saturação do genoma de 20 cM, os grupos de ligação 5, 6, 7 e 11 não apresentaram repetições com número de marcas inferior a 6. Para as populações com 300 indivíduos e nível de saturação do genoma de 20 cM, apenas os grupos de ligação 3 e 11 apresentaram repetição com menos de 6 marcas (Quadro 11).

O número de marcas não ligadas nos casos acima descritos foram apenas de uma ou duas marcas, sendo que o caso de duas marcas não ligadas só foi obtido para o grupo de ligação 4 nas populações com 154 indivíduos derivada do genoma

com nível de saturação de 20 cM (Quadro 12). Ressalta-se que o exposto no Quadro 8 é diferente do apresentado no Quadro 12, uma vez que na primeira foram incluídas todas as 100 repetições, independente do número de grupos de ligação formado. Na segunda, foram analisadas somente as repetições com formação de 11 grupos de ligação.

A formação de 11 grupos de ligação e sem a presença de marcas não ligadas em todas as repetições poderia servir como um indicativo do tamanho mínimo de população na qual os mapas obtidos seriam representativos do genoma verdadeiro. Tomando essa premissa como verdadeira, para o nível de saturação de 5 cM do genoma utilizado para geração de populações, o número mínimo de indivíduos para o mapeamento genético seria de 100 indivíduos, de acordo com o obtido neste trabalho de simulação. Para o nível de saturação de 10 cM do genoma, o número mínimo seria de 154 indivíduos na população segregante. Para o nível de saturação de 20 cM do genoma utilizado para geração de populações, o número mínimo de indivíduos para o mapeamento genético seria de 500 indivíduos na população segregante.

Quadro 11 - Número de ocorrências em que o total de marcas em cada grupo de ligação foi igual a 21, 11 e 6, para as saturações de genoma de 5 cM, 10 cM e 20 cM, respectivamente. Avaliação feita apenas nas repetições em que houve a formação de 11 grupos de ligação no mapeamento das populações simuladas, conforme apresentado no Quadro 7

Saturação do genoma	Tamanho de população	NRA[1]	Grupo de ligação										
			1	2	3	4	5	6	7	8	9	10	11
5 cM	50	70	70 (0)[2]	70 (0)	70 (0)	70 (0)	70 (0)	70 (0)	70 (0)	70 (0)	69 (1)	70 (0)	70 (0)
	100	100	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)
	154	100	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)
	200	100	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)
	300	100	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)
	500	100	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)
	800	100	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)
10 cM	100	98	98 (0)	98 (0)	97 (1)	98 (0)	97 (1)	98 (0)	98 (0)	97 (1)	98 (0)	98 (0)	98 (0)
	154	100	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)
	200	100	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)
	300	100	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)
	500	100	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)
	800	100	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)
20 cM	154	43	40 (3)	41 (2)	41 (2)	37 (6)	40 (3)	41 (2)	42 (1)	42 (1)	42 (1)	36 (7)	40 (3)
	200	86	85 (1)	85 (1)	85 (1)	85 (1)	86 (0)	86 (0)	86 (0)	85 (1)	85 (1)	85 (1)	86 (0)
	300	98	98 (0)	98 (0)	97 (1)	98 (0)	98 (0)	98 (0)	98 (0)	98 (0)	98 (0)	98 (0)	97 (1)
	500	100	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)
	800	100	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)

[1] NRA = número de repetições avaliadas, ou seja, aquelas com formação de 11 grupos de ligação no mapeamento;
[2] O número de ocorrências em que o total de marcas em cada grupo de ligação foi diferente de 21, 11 e 6, respectivamente, para os níveis de saturação de 5, 10 e 20 cM, está entre parênteses.

Quadro 12 - Número máximo de marcas não ligadas por grupo de ligação nas populações simuladas em que houve a formação de onze grupos de ligação. Avaliação feita somente para os grupos de ligações em que houve alguma repetição com marcas não ligadas, conforme apresentado no Quadro 11

Saturação do genoma	Tamanho de população	N.R.	Grupo de ligação										
			1	2	3	4	5	6	7	8	9	10	11
5 cM	50	70	-	-	-	-	-	-	-	-	1	-	-
10 cM	100	98	-	-	1	-	1	-	-	1	-	-	-
20 cM	154	43	1	1	1	**2**	1	1	1	1	1	1	1
	200	86	1	1	1	1	-	-	-	1	1	1	-
	300	98	-	-	1	-	-	-	-	-	-	-	1

Nota:

a) Os traços (-) no quadro acima indicam os grupos de ligação em que todas as repetições analisadas apresentaram o número de marcas esperadas, ou seja, 21, 11 e 6 marcas por grupo de ligação nos genomas com saturação de 5, 10 e 20 cM, respectivamente.

b) A correta análise deste quadro faz-se da seguinte maneira: por exemplo, volte ao Quadro 11 e observe no nível de saturação do genoma de 20 cM o tamanho de população de 154 indivíduos. Verifica-se que todos os grupos de ligação apresentam repetições com número de marcas inferior ao esperado (6 marcas), e que para o grupo de ligação 4 houve seis repetições. Para saber qual foi o número máximo de marcas não ligadas nestas repetições recorre-se ao Quadro 12, verificando-se que para o grupo de ligação 4 das seis repetições que apresentaram marcas não ligadas, a repetição que apresentou maior número de marcas não ligadas foi de 2 marcas.

4.3 Correlação de Spearman

A correlação de Spearman foi utilizada para identificação de inversão da posição de marcas dentro de cada um dos 11 grupos de ligação formados no mapeamento. Esta avaliação só foi realizada nas repetições em que foram formados 11 grupos de ligação (vide Quadro 7).

Valores de correlação de Spearman iguais à unidade indicam que a ordem das marcas no grupo de ligação obtido no mapeamento das populações segregantes não foi alterada em relação à ordem previamente conhecida no genoma o qual foi utilizado para a geração das populações. Por outro lado, se os valores de correlação de Spearman são diferentes da unidade, ou seja, menores do que 1, indicam que a ordem das marcas no grupo de ligação obtido no mapeamento das populações segregantes foi alterada em relação a ordem previamente conhecida no genoma o qual foi utilizado para a geração das populações.

Repetições com presença de valores de correlação de Spearman diferente da unidade para alguns dos 11 grupos de ligação foram obtidas para as populações segregantes nos três níveis de saturação de genoma estudados (Quadro 13).

Para o nível de saturação de 5 cM, os tamanhos de população de 50, 100 e 154 indivíduos apresentaram repetições com valores de correlação de Spearman diferente da unidade. Sendo, que a presença de marcas invertidas no grupo de ligação foi maior nas populações de menor tamanho, uma vez que para as populações de tamanho de 50 indivíduos todos os grupos de ligação apresentaram repetições com inversão de ordem de marcas, enquanto que para o tamanho de população de 100 indivíduos apenas os grupos de ligação 1, 4, 5, 10 e 11 apresentaram repetições com inversão de ordem de marcas. E, finalmente, para o tamanho de população de 154 indivíduos somente o grupo de ligação 6 apresentou uma repetição com inversão de ordem de marcas.

Para o nível de saturação do genoma de 10 cM, apenas os tamanhos de população com 100 e 154 indivíduos apresentaram repetições com inversão na ordem de marcas. Sendo, também, observado maior freqüência de inversão de marcas nas populações com menor número de indivíduos, uma vez que para o tamanho de população de 100 indivíduos os grupos de ligação 1, 2, 8, 10 e 11 apresentaram repetições com marcas invertidas. E, para o tamanho de população de 154 indivíduos,

somente o grupo de ligação 2 apresentou uma repetição com inversão de ordem de marcas.

Resultado inesperado foi observado para as populações geradas a partir do genoma com saturação de 20 cM, nas quais foi observada inversão de ordem de marcas apenas para o tamanho de população de 200 indivíduos nos grupos de ligação 5, 6 e 9. A não ocorrência de inversão para o tamanho de população de 154 indivíduos e a ocorrência para o de 200 indivíduos pode ser devida à menor quantidade de repetições avaliadas no primeiro tamanho de população (43 repetições) em relação ao segundo tamanho (86 repetições) (Quadro 13).

Numa avaliação rápida, sem muitos critérios, poder-se-ia incorrer no erro de afirmar que o tamanho de população de 154 indivíduos, derivada do genoma com nível de saturação de 20 cM, seria mais adequada para mapeamento do que as populações com 154 indivíduos geradas a partir dos genomas com níveis de saturação de 5 e 10 cM. Uma vez que as populações de 154 indivíduos geradas a partir do genoma com nível de saturação de 20 cM não apresentaram nenhuma inversão de ordem de marcas nas repetições avaliadas, enquanto que para as populações com 154 indivíduos geradas a partir dos genomas com níveis de saturação de 5 e 10 cM houve uma repetição com inversão de ordem de marcas, nos grupos de ligação 6 e 2, respectivamente (Quadro 13). Mas, a avaliação da qualidade do tamanho de população em formar bons mapas não deve ser baseada apenas na correlação de Spearman. Na avaliação do número de repetições com presença de marcas não ligadas, as populações com 154 indivíduos geradas a partir do genoma com nível de saturação de 20 cM foram as que apresentaram maior número de repetições com marcas não ligadas (Quadro 11), portanto com esta avaliação não se cometeria o erro em afirmar que o tamanho de população de 154 indivíduos derivado do genoma menos saturado (20 cM) seria melhor do que as populações de 154 indivíduos derivadas dos genomas mais saturados (5 e 10 cM).

A inversão de ordem de marcas dentro de grupos de ligação foi maior tanto quanto menores foram os tamanhos de populações utilizadas no mapeamento. Portanto, a utilização de populações de tamanhos inadequados nos trabalhos de mapeamento pode levar a graves problemas relacionados a posicionamento de marcas nos grupos de ligação encontrados, podendo por exemplo, levar a resultados equivocados nos processos de mapeamento de QTLs e seleção assistida por marcadores moleculares.

Quadro 13 - Número de repetições em que a correlação de Spearman em cada grupo de ligação foi igual à unidade. Avaliação feita apenas nas repetições em que houve a formação de 11 grupos de ligação no mapeamento das populações simuladas, conforme apresentado no Quadro 7

Saturação do genoma	Tamanho de população	NRA[1]	Grupo de ligação										
			1	2	3	4	5	6	7	8	9	10	11
5 cM	50	70	58 (12)[2]	51 (19)	60 (10)	53 (17)	52 (18)	50 (20)	50 (20)	45 (25)	61 (9)	48 (22)	57 (13)
	100	100	98 (2)	100 (0)	100 (0)	98 (2)	99 (1)	100 (0)	100 (0)	100 (0)	100 (0)	99 (1)	98 (2)
	154	100	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	99 (1)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)
	200	100	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)
	300	100	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)
	500	100	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)
	800	100	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)
10 cM	100	98	97 (1)	97 (1)	98 (0)	98 (0)	98 (0)	98 (0)	98 (0)	97 (1)	98 (0)	97 (1)	96 (2)
	154	100	100 (0)	99 (1)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)
	200	100	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)
	300	100	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)
	500	100	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)
	800	100	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)
20 cM	154	43	43 (0)	43 (0)	43 (0)	43 (0)	43 (0)	43 (0)	43 (0)	43 (0)	43 (0)	43 (0)	43 (0)
	200	86	86 (0)	86 (0)	86 (0)	86 (0)	85 (1)	85 (1)	86 (0)	86 (0)	85 (1)	86 (0)	86 (0)
	300	98	98 (0)	98 (0)	98 (0)	98 (0)	98 (0)	98 (0)	98 (0)	98 (0)	98 (0)	98 (0)	98 (0)
	500	100	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)
	800	100	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)

[1]\ NRA = número de repetições avaliadas, ou seja, aquelas com formação de 11 grupos de ligação no mapeamento;
[2] \ Número de repetições em que a correlação de Spearman foi diferente de 1 está entre parênteses.

4.4 Tamanho dos grupos de ligação

O tamanho de cada grupo de ligação foi obtido pela média aritmética dos tamanhos dos grupos de ligação obtidos nas repetições com formação de 11 grupos de ligação. Os valores de tamanho médio dos grupos de ligação e seus desvios padrão obtidos para os vários tamanhos de população segregantes utilizadas no mapeamento genético estão apresentados no Quadro 14. O tamanho de cada grupo de ligação esperado após o mapeamento das populações segregantes era de 100 cM, pois, este era o tamanho de cada grupo de ligação nos três níveis de saturação do genoma utilizados para geração das populações segregantes.

Analisando-se o efeito do tamanho de população dentro do nível de saturação do genoma de 5 cM verifica-se a aproximação do tamanho médio dos grupos ao valor esperado de 100 cM de ligação à medida que aumenta-se o tamanho das populações segregantes de 50 para 800 indivíduos. Por exemplo, para o grupo de ligação 1, tem-se que o tamanho médio do grupo de ligação foi de 105,3 e 102,8 cM, para os tamanhos de população de 50 e 800 indivíduos, respectivamente (Quadro 14). Exceção a este comportamento foi observado apenas para o grupo ligação 6, onde o tamanho médio do grupo de ligação foi de 102,1 e 102,9 cM, para os tamanhos de população de 50 e 800 indivíduos, respectivamente. Uma explicação para aproximação do tamanho médio dos grupos de ligação ao valor esperado de 100 cM à medida que se aumenta o tamanho da população segregante é a redução da amplitude de variação dos tamanhos de grupos de ligação obtidos entre repetições. Esta redução da amplitude de variação dos tamanhos de grupos ligação em torno da média pode ser observada pela redução do desvio padrão com o aumento do número de indivíduos nas populações segregantes. Por exemplo, para o grupo de ligação 1 tem-se que o desvio padrão foi de 11,7 e 2,9 cM, para os tamanhos de população de 50 e 800 indivíduos, respectivamente. Além, desta explicação em termos estatísticos, há uma explicação biológica dada pela melhoria das estimativas das freqüências de recombinação à medida que se aumenta o tamanho das populações segregantes utilizadas no mapeamento, conseqüentemente melhorando as estimativas de freqüência de recombinação, como já discutido em tópicos anteriores.

Quadro 14 - Tamanho médio de grupos de ligação e seu desvio padrão em função do tamanho de população e nível de saturação do genoma. Avaliação feita apenas nas repetições em que houve a formação de 11 grupos de ligação no mapeamento das populações simuladas, conforme apresentado no Quadro 7

S.G.	T.P.	N.R.	Grupo de ligação											Média [2]
			1	2	3	4	5	6	7	8	9	10	11	
5 cM	50	70	105,3 a (11,7)[1]	107,8 a (13,2)	105,6 a (10,4)	104,0 a (12,4)	105,3 a (11,8)	102,1 a (12,6)	104,7 a (10,4)	107,5 a (12,9)	105,4 a (10,4)	104,0 a (12,9)	106,3 a (12,5)	105,3 a (12,0)
	100	100	103,9 a (9,0)	103,7 b (7,9)	102,2 b (7,8)	104,0 a (8,5)	103,4 a (7,7)	102,0 a (9,1)	102,4 a (8,4)	102,8 b (7,1)	104,4 a (7,5)	104,9 a (7,8)	102,5 b (7,4)	104,2 b (8,1)
	154	100	103,7 a (6,3)	104,3 ab (7,3)	105,3 ab (6,40)	102,8 a (5,7)	104,8 a (7,2)	105,1 a (6,7)	104,1 a (5,9)	104,7 ab (6,4)	104,1 a (6,7)	102,9 a (6,8)	104,0 ab (6,1)	103,3 bc (6,5)
	200	100	103,0 a (5,8)	102,3 b (5,7)	102,8 ab (5,5)	102,6 a (5,7)	102,2 a (5,0)	103,4 a (5,9)	102,5 a (6,0)	103,3 b (6,3)	102,5 a (5,9)	103,6 a (6,0)	103,0 ab (5,4)	103,2 c (5,7)
	300	100	102,5 a (3,7)	102,2 b (4,5)	103,2 ab (5,0)	102,0 a (3,9)	103,4 a (4,6)	103,2 a (5,0)	102,7 a (4,4)	102,9 b (4,3)	102,8 a (5,0)	102,9 a (4,9)	103,6 ab (4,5)	102,9 c (4,6)
	500	100	103,1 a (3,2)	103,2 b (3,4)	102,8 ab (3,2)	103,6 a (3,2)	102,9 a (3,5)	103,8 a (3,0)	103,2 a (3,7)	103,2 b (3,2)	102,9 a (3,6)	103,6 a (3,0)	102,4 b (3,8)	102,8 c (3,4)
	800	100	102,8 a (2,9)	102,6 b (2,5)	102,4 ab (2,6)	102,9 a (2,8)	102,8 a (2,7)	102,9 a (3,0)	103,2 a (2,9)	103,2 b (2,9)	102,6 a (2,5)	103,0 a (2,6)	102,3 b (2,7)	102,8 c (2,7)
10 cM	100	98	104,3 a (9,1)	105,1 a (9,4)	103,7 a (8,9)	105,0 a (9,0)	105,0 a (8,5)	105,6 a (8,2)	103,2 a (8,2)	103,6 a (9,0)	104,6 a (7,9)	104,0 a (9,1)	104,8 a (9,1)	104,4 a (8,8)
	154	100	103,2 a (7,2)	103,2 a (7,4)	103,2 a (7,2)	103,8 a (6,6)	103,2 a (7,2)	102,9 b (6,4)	104,2 a (7,2)	103,8 a (8,0)	104,8 a (7,1)	104,4 a (7,5)	103,5 a (6,5)	103,6 ab (7,1)
	200	100	103,0 a (6,5)	103,1 a (7,1)	103,3 a (6,5)	102,7 a (5,6)	104,0 a (6,6)	102,9 b (5,2)	102,4 a (5,9)	102,1 a (5,6)	104,7 a (5,9)	103,0 a (6,9)	103,2 a (6,0)	103,1 bc (6,2)
	300	100	102,4 a (5,4)	103,6 a (5,3)	103,1 a (5,0)	103,2 a (4,7)	102,9 a (5,4)	103,9 ab (4,9)	103,2 a (5,0)	102,7 a (4,5)	102,8 a (4,8)	102,7 a (4,6)	103,0 a (4,8)	103,1 bc (4,9)
	500	100	102,8 a (4,1)	103,0 a (4,0)	102,8 a (3,7)	103,1 a (4,0)	102,9 a (4,2)	103,0 ab (3,9)	103,1 a (4,1)	102,7 a (3,9)	103,1 a (3,8)	102,8 a (3,7)	102,7 a (3,8)	102,9 bc (3,9)
	800	100	102,3 a (3,0)	102,7 a (2,7)	103,4 a (2,9)	102,4 a (2,9)	102,6 a (3,0)	103,1 ab (3,2)	103,1 a (3,1)	103,2 a (3,0)	102,5 a (3,2)	103,0 a (3,2)	102,2 a (2,8)	102,8 c (3,0)
20 cM	154	43	98,8 a (10,4)	99,6 ab (8,0)	97,9 b (6,9)	97,5 b (11,5)	98,7 a (8,6)	100,2 a (6,7)	99,3 a (7,7)	98,4 a (8,3)	98,6 a (6,8)	96,9 b (10,1)	97,4 b (8,6)	98,5 b (8,6)
	200	86	100,9 a (7,5)	98,4 b (7,2)	101,8 a (6,8)	100,2 ab (8,1)	100,6 a (7,1)	101,9 a (7,1)	101,4 a (6,9)	100,9 a (7,6)	101,2 a (7,8)	101,2 a (7,3)	100,2 ab (6,6)	100,8 a (7,3)
	300	98	100,5 a (4,7)	101,4 a (6,1)	101,0 ab (6,2)	101,1 ab (6,0)	101,1 a (5,7)	100,6 a (5,7)	102,3 a (6,0)	100,5 a (5,49)	101,3 a (6,3)	101,5 a (5,8)	101,0 a (6,5)	101,1 a (5,9)
	500	100	101,2 a (3,9)	100,7 ab (3,9)	101,0 ab (4,2)	101,6 a (4,5)	101,0 a (4,1)	100,8 a (4,8)	101,0 a (4,1)	100,6 a (4,9)	101,3 a (4,2)	100,8 a (4,3)	100,2 ab (4,3)	100,9 a (4,3)
	800	100	100,4 a (3,6)	100,9 ab (3,4)	101,2 ab (3,7)	100,4 ab (3,77)	100,5 a (3,4)	100,3 a (3,4)	101,2 a (3,6)	100,4 a (3,3)	101,0 a (3,2)	100,4 ab (3,7)	101,1 a (3,6)	100,7 a (3,5)

[1]\ Entre parênteses estão os valores de desvio padrão; [2]\ Média dos onze grupos de ligação;
S.G.= Saturação do genoma; T.P.= Tamanho da população; N.R.= número de repetições avaliadas;
Médias seguidas por letras iguais nas colunas, dentro de cada nível de saturação do genoma, não diferem entre si pelo teste de Tukey a 1% de probabilidade (P<0,01).

Uma melhor visualização do comportamento do tamanho médio dos grupos de ligação e do desvio padrão com o aumento do número de indivíduos no nível de saturação do genoma de 5 cM pode ser observado nas Figuras 14 e 15, que apesar de representarem apenas o comportamento observado para o grupo de ligação 1, é representativo do ocorrido para todos os grupos de ligação analisados dentro do nível de saturação do genoma de 5 cM. Fica evidente o efeito da redução da amplitude de variação dos tamanhos de grupos de ligação com o aumento do tamanho das populações segregantes. Esta redução também pode ser visualizada na Figura 15 pela análise do comportamento dos valores de desvio padrão apresentado nos topos de cada barra.

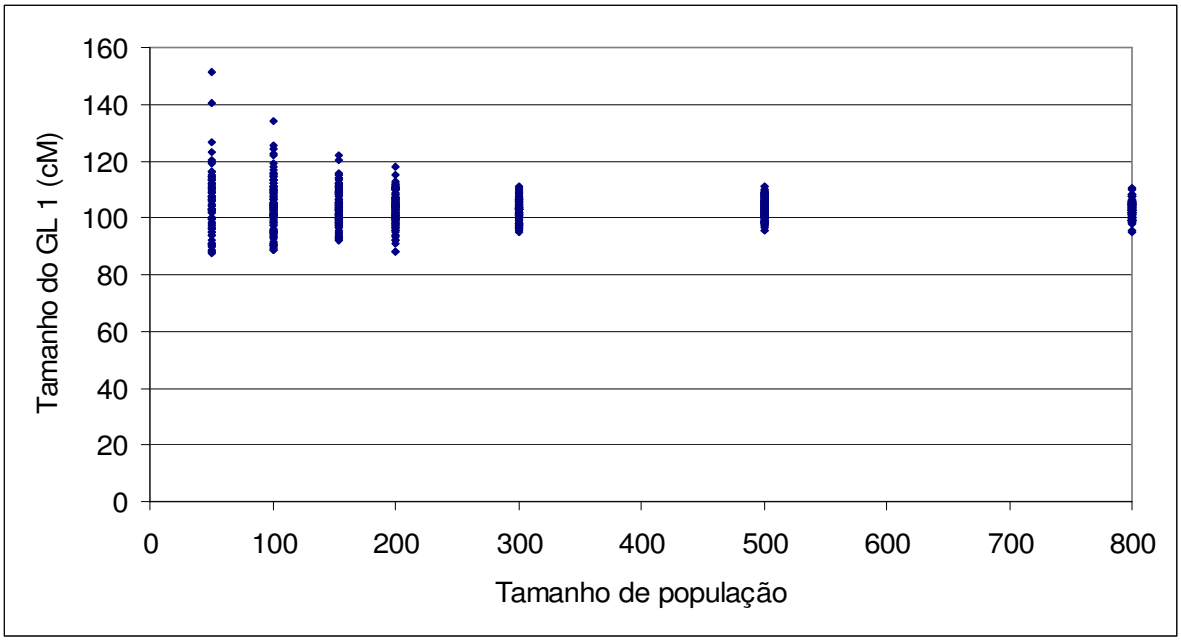


Figura 14 - Dispersão dos valores de tamanho do grupo de ligação 1 em função do tamanho de população segregante. Avaliação feita nas populações com 50, 100, 154, 200, 300, 500 e 800 indivíduos, simuladas a partir do genoma com nível de saturação de 5 cM.

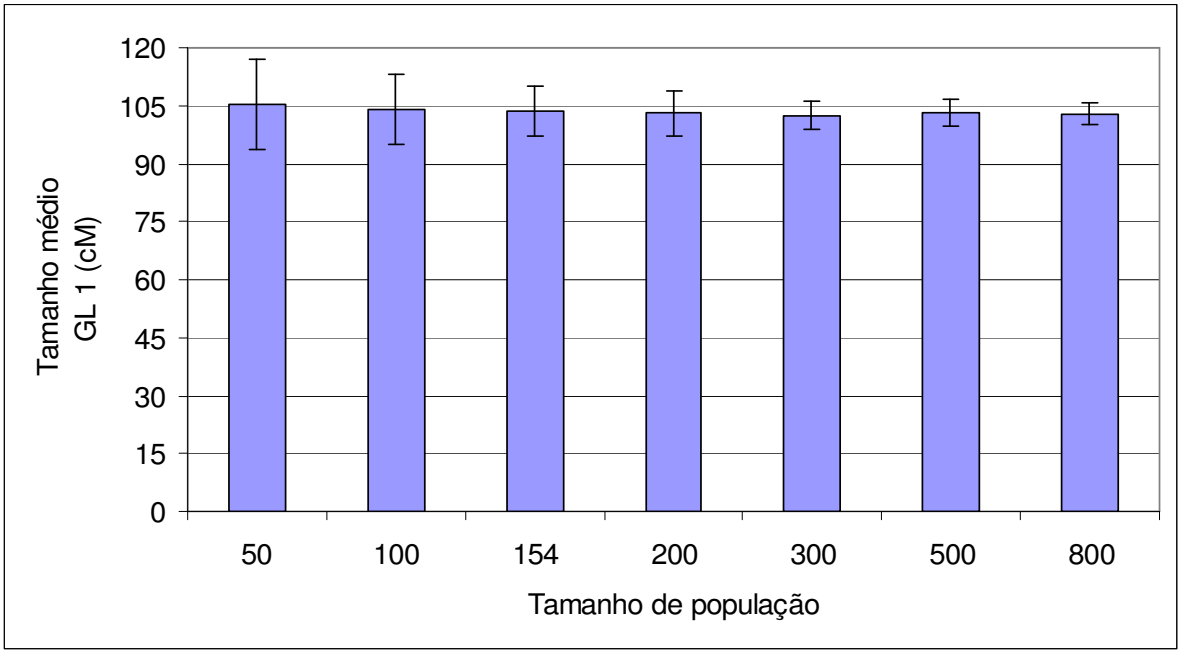


Figura 15 - Tamanho médio do grupo de ligação 1 e seu desvio padrão (linha vertical no topo de cada barra) em função do tamanho de população segregante. Avaliação feita nas populações com 50, 100, 154, 200, 300, 500 e 800 indivíduos, simuladas a partir do genoma com nível de saturação de 5 cM.

A realização do teste de Tukey (P< 0,01) permite a avaliação estatística dos efeitos de tamanho de população no tamanho dos grupos de ligação formados no mapeamento das populações simuladas. Para os grupos de ligação 1, 4, 5, 6, 7, 9 e 10 não houve diferença significativa entre médias para os diferentes tamanhos de população no nível de saturação do genoma de 5 cM. A análise utilizando-se a média geral (última coluna do Quadro 14) permite concluir que à medida que aumenta-se o tamanho de população há ao o tamanho do grupo de ligação aproxima-se do valor esperado de 100 cM. Sendo, que a partir do tamanho de população de 154 indivíduos não há mais diferenças significativas da médias do tamanho dos grupos de ligação.

Para as populações geradas a partir do genoma com nível de saturação de 10 cM houve, assim como para as populações geradas a partir do genoma de 5 cM, uma aproximação do tamanho médio dos grupos de ligação ao valor esperado de 100 cM à medida que aumentou-se o tamanho das populações segregantes de 100 para 800 indivíduos. Por exemplo, para o grupo de ligação 1, tem-se que o tamanho médio do grupo de ligação foi de 104,3 e 102,3 cM, para os tamanhos de população de 100 e 800 indivíduos, respectivamente (Quadro 14). Também, os valores de desvio padrão

diminuíram com o aumento do tamanho das populações. Por exemplo, para o grupo de ligação 1 tem-se que o desvio padrão foi de 9,1 e 3,0 cM, para os tamanhos de população de 100 e 800 indivíduos, respectivamente. As explicações para a aproximação do tamanho médio dos grupos de ligação ao valor esperado e a redução do desvio padrão são as mesmas apresentadas para as populações geradas a partir do genoma com nível de saturação de 5 cM. Uma melhor visualização do comportamento do tamanho médio dos grupos de ligação e desvio padrão com o aumento do número de indivíduos pode ser observado nas Figuras 16 e 17, que apesar de representarem apenas o comportamento observado para o grupo de ligação 1, é representativo do ocorrido para todos os grupos de ligação analisados dentro do nível de saturação do genoma de 10 cM.

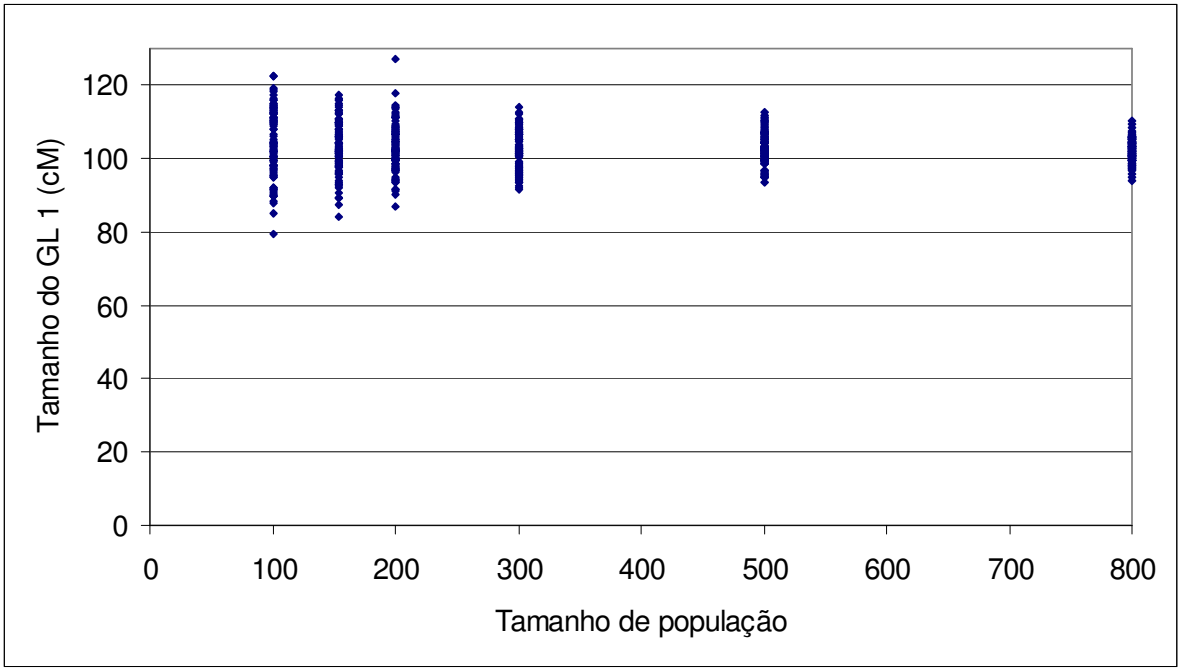


Figura 16 - Dispersão dos valores de tamanho do grupo de ligação 1 em função do número de indivíduos na população segregante. Avaliação feita nas populações com 100, 154, 200, 300, 500 e 800 indivíduos, simuladas a partir do genoma com nível de saturação de 10 cM.

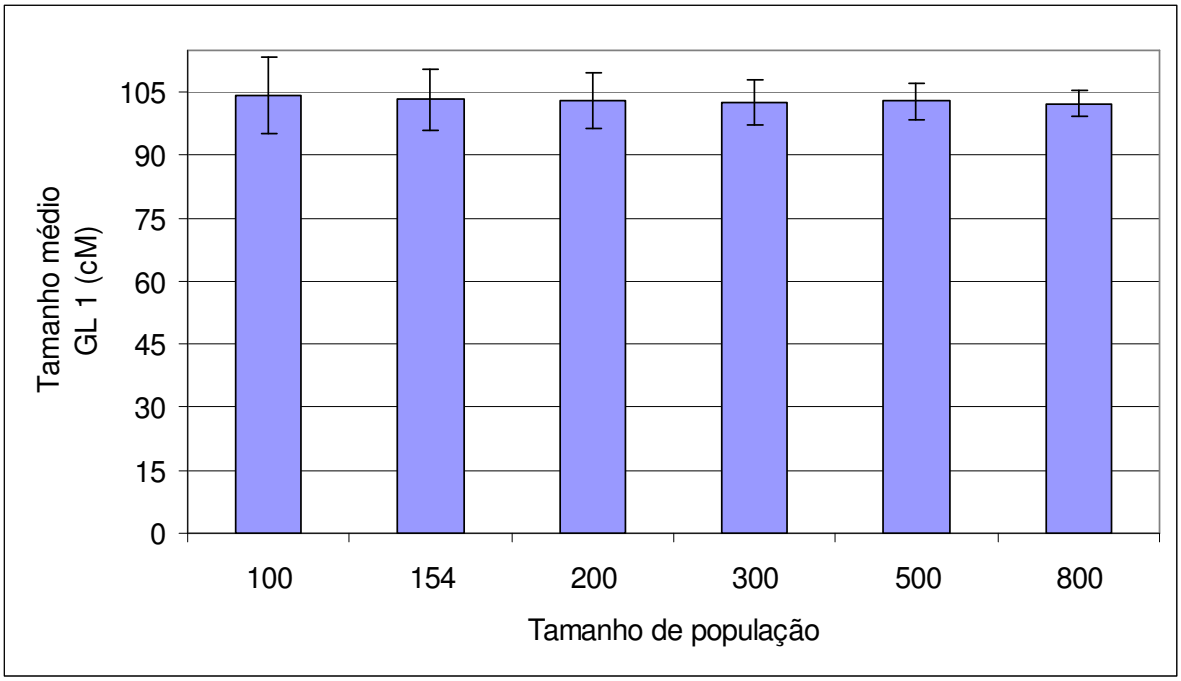


Figura 17 - Tamanho médio do grupo de ligação 1 e seu desvio padrão (linha vertical no topo de cada barra) em função do tamanho de população segregante. Avaliação feita nas populações com 100, 154, 200, 300, 500 e 800 indivíduos, simuladas a partir do genoma com nível de saturação de 10 cM.

No nível de saturação do genoma de 10 cM é possível concluir pelo teste de média que apenas para o grupo de ligação 6 houve diferença estatisticamente significativa entre médias dos diferentes tamanhos de populações. A análise da média geral (última coluna do Quadro 14), através do teste de média, mostra a aproximação do tamanho do grupo de ligação ao valor esperado com o aumento do tamanho de população.

Nas populações geradas a partir do genoma com nível de saturação de 20 cM, o tamanho médio de grupos de ligação apresentou um comportamento um pouco distinto daquele obtido nas populações geradas a partir dos genomas com níveis de saturação de 5 e 10 cM. Para o menor tamanho de população (154 indivíduos) avaliado no genoma de saturação de 20 cM foram obtidas menores médias de tamanhos de grupos de ligação em relação ao obtido para as populações com 800 indivíduos, em todos os grupos de ligação (Quadro 14). E, como visto anteriormente, os maiores tamanhos de grupos de ligação para os níveis de saturação de 5 e 10 cM foram obtidos para as menores populações. Por outro lado, o comportamento do desvio padrão não se apresentou diferente do observado para as populações de 5 e

10 cM, ou seja, o seu valor decresceu à media que aumentou-se o número de indivíduos nas populações segregantes, por exemplo, para o grupo de ligação 1, o valor de desvio padrão foi de 10,4 e 3,6 cM, para os tamanhos de população de 154 e 800 indivíduos, respectivamente (Quadro 14). Uma melhor visualização do comportamento do tamanho médio dos grupos de ligação e desvio padrão com o aumento do número de indivíduos pode ser observado nas Figuras 18 e 19, que apesar de representarem apenas o comportamento observado para o grupo de ligação 1, é representativo do ocorrido para todos os grupos de ligação analisados dentro do nível de saturação do genoma de 20 cM.

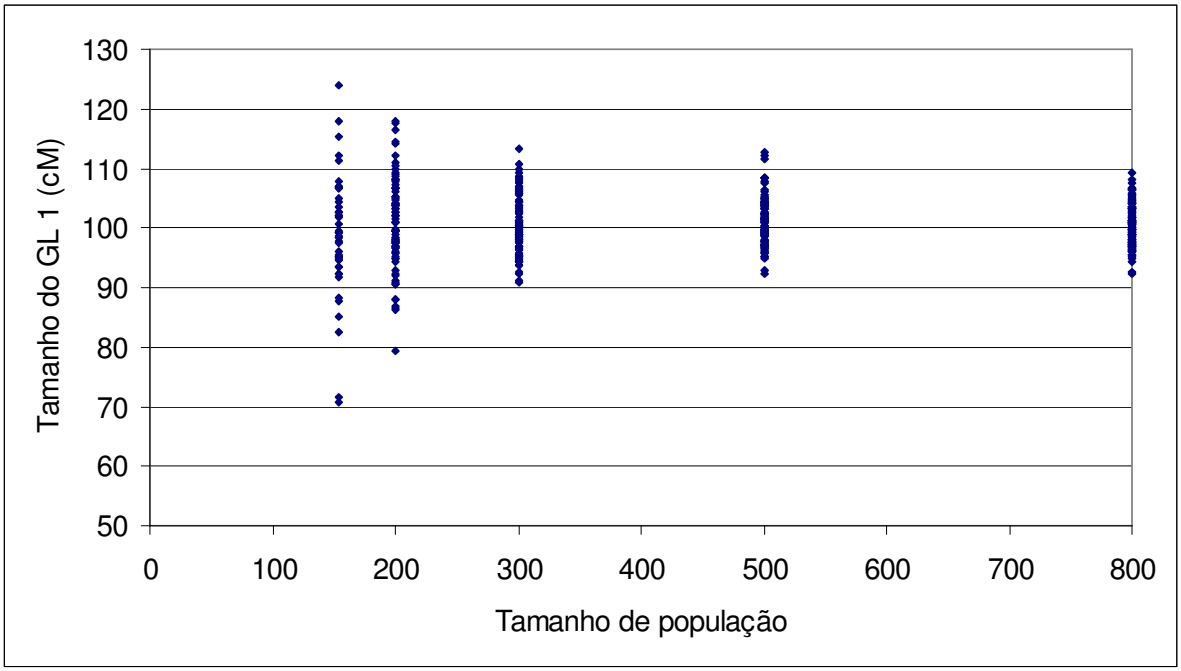


Figura 18 - Dispersão dos valores de tamanho do grupo de ligação 1 em função do número de indivíduos na população segregante. Avaliação feita nas populações com 154, 200, 300, 500 e 800 indivíduos, simuladas a partir do genoma com nível de saturação de 20 cM.

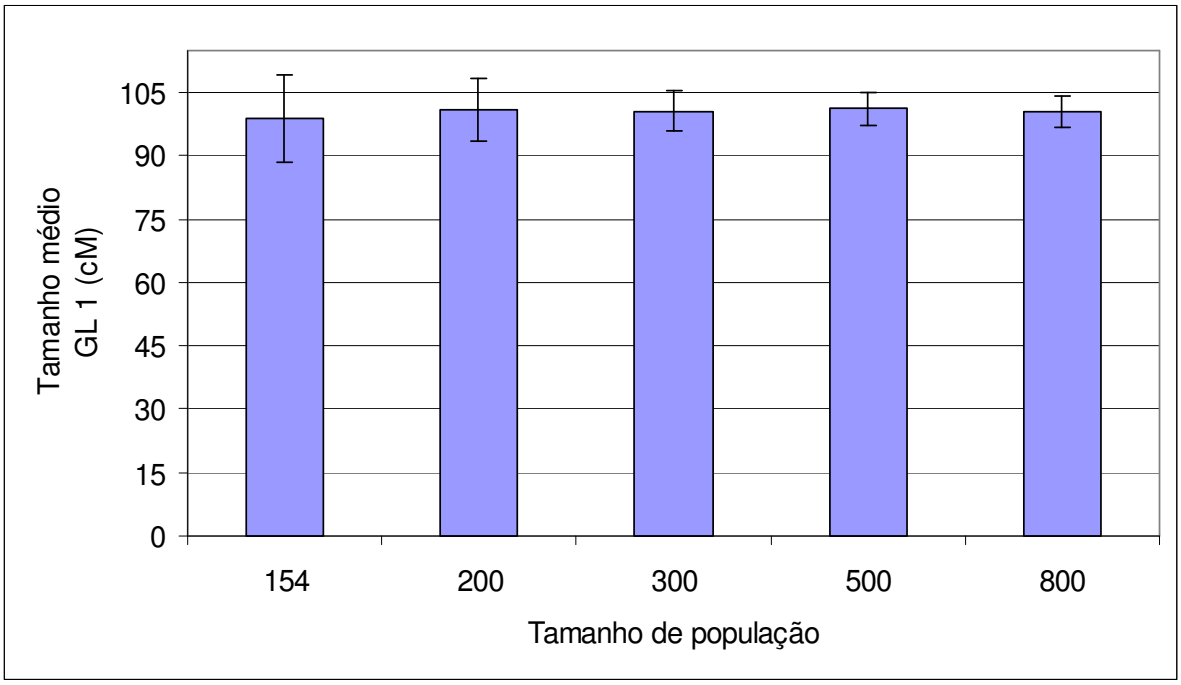


Figura 19 - Tamanho médio do grupo de ligação 1 e seu desvio padrão (linha vertical no topo de cada barra) em função do tamanho de população segregante. Avaliação feita nas populações com 154, 200, 300, 500 e 800 indivíduos, simuladas a partir do genoma com nível de saturação de 20 cM.

No nível de saturação de 20 cM, os grupos de ligação 2, 3, 4, 10 e 11 apresentaram diferença significativa pelo teste de comparação de médias entre os diferentes tamanhos de população analisados. O teste para a média geral (última coluna do Quadro 14) levou a um resultado adverso ao apresentado para os níveis de saturação de 5 e 10 cM. Pois, a população de menor número de indivíduos foi a que apresentou o menor tamanho médio de grupo de ligação.

Da interpretação dos resultados encontrados nas avaliações dos vários tamanhos de populações nos três níveis de saturação de genoma, pode-se inferir que a precisão obtida na recomposição do tamanho verdadeiro do genoma é maior tanto quanto forem maiores os tamanhos de populações utilizadas no processo de mapeamento, independente do nível de saturação do genoma por macas moleculares.

4.5 Distância média de marcas adjacentes

A distancia média de marcas adjacentes dentro de cada grupo de ligação foi obtida pela estimativa de duas médias, como descrito a seguir. Primeiramente, obteve-se uma média aritmética das distâncias entre marcas adjacentes dentro de cada grupo de ligação nas repetições em que houve a formação de 11 grupos de ligação no mapeamento. Posteriormente, obteve-se uma segunda média aritmética entre as médias aritméticas primeiramente obtidas em cada repetição.

Os valores de distância média entre marcas adjacentes e seu desvio padrão obtido para os vários tamanhos de população segregantes utilizadas no mapeamento genético estão apresentados no Quadro 15. A distância média entre marcas adjacentes esperado era de 5, 10 ou 20 cM, para as populações segregantes geradas a partir dos genomas com níveis de saturação de 5, 10 ou 20 cM, respectivamente.

Analisando-se o efeito do tamanho de população dentro do nível de saturação do genoma de 5 cM verifica-se a aproximação do valor de distância média entre marcas adjacentes ao valor esperado de 5 cM à medida que se aumenta o tamanho da população segregante de 50 para 800 indivíduos, para todos os grupos de ligação. Por exemplo, para o grupo de ligação 1 tem-se que a distância média foi de 5,26 e 5,14 cM, para os tamanhos de população de 50 e 800 indivíduos, respectivamente (Quadro 15). Exceção a este comportamento foi observado apenas para o grupo ligação 6, onde a distância média foi de 5,10 e 5,14 cM, para os tamanhos de população de 50 e 800 indivíduos, respectivamente. Uma explicação para a aproximação do valor da distância média entre marcas adjacentes ao valor esperado quando se aumenta de 50 para 800 indivíduos é a redução da amplitude de variação dos valores de distância média obtidos entre repetições. Esta redução da amplitude de variação dos valores de distância em torno da média pode ser observada pela redução dos valores de desvio padrão à medida que se aumentou o número de indivíduos nas populações segregantes. Por exemplo, para o grupo de ligação 1, tem-se que o desvio padrão foi de 0,58 e 0,14 cM, para os tamanhos de população de 50 e 800 indivíduos, respectivamente.

Quadro 15 - Distância média entre marcas adjacentes e seu desvio padrão em função do tamanho de população e nível de saturação do genoma. Avaliação feita apenas nas repetições em que houve a formação de 11 grupos de ligação no mapeamento das populações simuladas, conforme apresentado no Quadro 7

S.G.	T.P.	N.R.	Grupo de ligação											Média [2]
			1	2	3	4	5	6	7	8	9	10	11	
5 cM	50	70	5,26 a (0,58)	5,39 a (0,66)	5,28 a (0,52)	5,20 a (0,62)	5,26 a (0,59)	5,10 a (0,63)	5,23 a (0,52)	5,37 a (0,64)	5,27 a (0,52)	5,20 a (0,64)	5,31 a (0,62)	5,26 a (0,63)
	100	100	5,19 a (0,45)	5,18 b (0,39)	5,11 b (0,39)	5,20 a (0,42)	5,17 a (0,38)	5,10 a (0,45)	5,12 a (0,42)	5,14 b (0,35)	5,22 a (0,37)	5,24 a (0,39)	5,12 b (0,37)	5,16 bc (0,40)
	154	100	5,18 a (0,31)	5,21 ab (0,36)	5,26 a (0,32)	5,14 a (0,28)	5,24 a (0,36)	5,25 a (0,33)	5,20 a (0,29)	5,23 ab (0,32)	5,20 a (0,33)	5,14 a (0,34)	5,20 ab (0,30)	5,21 b (0,32)
	200	100	5,15 a (0,29)	5,11 b (0,28)	5,14 ab (0,27)	5,13 a (0,28)	5,11 a (0,25)	5,17 a (0,29)	5,12 a (0,30)	5,16 b (0,31)	5,12 a (0,29)	5,18 a (0,29)	5,15 ab (0,27)	5,14 c (0,28)
	300	100	5,12 a (0,18)	5,11 b (0,22)	5,16 ab (0,25)	5,10 a (0,19)	5,17 a (0,23)	5,16 a (0,25)	5,13 a (0,22)	5,14 b (0,21)	5,14 a (0,25)	5,14 a (0,24)	5,18 ab (0,22)	5,14 c (0,23)
	500	100	5,15 a (0,16)	5,16 b (0,17)	5,14 ab (0,16)	5,18 a (0,16)	5,14 a (0,17)	5,19 a (0,15)	5,16 a (0,18)	5,16 b (0,16)	5,14 a (0,18)	5,18 a (0,15)	5,12 b (0,19)	5,16 c (0,17)
	800	100	5,14 a (0,14)	5,13 b (0,12)	5,12 ab (0,13)	5,14 a (0,14)	5,14 a (0,13)	5,14 a (0,15)	5,16 a (0,14)	5,16 b (0,14)	5,13 a (0,12)	5,15 a (0,13)	5,11 b (0,13)	5,14 c (0,13)
10 cM	100	98	10,43 a (0,91)	10,51 a (0,94)	10,38 a (0,89)	10,50 a (0,90)	10,51 a (0,84)	10,56 a (0,82)	10,32 a (0,82)	10,37 a (0,89)	10,46 a (0,79)	10,40 a (0,91)	10,73 a (0,91)	10,45 a (0,87)
	154	100	10,32 a (0,72)	10,32 a (0,74)	10,32 a (0,72)	10,38 a (0,66)	10,32 a (0,72)	10,29 b (0,64)	10,42 a (0,72)	10,38 a (0,80)	10,48 a (0,71)	10,44 a (0,75)	10,35 a (0,65)	10,36 ab (0,71)
	200	100	10,30 a (0,65)	10,31 a (0,71)	10,33 a (0,64)	10,27 a (0,56)	10,40 a (0,65)	10,29 b (0,52)	10,23 a (0,59)	10,20 a (0,55)	10,47 a (0,58)	10,30 a (0,69)	10,31 a (0,60)	10,31 bc (0,62)
	300	100	10,24 a (0,54)	10,36 a (0,53)	10,31 a (0,50)	10,32 a (0,47)	10,29 a (0,54)	10,39 ab (0,49)	10,32 a (0,50)	10,27 a (0,45)	10,28 a (0,48)	10,27 a (0,46)	10,30 a (0,48)	10,31 bc (0,49)
	500	100	10,28 a (0,41)	10,30 a (0,40)	10,28 a (0,37)	10,31 a (0,40)	10,29 a (0,42)	10,30 ab (0,39)	10,31 a (0,41)	10,27 a (0,39)	10,31 a (0,38)	10,28 a (0,37)	10,27 a (0,38)	10,29 bc (0,39)
	800	100	10,23 a (0,30)	10,27 a (0,27)	10,34 a (0,29)	10,24 a (0,29)	10,26 a (0,30)	10,31 ab (0,32)	10,31 a (0,31)	10,32 a (0,30)	10,25 a (0,32)	10,30 a (0,32)	10,22 a (0,28)	10,28 c (0,30)
20 cM	154	43	20,03 a (1,67)	20,12 ab (1,31)	19,79 a (1,36)	20,15 a (1,51)	20,02 a (1,45)	20,23 a (1,08)	19,95 a (1,40)	19,78 a (1,59)	19,83 a (1,29)	20,04 a (1,40)	19,77 a (1,45)	19,97 b (1,41)
	200	86	20,24 a (1,43)	19,74 b (1,46)	20,41 a (1,25)	20,08 a (1,54)	20,13 a (1,43)	20,38 a (1,41)	20,29 a (1,39)	20,22 a (1,42)	20,28 a (1,50)	20,29 a (1,35)	20,04 a (1,33)	20,19 a (1,41)
	300	98	20,10 a (0,94)	20,29 a (1,22)	20,25 a (1,17)	20,23 a (1,20)	20,22 a (1,15)	20,12 a (1,14)	20,46 a (1,20)	20,11 a (1,09)	20,27 a (1,26)	20,30 a (1,17)	20,25 a (1,25)	20,24 a (1,37)
	500	100	20,24 a (0,79)	20,15 ab (0,79)	20,21 a (0,85)	20,33 a (0,90)	20,20 a (0,82)	20,17 a (0,96)	20,20 a (0,82)	20,13 a (0,98)	20,26 a (0,84)	20,16 a (0,86)	20,05 a (0,87)	20,19 a (0,86)
	800	100	20,08 a (0,72)	20,19 ab (0,68)	20,24 a (0,75)	20,08 a (0,75)	20,11 a (0,69)	20,07 a (0,69)	20,25 a (0,73)	20,08 a (0,66)	20,20 a (0,64)	20,08 a (0,75)	20,23 a (0,73)	20,15 ab (0,71)

[1]\ Entre parênteses estão os valores de desvio padrão; [2]\ Média dos onze grupos de ligação.

S.G.= Saturação do genoma; T.P.= Tamanho da população; N.R.= número de repetições avaliadas.

Médias seguidas por letras iguais nas colunas, dentro de cada nível de saturação do genoma, não diferem entre si pelo teste de Tukey a 1% de probabilidade (P<0,01).

Uma melhor visualização do comportamento dos valores obtidos para a distância média entre marcas adjacentes e desvio padrão com o aumento do número de indivíduos nas populações segregantes, geradas a partir do genoma com nível de saturação de 5 cM, pode ser observado nas Figuras 20 e 21, que apesar de representarem apenas o comportamento observado para o grupo de ligação 1, é representativo do ocorrido para todos os grupos de ligação analisados dentro do nível de saturação do genoma de 5 cM. De acordo com os valores de dispersão das distâncias médias apresentados na Figura 20 fica evidente o efeito da redução da amplitude de variação das distâncias médias com o aumento do tamanho das populações segregantes. Esta redução também pode ser visualizada na Figura 21 pela análise do comportamento dos valores de desvio padrão apresentado nos topos de cada barra.

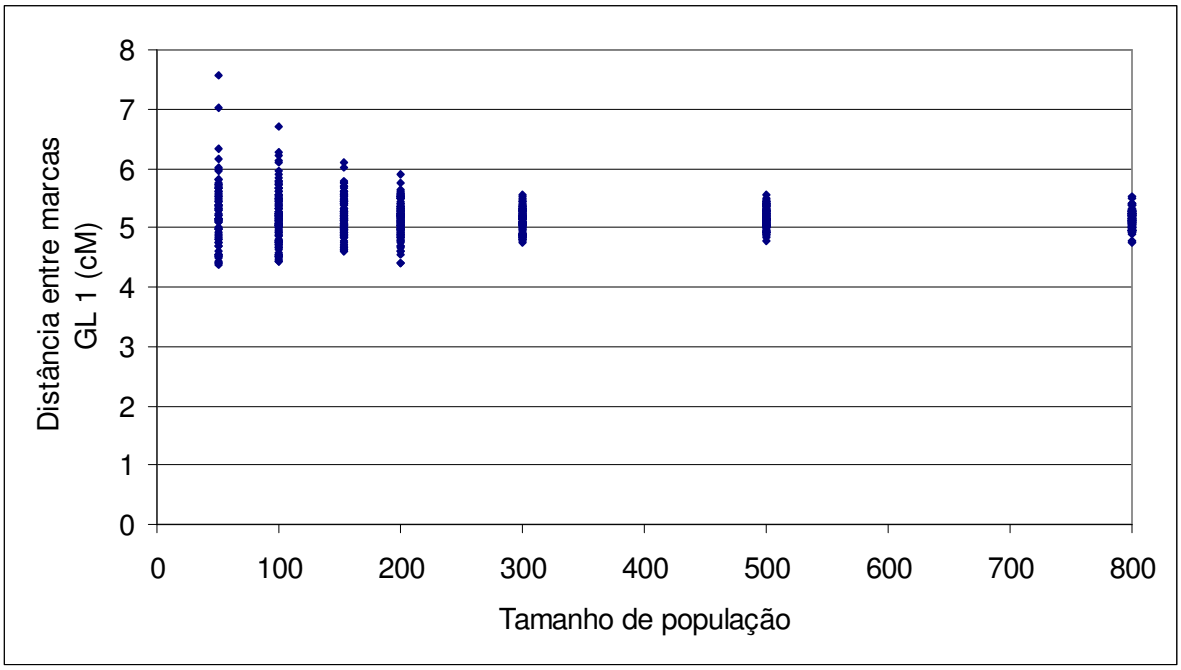


Figura 20 - Dispersão dos valores de distância entre marcas adjacentes no grupo de ligação 1 em função do tamanho de população segregante. Avaliação feita nas populações com 50, 100, 154, 200, 300, 500 e 800 indivíduos, simuladas a partir do genoma com nível de saturação de 5 cM.

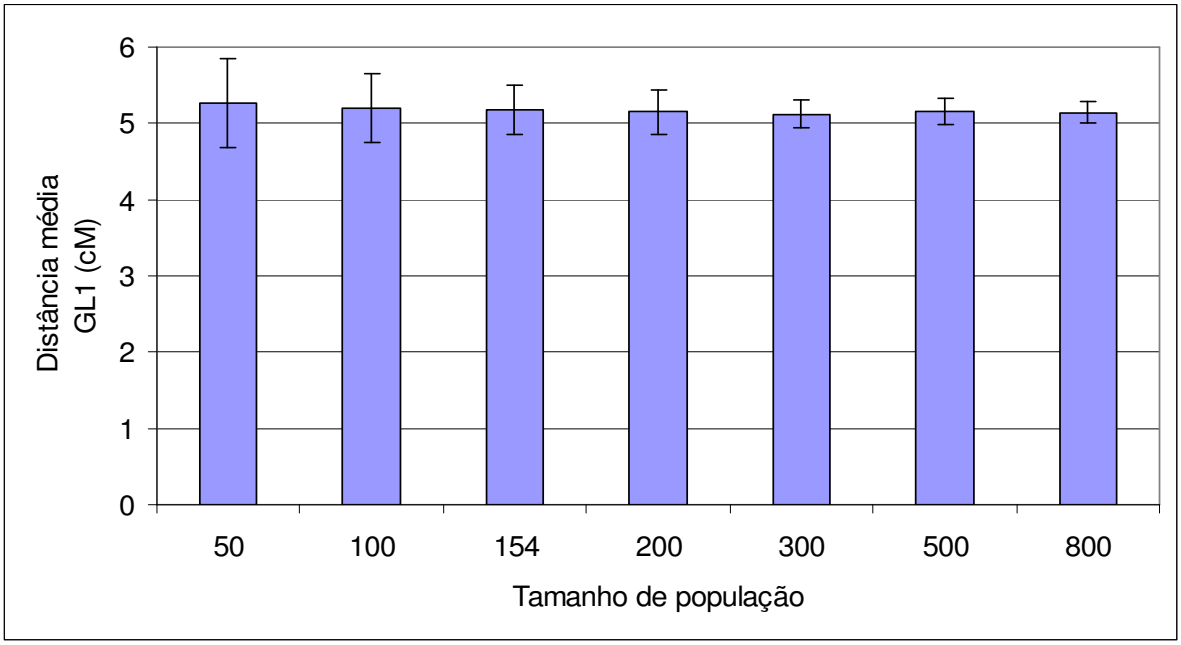


Figura 21 - Distância média entre marcas adjacentes no grupo de ligação 1 e seu desvio padrão (linha vertical no topo de cada barra) em função do tamanho de população segregante. Avaliação feita nas populações com 50, 100, 154, 200, 300, 500 e 800 indivíduos, simuladas a partir do genoma com nível de saturação de 5 cM.

Nos grupos de ligação 1, 4, 5, 6, 7, 9 e 10 as médias para os diferentes tamanhos de população no genoma de saturação de 5 cM não diferem significativamente pelo teste de média. Já, nos demais grupos de ligação há diferença, com os maiores tamanhos de população apresentando os menores valores de distâncias entre marcas adjacentes e aproximando-se do valor esperado de 5 cM. As médias gerais (última coluna do Quadro 15) diferem significativamente pelo teste de média e apresentam os menores valores associados as maiores populações. A partir do tamanho de população de 200 indivíduos as distâncias médias entre marcas adjacentes não si diferem pelo teste de médias.

Para as populações geradas a partir do genoma com nível de saturação de 10 cM houve, assim como para as populações geradas a partir do genoma de 5 cM, aproximação do valor de distância média entre marcas adjacentes ao valor esperado de 10 cM com o aumento do tamanho de população segregante de 100 para 800 indivíduos. Por exemplo, para o grupo de ligação 1, tem-se que a distância média foi de 10,43 e 10,23 cM, para os tamanhos de população de 100 e 800 indivíduos, respectivamente (Quadro 15). Também, os valores de desvio padrão diminuíram com o aumento do tamanho das populações. Por exemplo, para o grupo de ligação 1 tem-se que o desvio padrão foi de 0,91 e 0,30 cM, para os tamanhos de população de 100

e 800 indivíduos, respectivamente. As explicações para a aproximação do valor de distância média entre marcas adjacentes ao valor esperado em cada grupo de ligação.

Uma melhor visualização do comportamento dos valores obtidos para a distância média entre marcas adjacentes e desvio padrão com o aumento do número de indivíduos nas populações segregantes pode ser observada nas Figuras 22 e 23, que apesar de representar apenas o comportamento observado para o grupo de ligação 1, é representativo do ocorrido para todos os grupos de ligação analisados dentro do nível de saturação do genoma de 10 cM. De acordo com os valores de dispersão das distâncias médias apresentados na Figura 22 fica evidente o efeito da redução da amplitude de variação das distâncias médias com o aumento do tamanho das populações segregantes. Esta redução também pode ser visualizada na Figura 23 pela análise do comportamento dos valores de desvio padrão apresentado nos topos de cada barra.

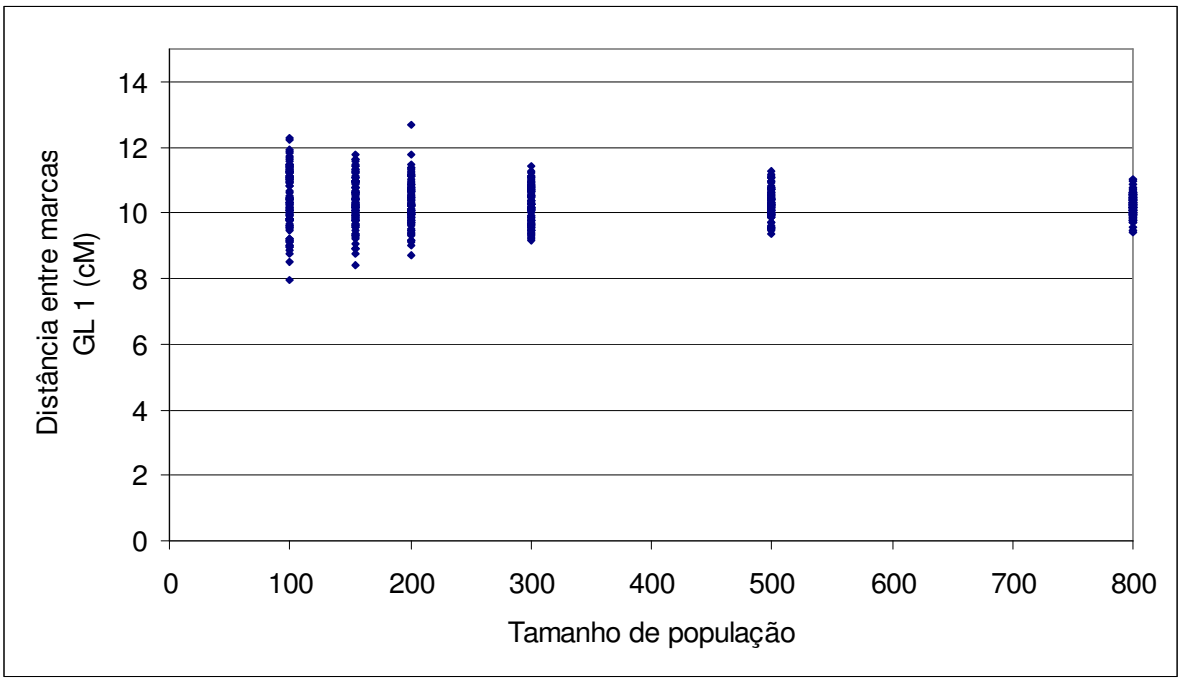


Figura 22 - Dispersão dos valores de distância entre marcas adjacentes no grupo de ligação 1 em função do tamanho de população. Avaliação feita nas populações com 100, 154, 200, 300, 500 e 800 indivíduos, simuladas a partir do genoma com nível de saturação de 10 cM.

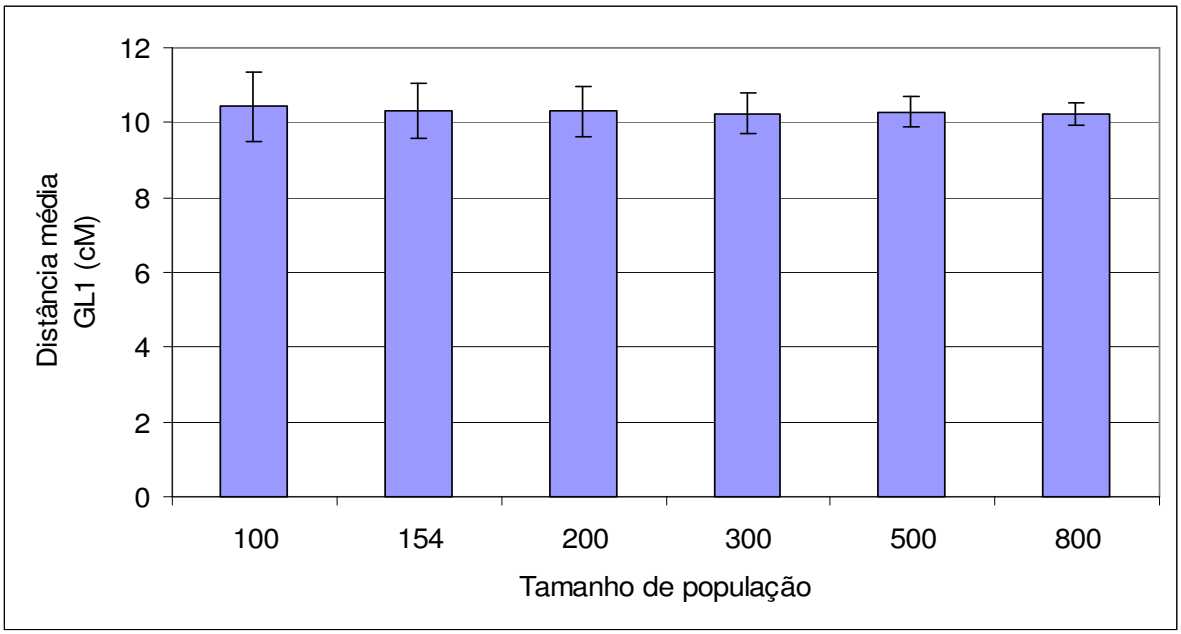


Figura 23 - Distância média entre marcas adjacentes no grupo de ligação 1 e seu desvio padrão (linha vertical no topo de cada barra) em função do tamanho de população segregante. Avaliação feita nas populações com 100, 154, 200, 300, 500 e 800 indivíduos, simuladas a partir do genoma com nível de saturação de 10 cM.

No nível de saturação do genoma de 10 cM apenas para o grupo de ligação 6 houve diferença significativa entre médias nos diferentes tamanhos de populações. A realização do teste de Tukey ($P< 0,01$) com as médias gerais (última coluna do Quadro 15) evidenciou diferenças significativas, com as menores médias de distâncias entre marcas adjacentes associadas aos maiores tamanhos de população e aproximando-se do valor esperado de 10 cM.

Nas populações geradas a partir do genoma com nível de saturação de 20 cM os valores de distância média entre marcas adjacentes não seguiram uma tendência de redução com o aumento do número de indivíduos nas populações segregante de 154 para 800 indivíduos, como ocorreu para a maioria dos grupos de ligação analisados nos níveis de saturação de genoma de 5 e 10 cM. Como visto anteriormente, os valores de distância média para os níveis de saturação de 5 e 10 cM foram maiores para as populações de 50 e 100 indivíduos, respectivamente, do que para as de 800 indivíduos, para a maioria dos grupos de ligação. Já, para o nível de saturação de 20 cM ocorreu o contrário, ou seja, as distâncias médias foram menores para as populações de 154 indivíduos do que para as de 800 indivíduos. As populações com 800 indivíduos só apresentaram menores valores de distância média do que as com 154 indivíduos para os grupos de ligação 4 e 6 (Quadro 15). Uma melhor visualização do comportamento do tamanho médio dos grupos de ligação e

desvio padrão com o aumento do número de indivíduos pode ser observado nas Figuras 24 e 25, que apesar de representarem apenas o comportamento observado para o grupo de ligação 1, é representativo do ocorrido para todos os grupos de ligação analisados dentro do nível de saturação do genoma de 20 cM.

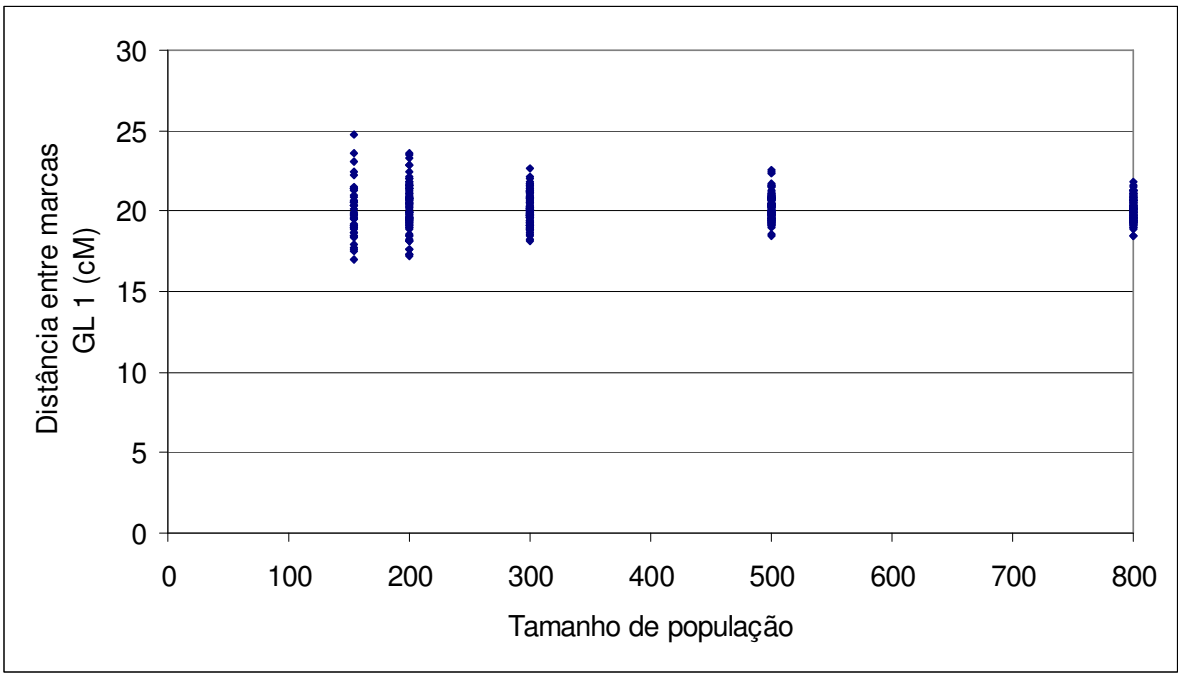


Figura 24 - Dispersão dos valores de distância entre marcas adjacentes no grupo de ligação 1 em função do tamanho de população. Avaliação feita nas populações com 154, 200, 300, 500 e 800 indivíduos, simuladas a partir do genoma com nível de saturação de 20 cM.

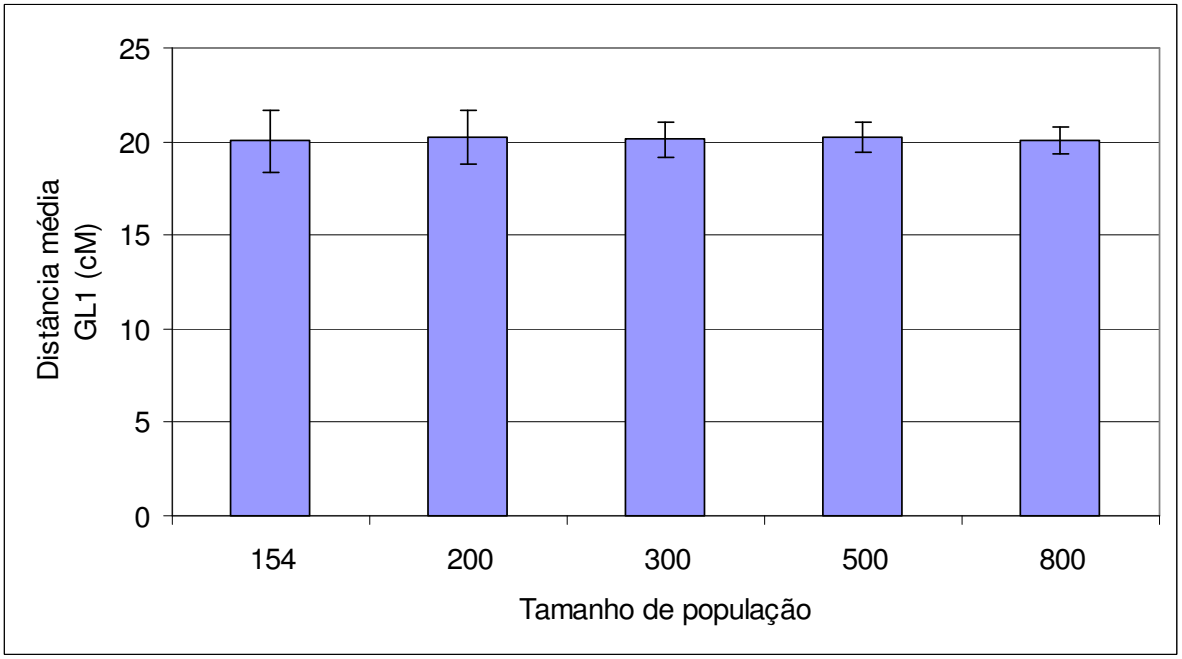


Figura 25 - Distância média entre marcas adjacentes no grupo de ligação 1 e seu desvio padrão (linha vertical no topo de cada barra) em função do tamanho de população segregante. Avaliação feita nas populações com 154, 200, 300, 500 e 800 indivíduos, simuladas a partir do genoma com nível de saturação de 20 cM.

No nível de saturação do genoma de 20 cM apenas as médias para o grupo de ligação 2 diferiram pelo teste de Tukey (P< 0,01). Resultado contrário encontrado para os genomas de 5 e 10 cM foi obtido para o genoma de 20 cM, no qual a média da população de menor tamanho apresentou-se significativamente menor do que as médias de algumas populações maiores (última coluna do Quadro 15). Mas, este resultado deve ser analisado com cuidado, dado ao menor número de repetições avaliadas para o tamanho de população de 154 indivíduos.

Da interpretação dos resultados encontrados nas avaliações dos vários tamanhos de populações nos três níveis de saturação de genoma quanto à distância média de marcas adjacentes nos grupos de ligação, pode-se inferir que a precisão obtida na estimativa das distâncias é maior tanto quanto maiores foram os tamanhos de populações utilizadas no processo de mapeamento, independente do nível de saturação do genoma por macas moleculares.

4.6 Variância das distâncias entre marcas adjacentes

A partir das distâncias de marcas adjacentes obtidas nos grupos de ligação foi estimada uma variância amostral. Os valores apresentados no Quadro 16 para cada de grupo de ligação são as médias aritméticas das variâncias obtidas em cada repetição em que houve a formação de 11 grupos de ligação no mapeamento das populações simuladas.

As estimativas esperadas de variância na análise dos mapas obtidos das populações segregantes simuladas eram de zero para qualquer tamanho de população e nível de saturação de genoma. Pois, nos genomas utilizados para geração das populações segregantes os marcadores estavam distribuídos de forma eqüidistante dentro dos 11 grupos de ligação. Sendo, que para o genoma com nível de saturação de 5 cM os marcadores encontravam-se distribuídos à distâncias eqüidistantes de 5 cM ao longo dos grupos de ligação. Para o genoma com nível de saturação de 10 cM os marcadores encontravam-se distribuídos à distância de 10 cM ao longo dos grupos de ligação. E, finalmente, para o genoma com nível de saturação de 20 cM os marcadores encontravam-se distribuídos à distância de 20 cM ao longo dos grupos de ligação.

Quanto menores forem os valores de variância, mais eqüidistantemente estão distribuídos os marcadores dentro do grupo de ligação. Portanto, a obtenção de

valores de variância pequenos e distâncias médias entre marcas adjacentes próximos aos valores esperados em cada nível de saturação do genoma, indicam boa recuperação do genoma com o mapeamento das populações segregantes.

Analisando-se o efeito do tamanho de população dentro do nível de saturação do genoma de 5 cM verifica-se uma redução da variância média das distâncias entre marcas adjacentes à medida que se aumenta o tamanho da população segregante, em todos os grupos de ligação. Por exemplo, para o grupo de ligação 1 tem-se que o valor de variância média foi de 7,64 e 0,41, para os tamanhos de população de 50 e 800 indivíduos, respectivamente (Quadro 16). Além, da redução da variância média ocorreu, também, uma redução da amplitude de variação dos valores de variância entre repetições à medida que se aumentou o tamanho da população segregante, o que é demonstrado pela redução do desvio padrão. Por exemplo, para o grupo de ligação 1, tem-se que o desvio padrão foi de 7,39 e 0,15 cM, para os tamanhos de população de 50 e 800 indivíduos, respectivamente.

Quadro 16 - Variância média das distâncias entre marcas adjacentes e seu desvio padrão em função do tamanho de população e nível de saturação do genoma. Avaliação feita apenas nas repetições em que houve a formação de 11 grupos de ligação no mapeamento das populações simuladas, conforme apresentado no Quadro 7 (Tukey a 1%)

S.G.	T.P.	N.R.	Grupo de ligação											Média [2]
			1	2	3	4	5	6	7	8	9	10	11	
5 cM	50	70	7,64 a (7,39)[1]	8,05 a (7,88)	6,29 a (2,43)	6,84 a (2,78)	7,68 a (9,94)	7,29 a (3,79)	7,00 a (3,05)	8,63 a (6,91)	6,38 a (2,00)	7,57 a (8,61)	8,39 a (12,14)	7,44 a (6,92)
	100	100	3,61 b (3,27)	3,35 b (1,24)	3,25 b (1,06)	3,51 b (3,27)	3,25 b (1,06)	3,11 b (0,99)	3,07 b (1,20)	3,27 b (1,24)	3,25 b (1,05)	3,32 b (1,14)	3,16 b (1,18)	3,29 b (1,73)
	154	100	2,08 c (0,62)	2,05 c (0,70)	2,05 c (0,62)	2,02 c (0,72)	2,05 bc (0,69)	2,07 c (0,71)	2,15 c (0,74)	2,09 c (0,77)	2,05 c (0,59)	1,99 bc (0,57)	1,90 bc (0,66)	2,05 c (0,68)
	200	100	1,53 cd (0,54)	1,60 cd (0,52)	1,60 c (0,52)	1,55 cd (0,56)	1,53 cd (0,57)	1,61 cd (0,57)	1,59 d (0,56)	1,73 cd (0,52)	1,66 c (0,58)	1,53 cd (0,48)	1,60 bc (0,54)	1,59 d (0,54)
	300	100	0,98 cd (0,34)	1,02 cd (0,33)	1,07 d (0,33)	1,06 de (0,32)	1,04 cd (0,34)	1,03 de (0,36)	1,03 de (0,32)	1,07 cde (0,36)	1,05 d (0,39)	1,06 cd (0,36)	1,05 c (0,33)	1,04 e (0,34)
	500	100	0,61 d (0,22)	0,60 d (0,19)	0,60 de (0,17)	0,64 e (0,19)	0,61 cd (0,17)	0,62 e (0,21)	0,62 ef (0,21)	0,60 de (0,19)	0,63 e (0,18)	0,62 cd (0,20)	0,61 c (0,21)	0,61 f (0,20)
	800	100	0,41 d (0,15)	0,37 d (0,11)	0,39 de (0,12)	0,39 e (0,14)	0,39 d (0,11)	0,38 e (0,11)	0,39 f (0,12)	0,39 e (0,13)	0,38 e (0,11)	0,42 d (0,14)	0,39 c (0,13)	0,39 f (0,12)
10 cM	100	98	7,78 a (4,05)	8,27 a (5,93)	8,47 a (4,55)	8,07 a (3,94)	7,79 a (3,69)	8,08 a (3,43)	7,76 a (3,28)	7,98 a (3,73)	7,62 a (3,79)	7,91 a (3,71)	7,14 a (3,78)	7,90 a (4,04)
	154	100	5,10 b (2,69)	4,58 b (2,32)	4,72 b (2,20)	4,74 b (2,16)	5,09 b (2,20)	4,94 b (2,51)	5,15 b (2,92)	5,32 b (2,76)	5,26 b (2,68)	4,54 b (1,84)	5,15 b (2,52)	4,96 b (2,46)
	200	100	3,79 c (1,99)	4,22 b (2,10)	3,91 b (1,76)	3,71 c (1,80)	3,68 c (1,56)	3,95 c (1,83)	3,59 c (1,78)	3,77 c (1,60)	3,77 c (1,79)	3,81 b (1,82)	3,83 c (2,24)	3,82 c (1,85)
	300	100	2,69 d (1,46)	2,37 c (0,95)	2,67 c (1,21)	2,49 d (1,14)	2,41 d (1,38)	2,77 d (1,38)	2,77 c (1,10)	2,54 d (1,17)	2,44 d (1,06)	2,45 c (1,17)	2,53 d (1,24)	2,56 d (1,22)
	500	100	1,43 e (0,60)	1,59 cd (0,70)	1,48 d (0,63)	1,40 e (0,65)	1,44 e (0,74)	1,40 e (0,67)	1,51 d (0,77)	1,40 e (0,62)	1,48 de (0,75)	1,53 cd (0,73)	1,47 e (0,69)	1,47 e (0,69)
	800	100	0,90 e (0,48)	0,91 d (0,42)	0,99 d (0,49)	0,98 e (0,51)	0,92 e (0,49)	0,91 e (0,43)	0,90 d (0,41)	0,89 e (0,43)	0,99 e (0,47)	0,82 d (0,37)	0,88 e (0,39)	0,92 f (0,45)
20 cM	154	43	11,10 a (8,57)	9,89 ab (6,93)	10,72 a (6,87)	11,55 a (6,30)	9,21 ab (6,17)	11,59 a (6,58)	12,33 a (6,40)	8,82 ab (5,64)	9,46 a (5,55)	10,86 a (6,73)	9,89 a (5,94)	10,49 a (6,58)
	200	86	10,31 a (6,83)	10,23 a (6,60)	11,05 a (7,90)	10,16 a (6,86)	9,62 a (6,55)	9,41 a (7,46)	9,64 a (6,81)	9,38 a (5,46)	9,37 a (5,96)	9,54 a (5,57)	9,36 a (5,45)	9,82 a (6,53)
	300	98	6,22 b (3,69)	7,43 b (6,04)	6,25 b (4,15)	6,20 b (4,40)	7,14 b (5,52)	6,39 b (4,47)	6,20 b (4,78)	6,79 b (4,32)	7,87 a (5,25)	6,91 b (4,56)	6,99 b (4,50)	6,76 b (4,75)
	500	100	3,80 c (3,12)	3,83 c (2,31)	3,75 c (2,74)	4,03 c (2,58)	3,85 c (2,69)	3,95 c (2,44)	3,64 c (2,66)	3,85 c (2,83)	3,84 b (2,73)	3,72 c (2,61)	3,79 c (2,59)	3,82 c (2,66)
	800	100	2,44 c (1,67)	2,21 c (1,71)	2,49 c (1,77)	2,49 c (1,81)	2,46 c (1,79)	2,57 c (1,79)	2,61 c (1,83)	2,83 c (1,96)	2,70 b (1,92)	2,66 c (2,14)	2,57 c (1,57)	2,55 d (1,82)

[1]\ Entre parênteses estão os valores de desvio padrão; [2]\ Média dos onze grupos de ligação.

S.G.= Saturação do genoma; T.P.= Tamanho da população; N.R.= número de repetições avaliadas.

Médias seguidas por letras iguais nas colunas, dentro de cada nível de saturação do genoma, não diferem entre si pelo teste de Tukey a 1% de probabilidade (P<0,01).

Uma melhor visualização do comportamento dos valores de variância e desvio padrão com o aumento do número de indivíduos pode ser observada nas Figuras 26 e 27, que apesar de representar apenas o comportamento observado para o grupo de ligação 1, é representativo do ocorrido para todos os grupos de ligação analisados dentro do nível de saturação do genoma de 5 cM. De acordo com os valores variância apresentados na Figura 26 fica evidente o efeito da redução da amplitude de variação dos valores de variância das distâncias entre marcas adjacentes com o aumento do tamanho das populações segregantes. Esta redução também pode ser visualizada na Figura 27 pela análise do comportamento dos valores de desvio padrão apresentado nos topos de cada barra.

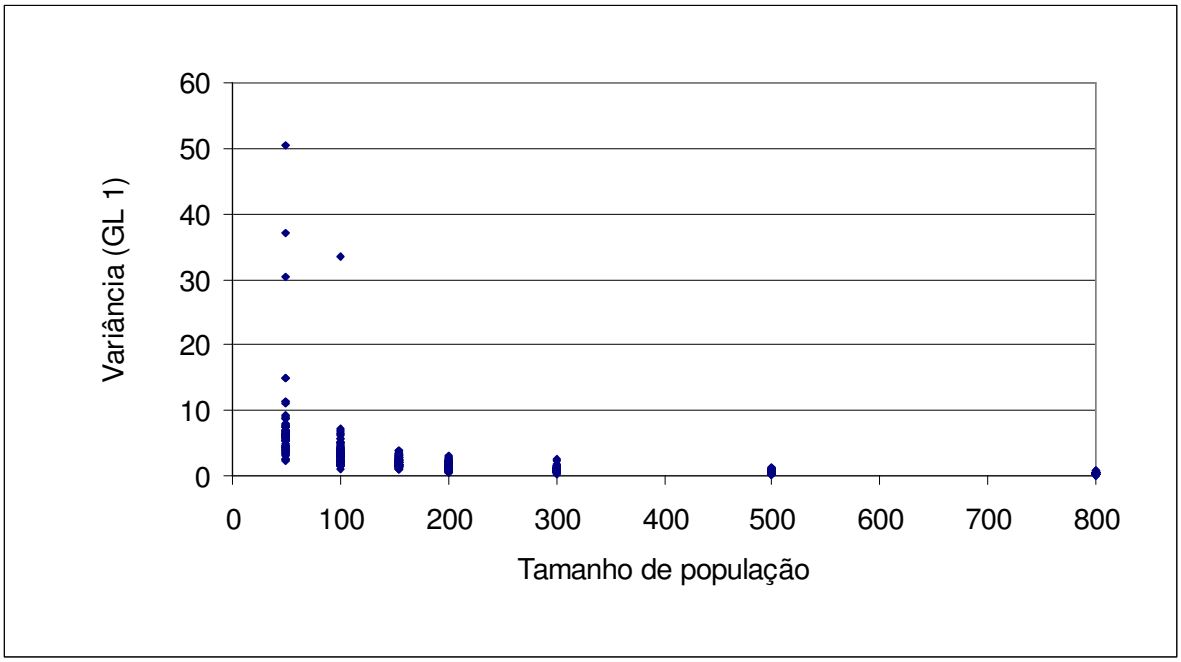


Figura 26 - Dispersão dos valores de variância das distâncias entre marcas adjacentes no grupo de ligação 1 em função do tamanho de população. Avaliação feita nas populações com 50, 100, 154, 200, 300, 500 e 800 indivíduos, simuladas a partir do genoma com nível de saturação de 5 cM.

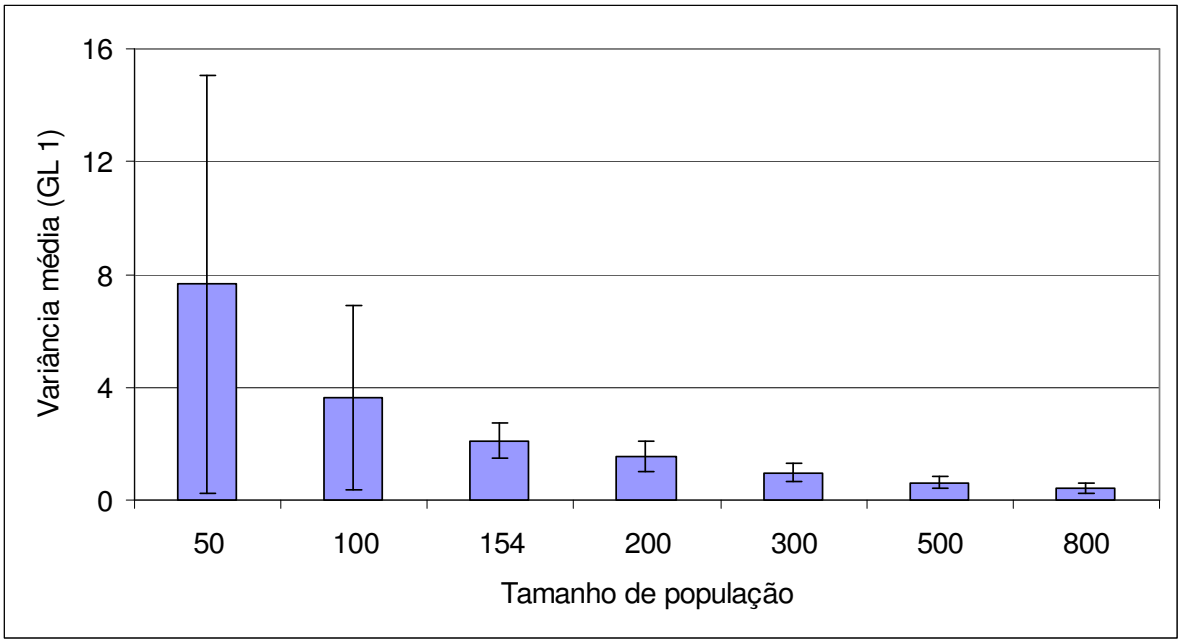


Figura 27 - Variância média das distâncias entre marcas adjacentes no grupo de ligação 1 e seu desvio padrão (linha vertical no topo de cada barra) em função do tamanho de população. Avaliação feita nas populações com 50, 100, 154, 200, 300, 500 e 800 indivíduos, simuladas a partir do genoma com nível de saturação de 5 cM.

A redução dos valores de variância com o aumento do tamanho da população é evidenciada pelas diferenças significativas dadas pelo teste de média, no qual as menores populações apresentam maiores valores do que as populações com mais indivíduos, para todos os grupos de ligação dentro do nível de saturação do genoma de 5 cM. Na análise da média geral (última coluna do Quadro 16) houve diferença significativa entre médias dos diferentes tamanhos de populações, sendo que as médias das populações de tamanhos 500 e 800 não diferiram significativamente.

Para os vários tamanhos de populações segregantes analisados dentro do nível de saturação do genoma de 10 cM ocorreu, assim como para o nível de saturação de 5 cM, uma redução da variância média das distâncias entre marcas adjacentes à medida que se aumentou o tamanho da população segregante em todos os grupos de ligação. Por exemplo, para o grupo de ligação 1 tem-se que o valor de variância média foi de 7,78 e 0,90, para os tamanhos de população de 100 e 800 indivíduos, respectivamente (Quadro 16). Além, da redução da variância média ocorreu, também, uma redução da amplitude de variação dos valores de variância entre repetições à medida que se aumentou o tamanho da população segregante, o que é demonstrado pela redução do desvio padrão. Isto tem uma grande implicação

na obtenção de maior confiabilidade no mapeamento quando da utilização de populações com maior número de indivíduos. Por exemplo, para o grupo de ligação 1 tem-se que o desvio padrão foi de 4,05 e 0,48 cM, para os tamanhos de população de 100 e 800 indivíduos, respectivamente.

Uma melhor visualização do comportamento dos valores de variância e desvio padrão com o aumento do número de indivíduos pode ser observada nas Figuras 28 e 29, que apesar de representar apenas o comportamento observado para o grupo de ligação 1, é representativo do ocorrido para todos os grupos de ligação analisados dentro do nível de saturação do genoma de 10 cM. De acordo com os valores de variância apresentados na Figura 28 fica evidente o efeito da redução da amplitude de variação dos valores de variância das distâncias entre marcas adjacentes com o aumento do tamanho das populações segregantes. Esta redução também pode ser visualizada na Figura 29 pela análise do comportamento dos valores de desvio padrão apresentado nos topos de cada barra.

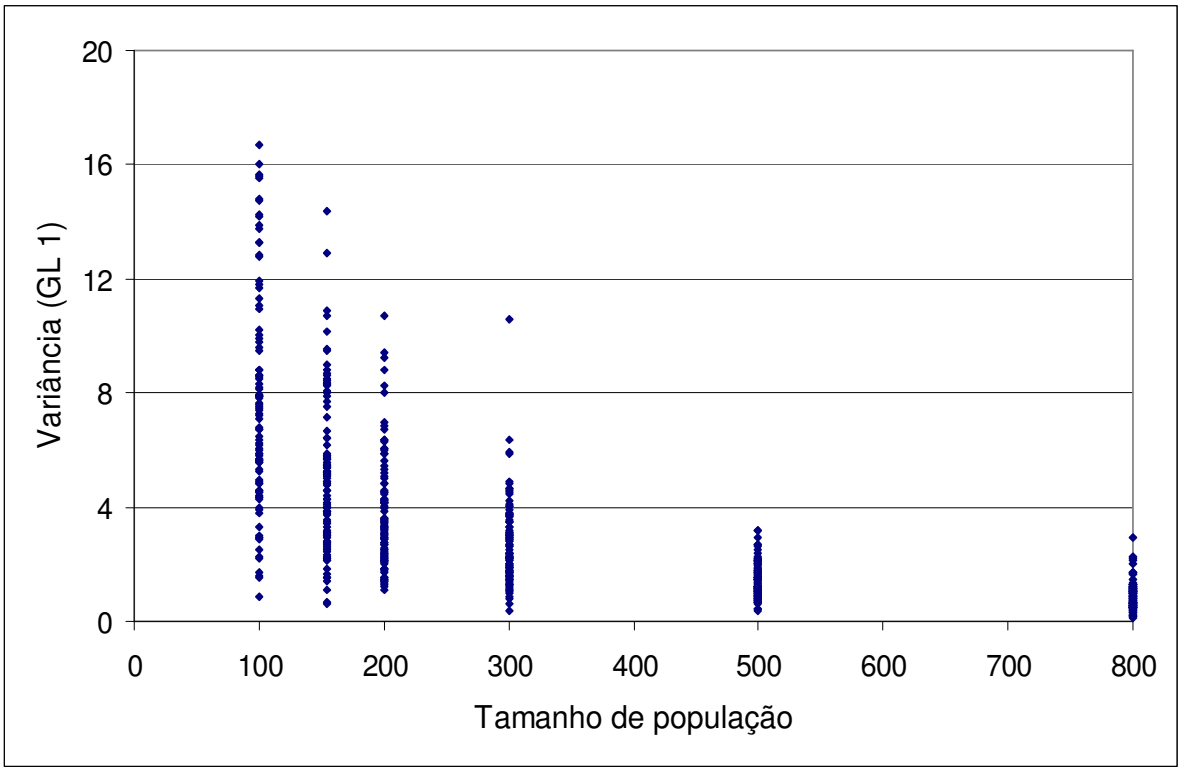


Figura 28 - Dispersão dos valores de variância das distâncias entre marcas adjacentes no grupo de ligação 1 em função do tamanho de população. Avaliação feita nas populações com 100, 154, 200, 300, 500 e 800 indivíduos, simuladas a partir do genoma com nível de saturação de 10 cM.

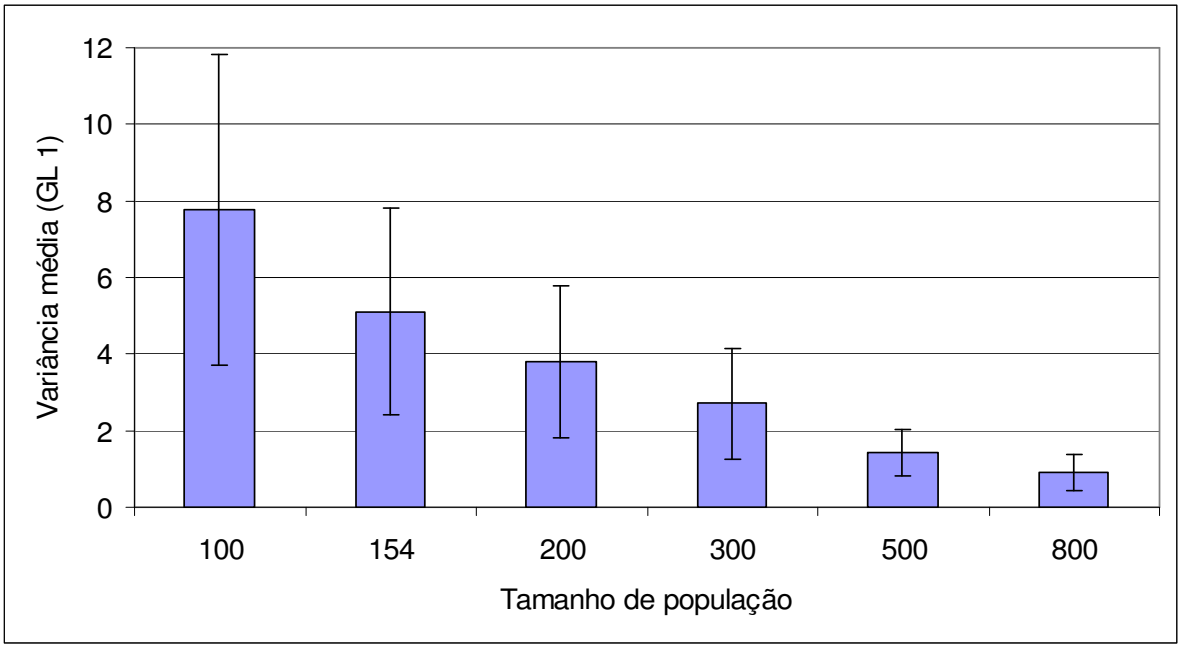


Figura 29 - Variância média das distâncias entre marcas adjacentes no grupo de ligação 1 e seu desvio padrão (linha vertical no topo de cada barra) em função do tamanho de população. Avaliação feita nas populações com 100, 154, 200, 300, 500 e 800 indivíduos, simuladas a partir do genoma com nível de saturação de 10 cM.

A redução dos valores de variância com o aumento do tamanho da população é evidenciada pelas diferenças significativas dadas pelo teste de média, no qual as menores médias estão associadas às maiores populações, em todos os grupos de ligação dentro do nível de saturação do genoma de 10 cM. As médias gerais (última coluna do Quadro 16) apresentaram diferenças significativas para todos os tamanhos de população, evidenciando que as menores populações apresentam maiores valores de variância das distâncias entre marcas adjacentes, portanto, levando a menor precisão no mapeamento genético.

Nas populações geradas a partir do genoma com nível de saturação de 20 cM os valores de variância média das distâncias entre marcas adjacentes seguiram o mesmo comportamento das populações geradas a partir dos genomas com 5 e 10 cM, ou seja, houve uma redução da variância média das distâncias entre marcas adjacentes à medida que se aumentou o tamanho da população segregante, em todos os grupos de ligação. Por exemplo, para o grupo de ligação 1 tem-se que o valor de variância média foi de 11,10 e 2,44, para os tamanhos de população de 154 e 800 indivíduos, respectivamente (Quadro 16). Além, da redução da variância média ocorreu também, uma redução da amplitude de variação dos valores de variância entre repetições à medida que se aumentou o tamanho da população segregante. Por

exemplo, para o grupo de ligação 1, tem-se que o desvio padrão foi de 8,57 e 1,67, para os tamanhos de população de 154 e 800 indivíduos, respectivamente.

Uma melhor visualização do comportamento dos valores de variância e desvio padrão com o aumento do número de indivíduos pode ser observada nas Figuras 30 e 31, que apesar de representar apenas o comportamento observado para o grupo de ligação 1, é representativo do ocorrido para todos os grupos de ligação analisados dentro do nível de saturação do genoma de 20 cM. De acordo com os valores de variância apresentados na Figura 30 fica evidente o efeito da redução da amplitude de variação dos valores de variância das distâncias entre marcas adjacentes com o aumento do tamanho das populações segregantes. Esta redução também pode ser visualizada na Figura 31 pela análise do comportamento dos valores de desvio padrão apresentado nos topos de cada barra.

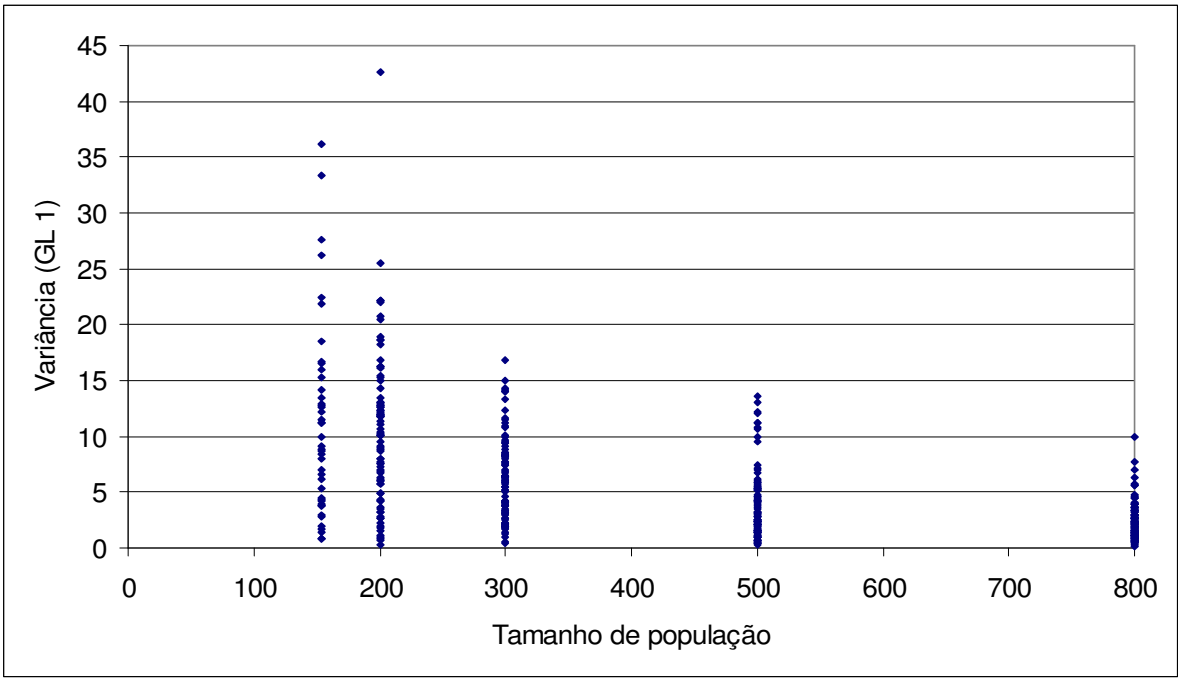


Figura 30 - Dispersão dos valores de variância das distâncias entre marcas adjacentes no grupo de ligação 1 em função do tamanho de população. Avaliação feita nas populações com 154, 200, 300, 500 e 800 indivíduos, simuladas a partir do genoma com nível de saturação de 20 cM.

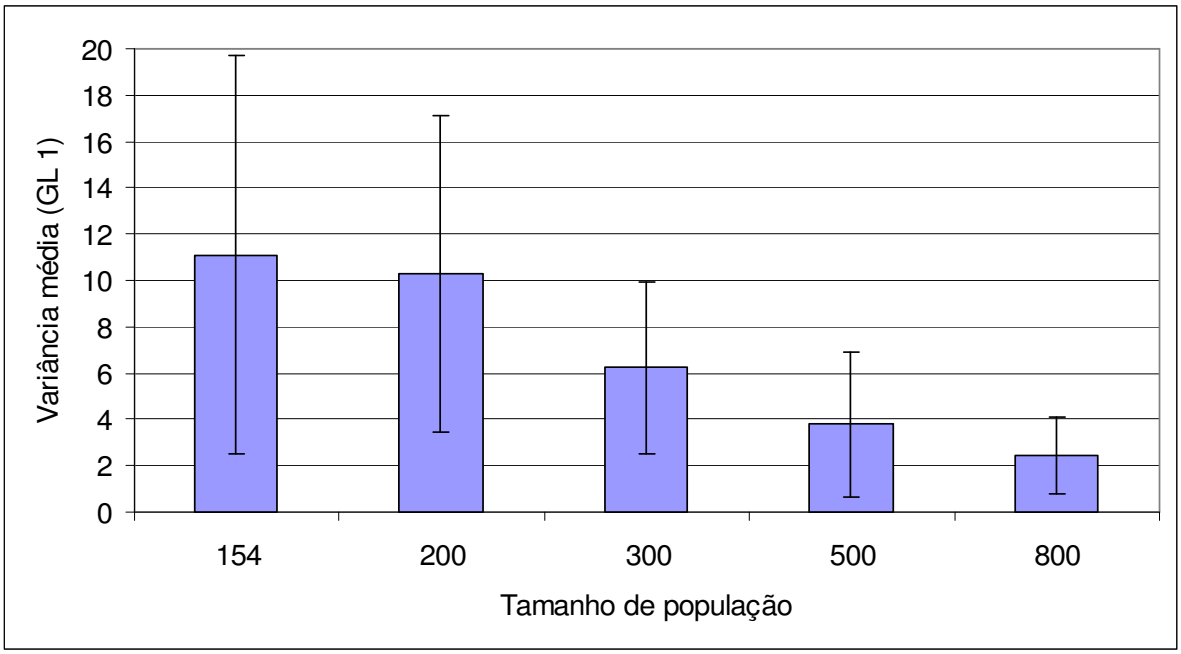


Figura 31 - Variância média das distâncias entre marcas adjacentes no grupo de ligação 1 e seu desvio padrão (linha vertical no topo de cada barra) em função do tamanho de população. Avaliação feita nas populações com 154, 200, 300, 500 e 800 indivíduos, simuladas a partir do genoma com nível de saturação de 20 cM.

Analogamente ao obtido para os níveis de saturação de genoma de 5 e 10 cM, os valores de variância para o nível de saturação de 20 cM reduziram com o aumento do tamanho de população, o que é evidenciado pelo teste de comparação de médias. Valores menores de variância estão associados às maiores populações, e maiores valores estão associados às menores populações, em todos os grupos de ligação. Este comportamento pode ser também observado pela análise da média geral (última coluna do Quadro 16), na qual as médias reduziram significativamente com o aumento do tamanho de população.

O nível de saturação do genoma por marcadores moleculares afetou os valores de variância média das distâncias entre marcas adjacentes. À medida que se reduziu o nível de saturação do genoma por marcas moleculares aumentaram-se os valores de variância média. Por exemplo, para o tamanho de população de 154 indivíduos os valores de variância média foram de 2,08, 5,10 e 11,10, para os níveis de saturação de 5, 10 e 20 cM, respectivamente. Os valores de desvio padrão também seguiram a mesma tendência, ou seja, para um mesmo tamanho de população aumentaram com a redução do nível de saturação do genoma.

4.7 Estresse

O valor de estresse expressa o grau de concordância dos valores de distância entre cada par de marcas adjacentes no grupo de ligação obtido no mapeamento das populações simuladas com as distâncias dos respectivos pares de marcas no genoma utilizado para simulação das populações. O procedimento para a obtenção da estimativa de estresse é apresentado no Material e Métodos (item 3.3.6).

Previamente à apresentação e discussão dos resultados referentes ao trabalho propriamente dito, faz-se necessário um esclarecimento a respeito do significado das estimativas de estresse em cada nível de saturação do genoma. A seguir está apresentado o procedimento utilizado para a obtenção dos valores de estresse em função do desvio médio num genoma utilizado para simulação com saturação de 5 cM.

Tomaremos o grupo de ligação com nível de saturação de 5 cM (Figura 32a) como o genoma utilizado para simulação de uma população segregante e, o grupo de ligação apresentado na Figura 32b como o grupo formado pelo mapeamento da população segregante simulada.

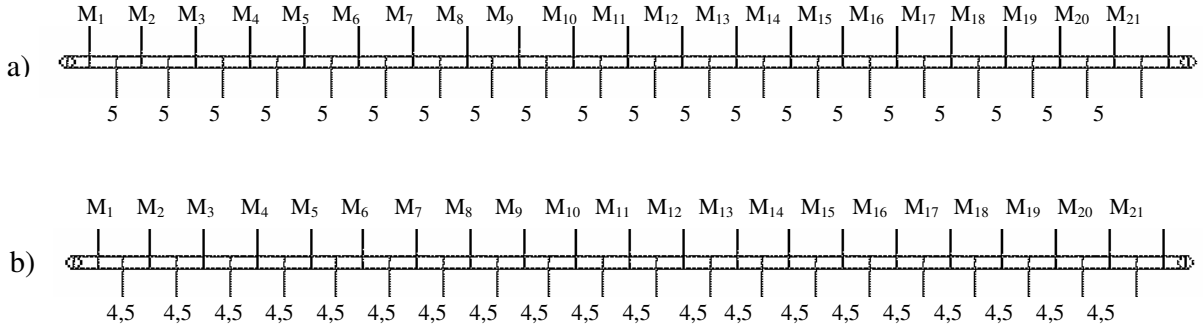


Figura 32 - (a) Grupo de ligação hipotético do genoma utilizado para simulação da população segregante, nível de saturação de 5 cM. (b) Grupo de ligação hipotético formado após o mapeamento da população segregante simulada.

De acordo com o apresentado no Material e métodos (item 3.3.6), a expressão que estima o estresse é:

$$S = 100.\sqrt{\frac{\sum_{k=1}^{m-1}(d_{ok} - d_k)^2}{\sum_{k=1}^{m-1} d^2_{ok}}} \qquad \text{Eq.(33)}$$

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

O termo $(d_{ok} - d_k)$ da expressão de estresse expressa os valores de desvio entre as distâncias no grupo de ligação utilizado para simulação e as distâncias no grupo de ligação obtido no mapeamento da população simulada, para cada intervalo de marcas adjacentes. Dado, que as distâncias dentro de ambos os grupos de ligação são constantes, isto é, no grupo de ligação do genoma utilizado para simulação todas as distâncias entre pares de marcas adjacentes são de 5 cM e, no grupo de ligação obtido no mapeamento da população simulada todas as distâncias são de 4,5 cM, os valores de desvio para cada intervalo de marcas adjacentes são de 0,5 cM ($d_{ok} - d_k = 5 - 4{,}5 = 0{,}5 = d_j$), ou seja, uma constante. Considerando que os desvios são constantes, a média aritmética destes desvios é igual ao próprio desvio, ou seja:

$\bar{d} = \frac{\sum_{j=1}^{m-1} d_j}{m-1} = d_j$; o qual será denotado por desvio médio ($\bar{d}$). Eq.(34)

Substituindo-se o desvio médio na expressão de estresse, tem-se:

$$S = 100.\sqrt{\sum_{k=1}^{m-1}(\bar{d})^2 \Big/ \sum_{k=1}^{m-1} d^2_{ok}} \qquad \text{Eq.(35)}$$

Como os valores de desvio médio ($\bar{d}$) e das distâncias entre marcas adjacentes (d_{ok}) são constantes, pode-se simplificar a expressão acima, como segue:

$$S = 100.\sqrt{(m-1)\left(\overline{d}\right)^2 / (m-1)d^2{}_{ok}} \qquad \text{Eq.(36)}$$

Considerando-se os valores de desvio médio em módulo e aplicando-se as devidas simplificações, obtém-se, finalmente, a equação 37, a qual foi utilizada para construção dos gráficos apresentados nas Figuras 33, 34 e 35 assumido-se diferentes valores de desvio médio e saturação do genoma utilizado para simulação de 5, 10 e 20 cM (d_{ok}).

$$S = 100.\frac{\overline{d}}{d_{ok}} \qquad \text{Eq.(37)}$$

em que: ($\overline{d}$) é o desvio médio e, (d_{ok}) é distância entre marcas adjacentes m_k e m_{k+1} no grupo de ligação utilizado para simulação.

Finalmente, para o exemplo apresentado na Figura 32, o valor de desvio médio é de 0,5 cM, o qual, para o genoma com saturação inicial de 5 cM proporcionará um valor de estresse de 10% (Figura 33).

$$S = 100.\frac{0{,}5}{5} = 10\%$$

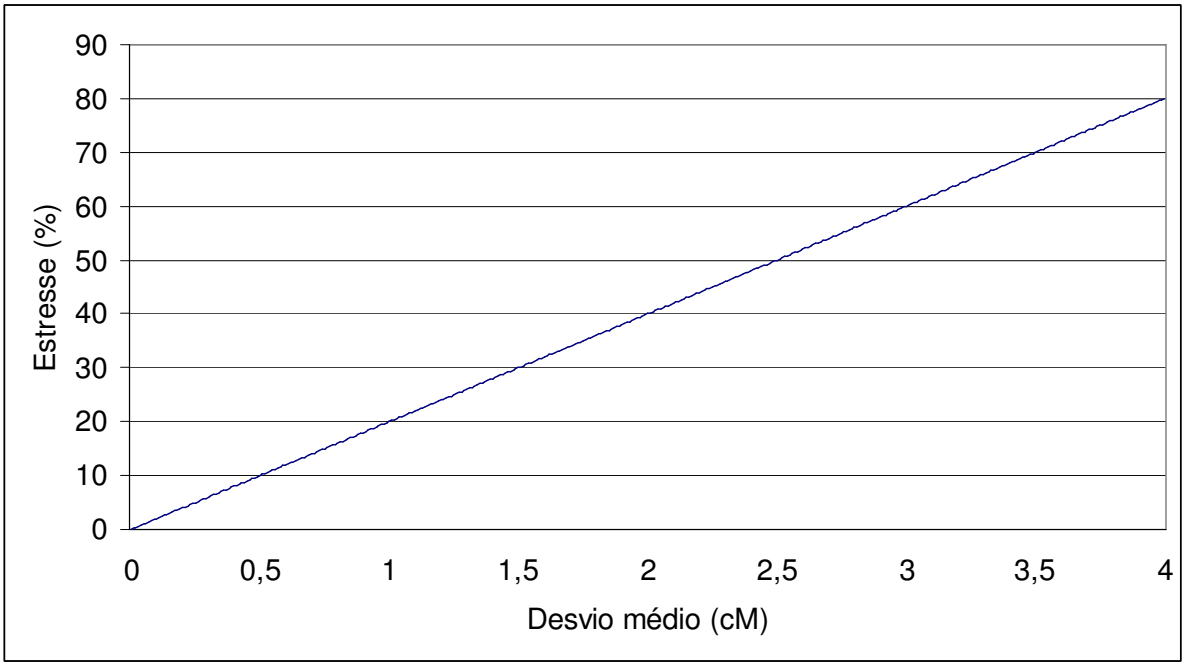


Figura 33 - Valores de estresse (%) expressos em função de desvios médios teóricos das distâncias entre marcas no mapa em relação às distâncias originais no genoma. As distâncias originais das marcas são de 5 cM. Um desvio médio de 0,5 cM implica que as distâncias entre marcas no mapa são de 4,5 cM.

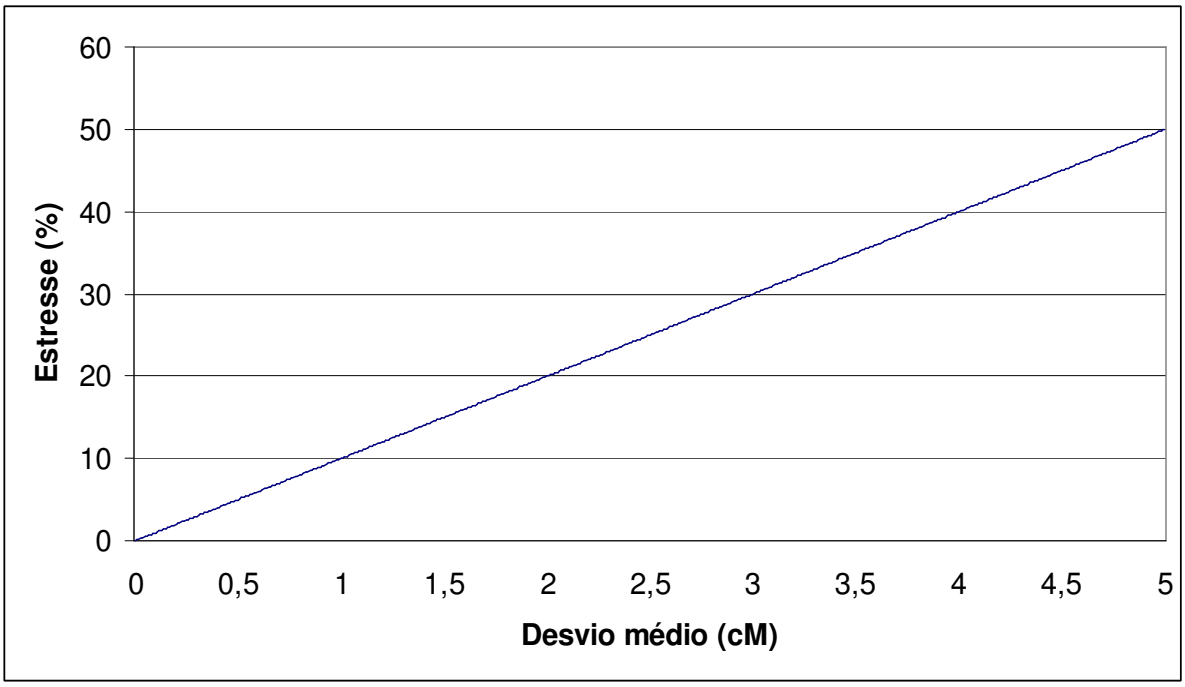


Figura 34 - Valores de estresse (%) expressos em função de desvios médios teóricos das distâncias entre marcas no mapa em relação às distâncias originais no genoma. As distâncias originais das marcas são de 10 cM. Um desvio médio de 0,5 cM implica que as distâncias entre marcas no mapa são de 9,5 cM.

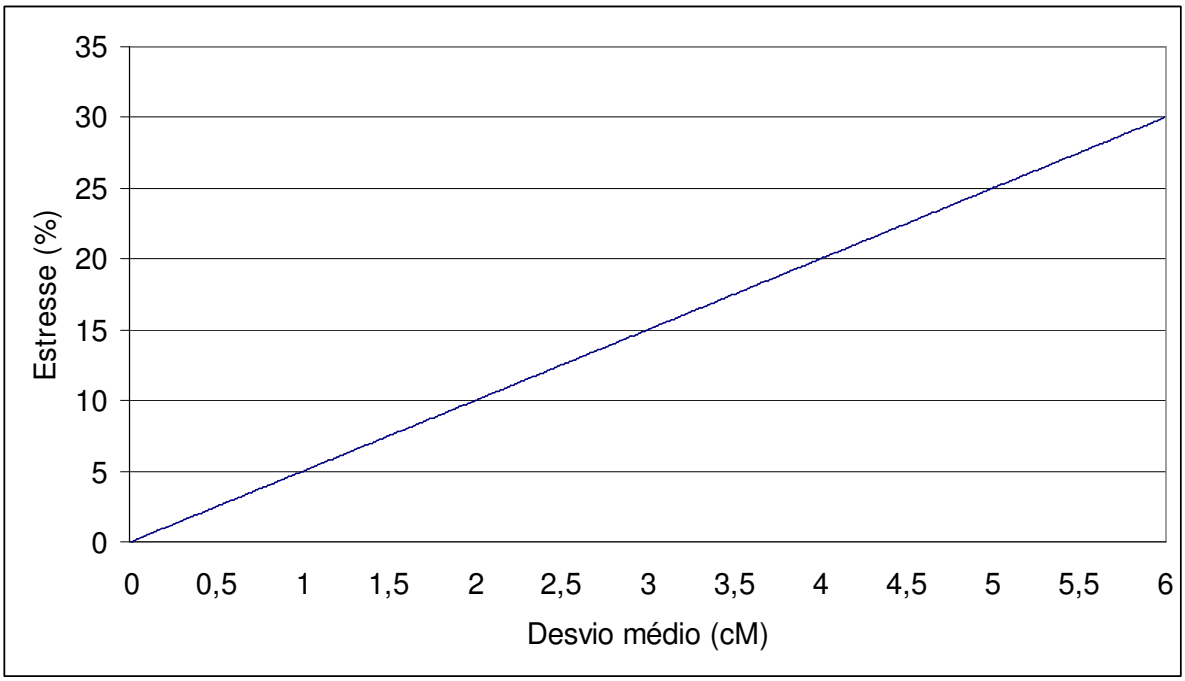


Figura 35 - Valores de estresse (%) expressos em função de desvios médios teóricos das distâncias entre marcas no mapa em relação às distâncias originais no genoma. As distâncias originais das marcas são de 20 cM. Um desvio médio de 0,5 cM implica que as distâncias entre marcas no mapa são de 19,5 cM.

O valor de estresse possui significados diferentes quando se comparam genomas com níveis de saturação distintos. Por exemplo, para uma mesma estimativa de estresse igual a 20%, o desvio médio referente a este estresse pode assumir valores de 1, 2 ou 4 cM, se estiver associado ao genoma com nível de saturação de 5, 10 ou 20 cM, respectivamente (Figuras 33, 34 e 35, ou de acordo com a Equação 37). Portanto, o leitor, quando da análise dos resultados apresentados neste trabalho deve ficar atento ao significado das estimativas de estresse quando da comparação de populações segregantes simuladas a partir de genomas com níveis diferentes de saturação.

Os valores de estresse médio para cada grupo de ligação estão apresentados no Quadro 17. O estresse médio é uma média aritmética dos valores de estresse obtidos nas repetições em que houve a formação de 11 grupos de ligação no mapeamento. Cabe aqui, ressaltar, que de agora em diante os valores de estresse apresentados foram obtidos pela equação apresentada no Material e métodos (item 3.3.6), não podendo ser empregada a equação 37, uma vez que os valores de desvio $(d_{ok} - d_k)$ não foram constantes dentro dos grupos de ligação formados no mapeamento das populações simuladas, diferentemente do apresentado na Figura 32b.

Analisando-se o efeito do tamanho de população dentro do nível de saturação do genoma de 5 cM nota-se uma redução dos valores de estresse médio à medida que se aumenta o tamanho da população segregante, para todos os grupos de ligação. Por exemplo, para o grupo de ligação 1 tem-se que o valor de estresse médio foi de 48,06 e 12,99 %, para os tamanhos de população de 50 e 800 indivíduos, respectivamente (Quadro 17). Estimativas de estresse médio de 48,06 e 12,99 % implicam, em termos de desvio médio, em valores de 2,40 e 0,64 cM, respectivamente (Eq. 37 ou Figura 33). Além, da redução dos valores de estresse médio, também, ocorreu uma redução da amplitude de variação dos valores de estresse entre repetições à medida que se aumentou o tamanho da população segregante, o que é representado pela redução dos valores de desvio padrão. Demonstrando uma maior confiabilidade quando do mapeamento de populações com maior número de indivíduos. Por exemplo, para o grupo de ligação 1 tem-se que o desvio padrão foi de 9,39 e 2,42 %, para os tamanhos de população de 50 e 800 indivíduos, respectivamente.

Quadro 17 - Estresse médio e seu desvio padrão em função do tamanho de população e nível de saturação do genoma. Avaliação feita apenas nas repetições em que houve a formação de 11 grupos de ligação no mapeamento das populações simuladas, conforme apresentado no Quadro 7 (Tukey a 1%)

S.G.	T.P.	N.R.	Grupo de ligação											Média [2]
			1	2	3	4	5	6	7	8	9	10	11	
5 cM	50	70	48,06 a (9,39)	51,05 a (9,29)	49,31 a (9,55)	50,60 a (10,31)	48,92 a (8,23)	49,82 a (8,30)	49,76 a (9,09)	49,63 a (9,87)	49,91 a (8,19)	49,31 a (9,15)	51,65 a (9,45)	49,82 a (9,18)
	100	100	35,88 b (7,15)	36,15 b (6,62)	35,62 b (5,80)	35,50 b (6,41)	35,68 b (5,75)	35,25 b (5,49)	34,66 b (6,49)	35,45 b (6,79)	35,73 b (6,02)	36,17 b (6,02)	34,98 b (6,19)	35,55 b (6,25)
	154	100	28,77 c (4,38)	28,77 c (5,10)	28,77 c (4,64)	28,03 c (5,17)	28,79 c (5,09)	28,87 c (5,13)	29,13 c (4,99)	28,83 c (5,40)	28,72 c (4,43)	28,23 c (4,12)	27,47 c (4,68)	28,58 c (4,84)
	200	100	24,62 d (4,45)	25,15 d (4,06)	25,18 d (3,88)	24,65 d (4,55)	24,31 d (4,59)	25,29 d (4,71)	25,06 d (4,47)	26,34 c (3,98)	25,60 d (4,38)	24,81 d (4,05)	25,07 c (4,34)	25,10 d (4,33)
	300	100	19,56 e (3,20)	20,12 e (3,28)	20,83 e (3,22)	20,41 e (3,00)	20,43 e (3,41)	20,40 e (3,64)	20,26 e (3,18)	20,58 d (3,72)	20,51 e (3,80)	20,57 e (3,63)	20,51 d (3,48)	20,38 e (3,42)
	500	100	15,64 f (2,85)	15,68 f (2,67)	15,62 f (2,39)	16,17 f (2,54)	15,76 f (2,32)	15,96 f (2,77)	15,88 f (2,80)	15,59 e (2,56)	16,08 f (2,54)	15,93 f (2,56)	15,69 e (2,69)	15,82 f (2,62)
	800	100	12,99 g (2,42)	12,33 g (2,02)	12,55 g (2,03)	12,65 g (2,41)	12,76 g (2,00)	12,69 g (2,01)	12,86 g (2,07)	12,90 f (2,13)	12,47 g (1,81)	13,08 g (2,28)	12,56 f (2,11)	12,71 g (2,13)
10 cM	100	98	27,49 a (7,00)	27,60 a (6,95)	28,36 a (7,04)	28,15 a (6,44)	27,48 a (6,63)	28,12 a (6,13)	27,32 a (5,53)	27,76 a (6,33)	27,00 a (6,49)	27,42 a (6,07)	26,51 a (6,78)	27,57 a (6,49)
	154	100	22,15 b (5,56)	21,20 b (5,27)	21,44 b (5,28)	21,51 b (4,80)	22,34 b (4,76)	21,55 b (5,52)	22,23 b (6,29)	22,84 b (6,09)	22,70 b (5,64)	21,50 b (4,80)	22,12 b (5,49)	21,96 b (5,42)
	200	100	19,21 c (4,90)	20,30 b (5,23)	19,66 b (4,34)	18,71 c (4,77)	19,18 c (4,74)	19,29 c (4,50)	18,57 c (4,42)	18,95 c (3,93)	19,33 c (4,70)	19,35 c (5,01)	19,08 c (5,27)	19,24 c (4,72)
	300	100	16,16 d (4,15)	15,67 c (3,03)	16,19 c (3,63)	15,57 d (3,85)	15,36 d (4,41)	16,61 d (3,66)	16,55 c (3,34)	15,65 d (3,56)	15,54 d (3,20)	15,38 d (3,59)	15,71 d (3,78)	15,85 d (3,68)
	500	100	12,12 e (2,73)	12,71 d (2,75)	12,23 d (2,56)	12,00 e (2,78)	12,11 e (3,12)	11,93 e (2,89)	12,43 d (2,98)	11,96 e (2,63)	12,16 e (3,22)	12,33 e (2,88)	12,10 e (2,91)	12,19 e (2,86)
	800	100	9,53 f (2,37)	9,61 e (2,31)	10,18 d (2,54)	9,88 f (2,37)	9,63 f (2,61)	9,87 e (2,30)	9,82 e (2,27)	9,74 f (2,39)	10,06 e (2,46)	9,41 f (2,16)	9,37 f (2,19)	9,74 f (2,36)
20 cM	154	43	15,57 a (5,45)	14,64 ab (5,02)	15,59 a (4,21)	16,34 a (4,31)	14,80 ab (4,15)	15,64 a (4,26)	16,76 a (3,71)	14,80 a (4,83)	14,76 a (4,09)	15,22 a (4,60)	15,14 a (3,62)	15,39 a (4,34)
	200	86	15,29 a (5,10)	15,40 a (4,69)	15,53 a (4,79)	15,65 a (4,45)	14,83 a (4,78)	14,66 a (4,54)	14,75 a (5,08)	14,93 a (4,09)	14,70 a (5,26)	14,90 a (4,04)	14,54 a (4,45)	15,02 a (4,66)
	300	98	11,61 b (3,51)	12,84 b (4,81)	12,09 b (3,81)	12,12 b (3,86)	12,55 b (4,46)	12,12 b (3,74)	12,16 b (4,19)	12,33 b (3,74)	13,46 a (4,26)	12,60 b (4,02)	12,78 b (4,14)	12,42 b (4,07)
	500	100	9,06 c (3,35)	9,25 c (2,71)	9,26 c (2,94)	9,75 c (2,96)	9,28 c (2,99)	9,78 c (2,70)	9,00 c (3,14)	9,53 c (3,29)	9,32 b (3,05)	9,20 c (3,01)	9,23 c (3,10)	9,33 c (3,02)
	800	100	7,56 c (2,19)	7,14 d (2,42)	7,70 c (2,52)	7,60 d (2,52)	7,36 d (2,70)	7,61 d (2,36)	7,74 c (2,72)	7,85 d (2,51)	7,71 c (2,49)	7,75 c (2,73)	7,73 c (2,50)	7,61 d (2,52)

[1]\ Entre parênteses estão os valores de desvio padrão; [2]\ Média dos onze grupos de ligação.

S.G.= Saturação do genoma; T.P.= Tamanho da população; N.R.= número de repetições avaliadas.

Médias seguidas por letras iguais nas colunas, dentro de cada nível de saturação do genoma, não diferem entre si pelo teste de Tukey a 1% de probabilidade (P<0,01).

Uma melhor visualização do comportamento dos valores de estresse e estresse médio com o aumento do número de indivíduos pode ser observada nas Figuras 36 e 37, respectivamente, que, apesar de representarem apenas o comportamento observado para o grupo de ligação 1 é representativo do ocorrido para todos os grupos de ligação analisados dentro do nível de saturação do genoma de 5 cM. De acordo com os valores de estresse apresentados na Figura 36 fica evidente o efeito da redução da amplitude de variação dos valores de estresse com o aumento do tamanho das populações segregantes. Esta redução também pode ser visualizada na Figura 37 pela análise do comportamento dos valores de desvio padrão apresentado nos topos de cada barra.

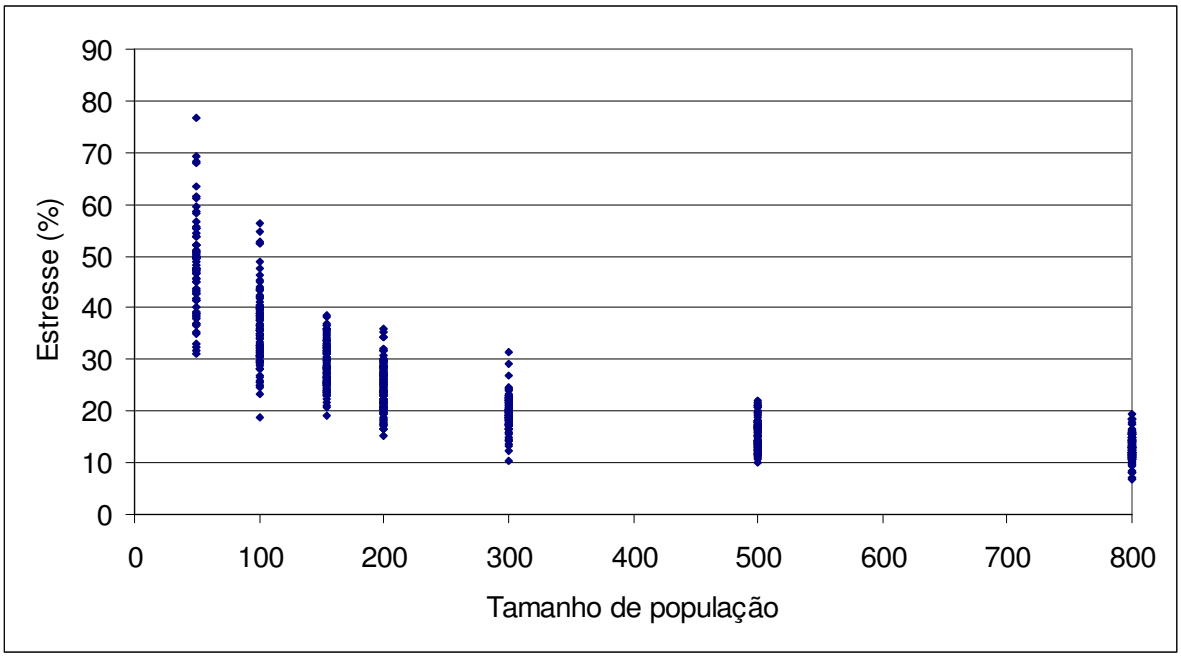


Figura 36 - Dispersão dos valores de Estresse para o grupo de ligação 1 em função do tamanho de população. Avaliação feita nas populações com 50, 100, 154, 200, 300, 500 e 800 indivíduos, simuladas a partir do genoma com nível de saturação de 5 cM.

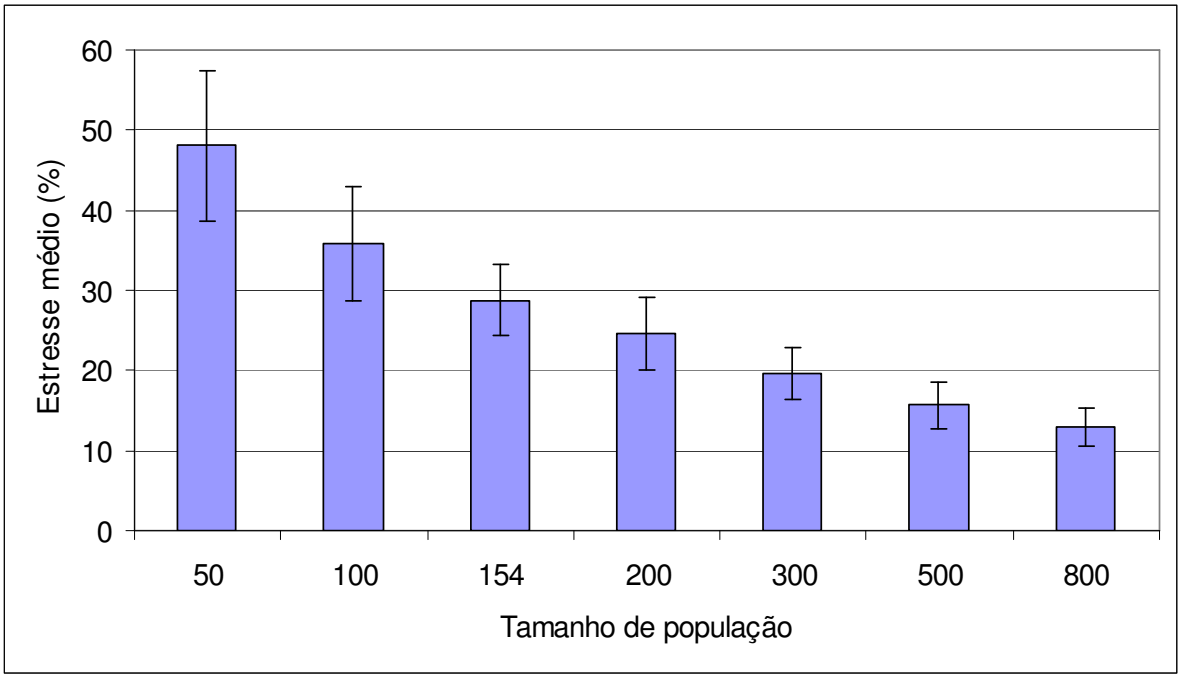


Figura 37 - Estresse médio para o grupo de ligação 1 e seu desvio padrão (linha vertical no topo de cada barra) em função do tamanho de população. Avaliação feita nas populações com 50, 100, 154, 200, 300, 500 e 800 indivíduos, simuladas a partir do genoma com nível de saturação de 5 cM.

A redução dos valores de estresses com o aumento do tamanho das populações dentro do nível de saturação do genoma é evidenciada pelas diferenças significativas obtidas com a utilização do teste de comparação de médias em todos os grupos de ligação, onde os menores valores estão associados as maiores populações. O efeito do aumento do tamanho de população na redução do estresse é claramente demonstrado pela análise das médias gerais (última coluna do Quadro 17), na qual as menores médias associadas as menores populações diferem significativamente das maiores médias associadas as menores populações.

Para os vários tamanhos de populações segregantes analisados dentro do nível de saturação do genoma de 10 cM verifica-se, assim como para o nível de saturação de 5 cM, uma redução dos valores de estresse médio à medida que se aumenta o tamanho das populações segregantes para todos os grupos de ligação. Por exemplo, para o grupo de ligação 1 tem-se que o valor de estresse médio foi de 27,49 e 9,53 %, para os tamanhos de população de 100 e 800 indivíduos, respectivamente (Quadro 17). Estimativas de estresse médio de 27,49 e 9,53 % implicam, em termos de desvio médio, em valores de 2,74 e 0,95 cM, respectivamente (Eq. 37 ou Figura 34). Além, da redução dos valores de estresse médio, ocorreu, também, uma redução da amplitude de variação dos valores de estresse nas

repetições à medida que se aumentou o tamanho da população segregante, o que é representado pela redução nos valores de desvio padrão. Por exemplo, para o grupo de ligação 1 tem-se que o desvio padrão foi de 7,00 e 2,37 %, para os tamanhos de população de 100 e 800 indivíduos, respectivamente.

Uma melhor visualização do comportamento dos valores de estresse e estresse médio com o aumento do número de indivíduos pode ser observada nas Figuras 38 e 39, que apesar de representarem apenas o comportamento observado para o grupo de ligação 1, é representativo do ocorrido para todos os grupos de ligação analisados dentro do nível de saturação do genoma de 10 cM. De acordo com os valores de estresse apresentados na Figura 38, fica evidente o efeito da redução da amplitude de variação dos valores de estresse com o aumento do tamanho das populações segregantes. Esta redução também pode ser visualizada na Figura 39, pela análise do comportamento dos valores de desvio padrão apresentado nos topos de cada barra.

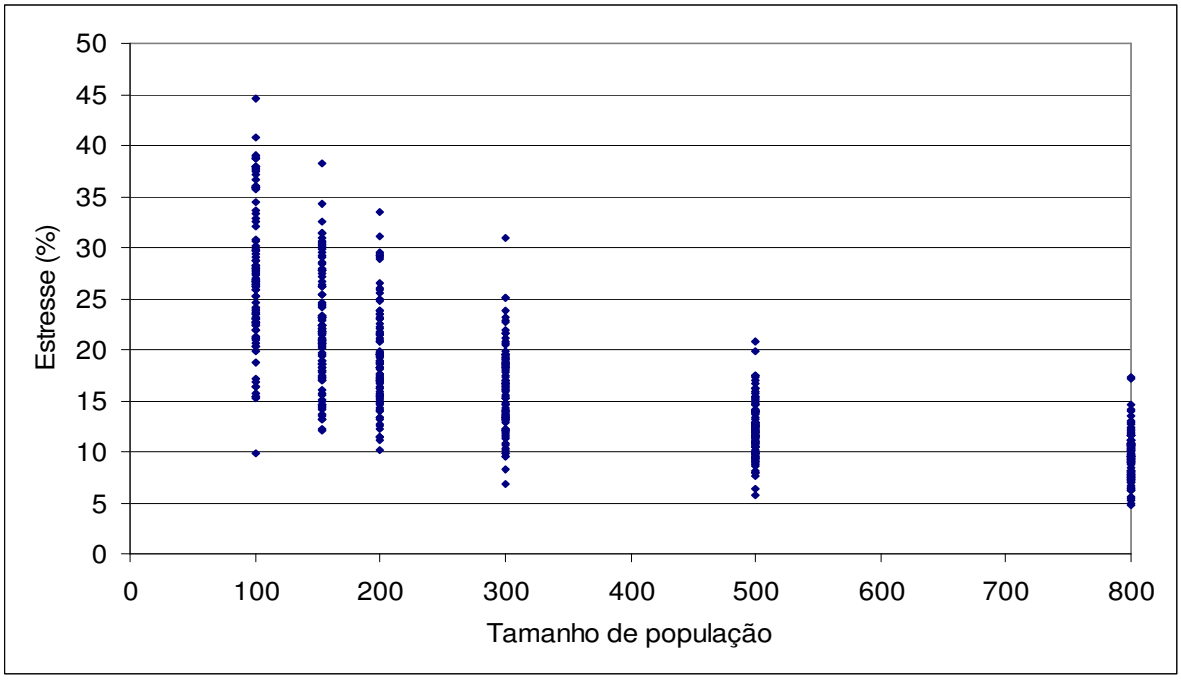


Figura 38 - Dispersão dos valores de Estresse para o grupo de ligação 1 em função do tamanho de população. Avaliação feita nas populações com 100, 154, 200, 300, 500 e 800 indivíduos, simuladas a partir do genoma com nível de saturação de 10 cM.

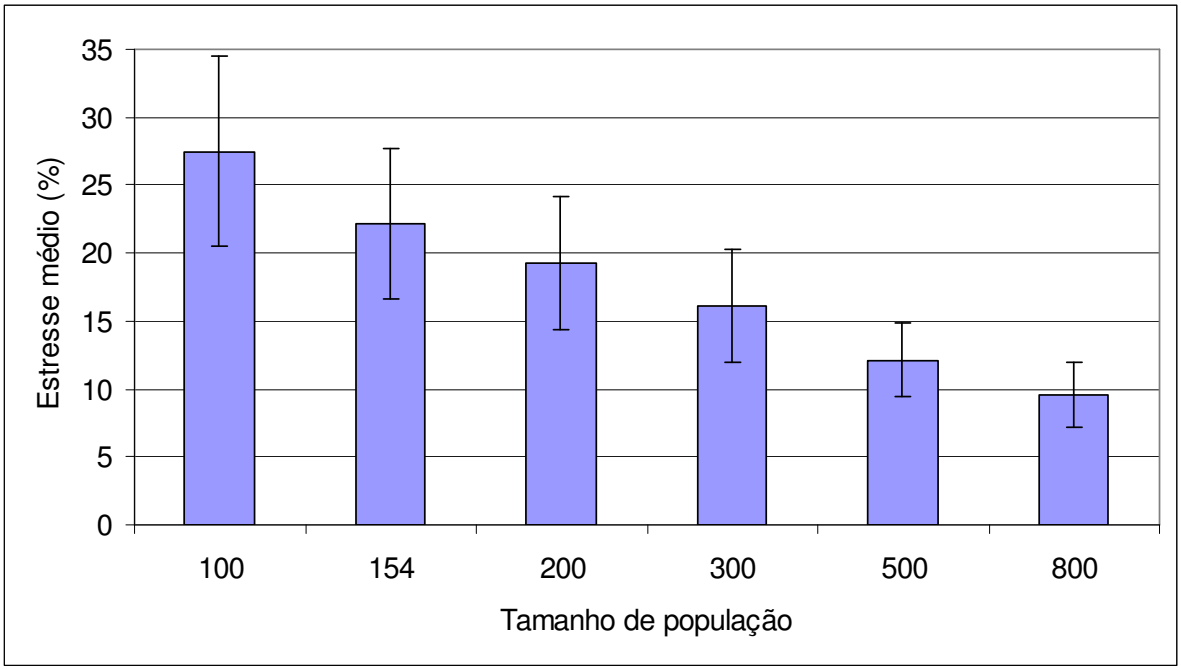


Figura 39 - Estresse médio para o grupo de ligação 1 e seu desvio padrão (linha vertical no topo de cada barra) em função do tamanho de população. Avaliação feita nas populações com 100, 154, 200, 300, 500 e 800 indivíduos, simuladas a partir do genoma com nível de saturação de 10 cM.

O efeito do tamanho de população dentro do nível de saturação do genoma de 10 cM na redução dos valores de estresse, assim com observado para o nível de saturação de 5 cM, é evidenciado pelo teste de média, em que os maiores valores de média estão significativamente associados as menores populações, enquanto os menores valores estão associados às maiores populações, em todos os grupos de ligação. A análise da média geral (última coluna do Quadro 17) permite uma melhor visualização da redução dos valores de estresse com o aumento do tamanho de população utilizado para o mapeamento.

Nas populações geradas a partir do genoma com nível de saturação de 20 cM os valores de estresse médio seguiram o mesmo comportamento das populações geradas a partir dos genomas com 5 e 10 cM, ou seja, houve uma redução das estimativas de estresse médio à medida que se aumentou o tamanho da populações segregantes para todos os grupos de ligação. Exceção a essa redução foi observado apenas para os grupos de ligação 2, 5, e 8 em referência ao aumento do tamanho de população de 154 para 200, nos quais, ao invés de redução houve aumento do valor de estresse, porém, este aumento não foi significativo pelo teste de média.

Como exemplo da redução dos valores de estresse no nível de saturação do genoma de 20 cM, tem-se que para o grupo de ligação 1 o valor de estresse médio foi

de 15,57 e 7,56 %, para os tamanhos de população de 154 e 800 indivíduos, respectivamente (Quadro 17). Estimativas de estresse médio de 15,57 e 7,56 % implicam, em termos de desvio médio, em valores de 3,11 e 1,51 cM, respectivamente (Eq. 37 ou Figura 35). Além, da redução dos valores de estresse médio, ocorreu, também, uma redução da amplitude de variação dos valores de estresse nas repetições à medida que se aumentou o tamanho da população segregante. Por exemplo, para o grupo de ligação 1 tem-se que o desvio padrão foi de 5,45 e 2,19 %, para os tamanhos de população de 154 e 800 indivíduos, respectivamente.

O comportamento dos valores de estresse e estresse médio com o aumento do número de indivíduos pode ser observado nas Figuras 40 e 41, respectivamente, que apesar de representarem apenas o comportamento observado para o grupo de ligação 1, é representativo do ocorrido para todos os grupos de ligação analisados dentro do nível de saturação do genoma de 20 cM. De acordo com os valores de estresse apresentados na Figura 40, fica evidente o efeito da redução da amplitude de variação dos valores de estresse com o aumento do tamanho das populações segregantes. Esta redução também pode ser visualizada na Figura 41, pela análise do comportamento dos valores de desvio padrão apresentado nos topos de cada barra.

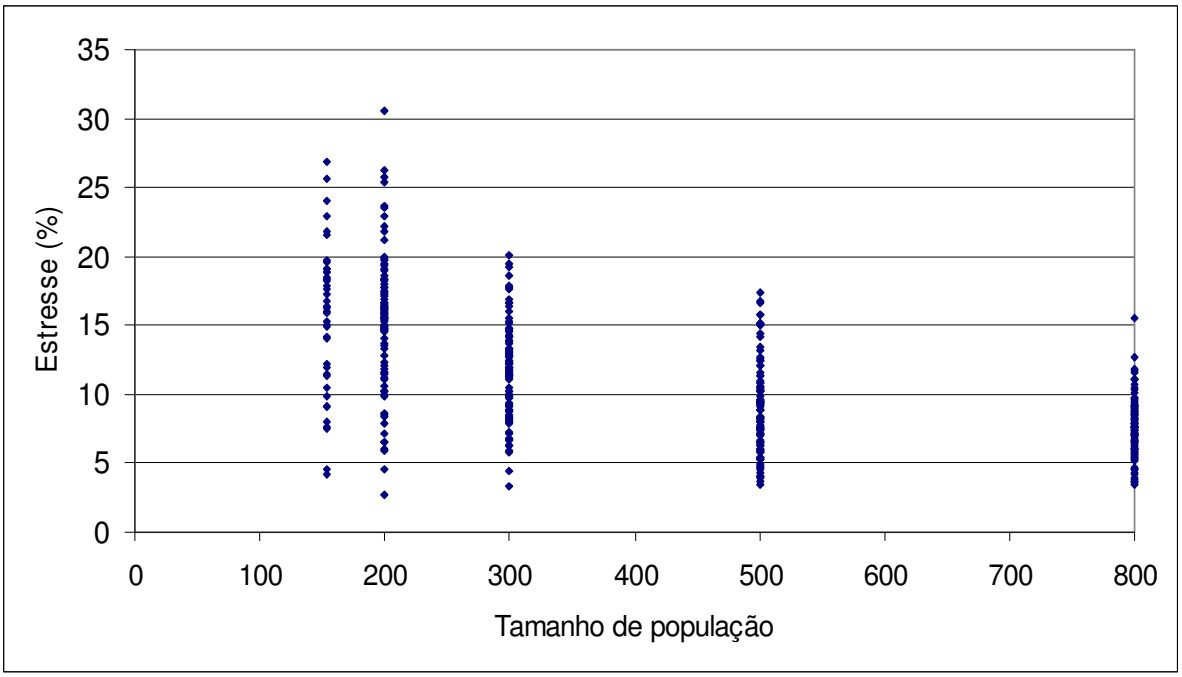


Figura 40 - Dispersão dos valores de Estresse para o grupo de ligação 1 em função do tamanho de população. Avaliação feita nas populações com 154, 200, 300, 500 e 800 indivíduos, simuladas a partir do genoma com nível de saturação de 20 cM.

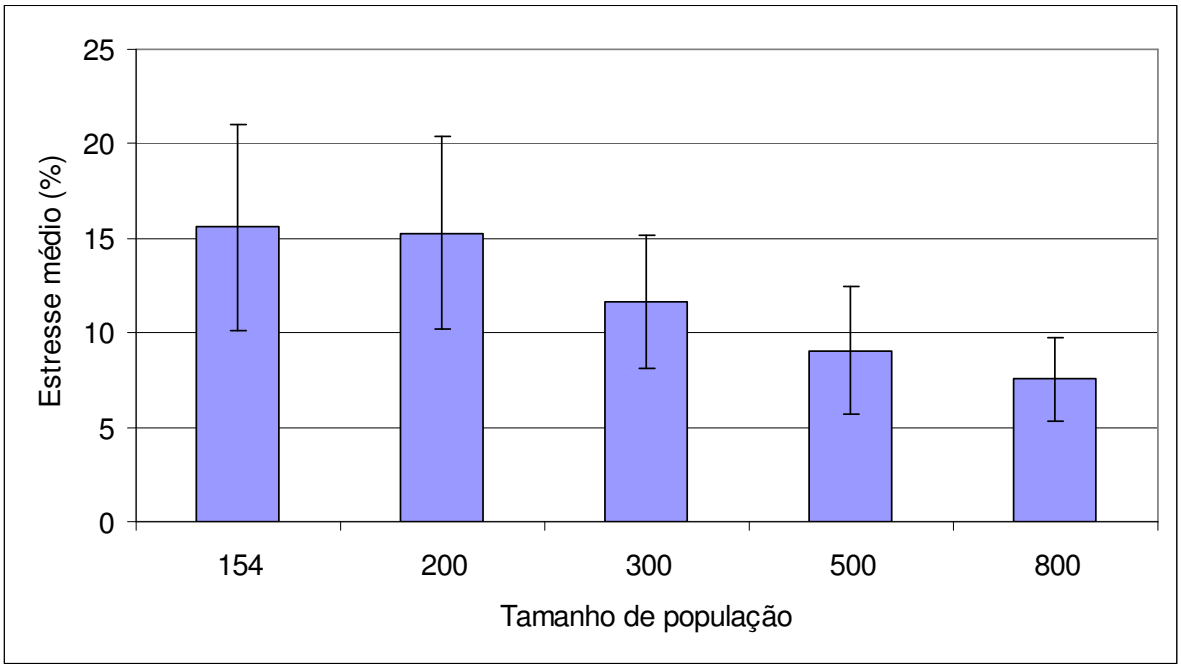


Figura 41 - Estresse médio para o grupo de ligação 1 e seu desvio padrão (linha vertical no topo de cada barra) em função do tamanho de população. Avaliação feita nas populações com 154, 200, 300, 500 e 800 indivíduos, simuladas a partir do genoma com nível de saturação de 20 cM.

Como discutido anteriormente, a comparação dos valores de estresse entre populações com genomas de níveis diferenciados de saturação deve ser cautelosa. No Quadro 18 e Figura 42 estão apresentados os valores de estresse médio de acordo com o apresentado na última coluna do Quadro 17 e seus respectivos significados em termos de desvio médio aproximado de acordo com a Equação 37. Dentro de um mesmo tamanho de população os valores de estresse são maiores para os genomas mais saturados, como exemplo, para o tamanho de população de 154 indivíduos os valores são de 28,58, 21,96 e 15,39%, para os níveis de saturação de 5, 10 e 20 cM, respectivamente. Contudo, quando os valores de estresse são decodificados em valores de desvio médio aproximados, verifica-se que os menores valores de desvio estão associados aos genomas mais saturados, o que pode ser facilmente visualizado na Figura 42. Como exemplo, os valores de estresse citados acima implicam em desvios médios de 1,42, 2,19 e 3,07 cM para os genomas de saturação de 5, 10 e 20 cM, respectivamente.

Quadro 18 - Valores de desvio médio aproximados (cM) dados em função do estresse médio e tamanho de população nos níveis de saturação do genoma de 5, 10 e 20 cM

Saturação do genoma	Tamanho de população segregante							
	Parâmetros	50	100	154	200	300	500	800
5 cM	Estresse médio (%)[1]	49,82	35,55	28,58	25,10	20,38	15,82	12,71
	Desvio médio (cM) [2]	2,49	1,77	1,42	1,25	1,01	0,79	0,63
10 cM	Estresse médio (%)	-	25,57	21,96	19,24	15,85	12,19	9,74
	Desvio médio (cM)	-	2,55	2,19	1,92	1,58	1,21	0,97
20 cM	Estresse médio (%)	-	-	15,39	15,02	12,42	9,33	7,61
	Desvio médio (cM)	-	-	3,07	3,0	2,48	1,86	1,52

[1]\ Valores de estresse médio apresentados na última coluna do Quadro 17;

[2]\ Estimativas de desvio médio dados pela Equação (37) [$S = 100.\left(\bar{d} / d_{ok}\right)$].

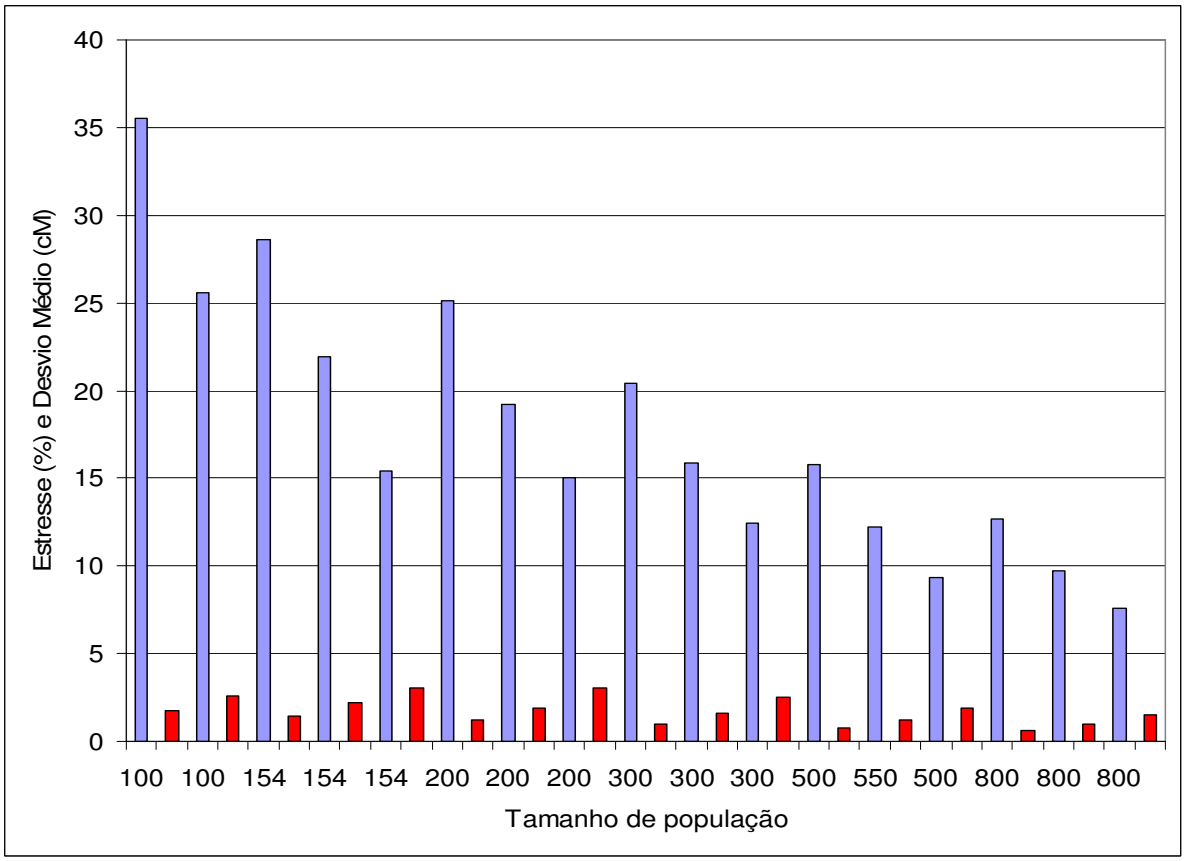


Figura 42 – Valores de Estresse em percentagem (barras azuis) e Desvio Médio aproximado em cM (barras vermelhas) em função dos tamanhos de população e níveis de saturação do genoma por marcadores moleculares. Para os dois tamanhos de população de 100 indivíduos, o primeiro da esquerda para a direita representa o genoma com saturação de 5 cM e o segundo representa o nível de saturação de 10 cM. Para os demais casos onde há três populações de tamanhos iguais, o primeiro, o segundo e o terceiro da esquerda para a direita representam os níveis de saturação do genoma de 5, 10 e 20 cM, respectivamente.

A disponibilidade de mapas genéticos fidedignos depende do número de marcas moleculares e número de indivíduos analisados. Porém, o tamanho de população RIL, assim como o número mínimo de marcas para representar os cromossomos nos grupos de ligação ainda não haviam sido alvo específico de estudos, sendo, determinado, primordialmente pela disponibilidade de material e recursos. Como resultado, os mapas disponíveis na literatura foram construídos, cada um, com um número diferente de indivíduos e marcas. Portanto, sem atenção necessária aos efeitos prejudiciais causados pelo número insuficiente de indivíduos e marcas utilizados no mapeamento, conforme demonstrado neste trabalho de simulação. Todavia, espera-se, que este trabalho contribua para um melhor entendimento dos efeitos do número de indivíduos e marcas no mapeamento, assim como para fornecer subsídios para a escolha adequada destas variáveis quando da condução de futuros trabalhos.

5 CONCLUSÕES

A partir da simulação de três genomas com diferentes níveis de saturação e sua análise e comparação com 2.100 genomas foi possível concluir que:

O tamanho da população e o número de marcas determinam a confiabilidade do mapa de ligação obtido;

Mapas com sérias distorções são obtidos com o uso de populações de tamanhos insuficientes, mesmo com grande quantidade de marcas;

Mapas com sérias distorções são obtidos com o uso de pequena quantidade de marcas, mesmo com populações com grande número de indivíduos;

Tamanhos mínimos de 100, 154 e 500 indivíduos foram necessários para a obtenção de mapas com o mesmo número de marcas por grupo de ligação do genoma original, nos casos de saturação alta (distância média de marcas de 5 cM), mediana (distância média de marcas de 10 cM) e baixa (distância média de marcas de 20 cM), respectivamente. Com o aumento do grau de saturação do mapa, pode-se reduzir substancialmente o número de indivíduos a ser genotipado em RILs;

Idealmente, o tamanho mínimo de população para a obtenção de mapas com boa confiabilidade deveria ser de 200, 300 e 500 indivíduos nos casos de saturação alta (distância média de marcas de 5 cM), mediana (distância média de marcas de 10 cM) e baixa (distância média de marcas de 20 cM), respectivamente.

6 REFERÊNCIAS BIBLIOGRÁFICAS

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

BALLVORA, A., HESSELBACH, J., NIEWOHNER, J., LEISTER, D., SALAMINI, F., GEBHARDT, C. Marker enrichment and high-resolution map of the segment of potato chromosome VII harbouring the nematode resistance gene. Molecular and General Genetics, v. 249, p. 82-90, 1995.

BOREVITZ, J.O., MALLOF, J.N., LUTES, J., DABI, T., REFERN, J.L., TRAINER, G.T., WENER, J.D., ASAMI, T., BERRY, C.C., WEIGEL, D., CHORY, J. Quantitative trait loci controlling light and hormone response in two accessions of *Arabidopsis thaliana*. Genetics, v. 160, p. 683-696, 2002.

BURNHAM, K.D., DORRANCE, A.E., VANTOAI, T.T., MARTIN, S.K.St. Quantitative trait loci for partial resistance to *Phytophthora sojae* in soybean. Crop Science, v. 43, p. 1610-1617, 2003.

BURR, B., BURR, F. Recombinant inbreds for molecular mapping in maize: theoretical and practical considerations. Elsevier Science Publishers, Reviews, v. 7, nº 2, p. 55-60. 1991.

BURR, B., BURR, F.A., THOMPSON, K.H., ALBERTSON, M., STUBER, C.W. Gene mapping with recombinant inbreds in maize. Genetics, v. 118, p. 519-526, 1988.

CAI, H.W., MORISHIMA, H. QTL clusters reflect character associations in wild and cultivated rice. Theoretical and Applied Genetics, v. 104, p. 1217-1228, 2002.

CARDINAL, A.J., LEE, M., SHAROPORA, N., WOODMAN-CLIKEMAN, W.L., LONG, M.J. Genetic mapping and analysis of quantitative trait loci for resistance to stalk tunneling by the European corn bores in maize. Crop Science, v. 41, p. 835-845, 2001.

CHANG, S.J.C., DOUBLER, T.W., KILO, V.Y., ABU-THREDEIH, J., PRABHU, R., FREIRE, V., SUTTNER, R., KLEIN, J., SCHMIDT, M.E., GIBSON, P.T., LIGHTFOOT, D.A. Association of loci underlying field resistance to soybean sudden death syndrome (SDS) and cyst nematode (SCN) race 3. Crop Science, v. 37, p. 965-971, 1997.

CHEN, H., WANG, S., XING, Y., XU, C., HAYES, P.M., ZHANG, O. Comparative analyses of genomic locations and race specificities of loci for quantitative resistance to *Pyricularia grisea* in rice and barley. PNAS, v. 100, nº 5, p. 2544-2549, 2003.

CHO, Y., NJITI, V.N., CHEN, X., TRIWATAYAKORN, K., KASSEM, M.A., MEKSEM, K., LIGHTFOOT, D.A., WOOD, A. Quantitative trait loci associated with foliar trigonelline accumulation in *Glycine max* L. Journal of Biomedicine and Biotechnology, 2:3, p. 151-157, 2002.

CLARKE, G. M. Statistics and experimental design: an introduction for biologists and biochemists. Third Edition, New York-Toronto,1994, 280p.

CRUZ, C. D. Programa para análises de dados moleculares e quantitativos – GQMOL. Viçosa: UFV, 2004.(Software em desenvolvimento).

CRUZ, C. D. GENES: Programa de análise e processamento de dados baseados em modelos de genética e estatística experimental. Versão 2004.2.1. Viçosa, UFV.

CRUZ, C.D., CARNEIRO, P. C.S. Modelos biométricos aplicados ao melhoramento genético. Vol. 2, Viçosa: Imprensa Universitária, 2003. 279p.

EUJAYL, I., BAUM, M. POWELL, W., ERSKINE, W., PEHU, E. A genetic linkage map of lentil (*Lens* sp.) based on RAPD and AFLP markers using recombinant inbred lines. Theoretical and Applied Genetics, v. 97, p. 83-89, 1998.

FALEIRO, F.G. Melhoramento e mapeamento genético do feijoeiro-comum: análise de características quantitativas, morfológicas, moleculares e de resistência a doenças. Viçosa, MG: UFV, 2000. 177p. (Dissertação Doutorado). Universidade Federal de Viçosa, 2000.

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

FERREIRA, A.R., FOUTZ, K.R., KEIM, P. Soybean genetic map of RAPD markers assigned to an existing scaffold RFLP map. The American Genetic Association, v. 91, p. 392-396, 2000.

FLOWERS, T.J., KOYAMA, M.L, FLOWERS, S.A., SUDHAKAR, C., SINGH, K.P., YEO, A.R. QTL: their place in engineering tolerance of rice to salinity. Journal of Experimental Botany, v. 51, nº 342, p. 99-106, 2000.

FOOLAD, M.R., ARULSEKAR, S. BECERRA, V., BLISS, F.A. A genetic map of *Prumus* based on an interspecific cross between peach and almond. Theoretical and Applied Genetics, v. 91, p. 262-269, 1995.

GRIFFITHS, A.J.F., MILLER, J.H., SUSUKI, D.T., LEWONTIN, R. C., GELBART, W.M. An introduction to genetic analysis. 6th ed. New York: W.H. Freeman, 1996. 916p.

GROH, S., GONZÁLEZ-DE-LEÓN, D., KHAIRALLAH, M. M., JIANG, C., BERGVINSON, D., BOHN, M., HOISINGTON, D. A., MELCHINGER, A. E. QTL Mapping in Tropical Maize: III. Genomic Regions for Resistance to *Diatraea* spp. and Associated Traits in Two RIL Populations. Crop Science, v. 38, p.1062–1072, 1998.

IQBAL, M.J., MEKSEM, K., NJITI, V.N., KASSEM, M.A., LIGHTFOOT, D.A. Microsatellite markers identify three additional quantitative trait loci for resistance to soybean sudden-death syndrome (SDS) in Essex X Forrest RILs. Theoretical and Applied Genetics, v. 102, p. 187-192, 2001.

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

HNETKOVSKY, N., CHANG, S.J.C., DOUBLER, T.W., GIBSON, P.T., LIGHTFOOT, D.A. Genetic mapping of loci underlying field resistance to soybean sudden death syndrome (SDS). Crop Science, v. 36, p. 393-400, 1996.

KY, C.-L., BARRE, P., LORIEUX, M., TROUSLOT, P., AKAFFOU, S., LOUARN, J., CHARRIER, A., HAMON, S., NOIROT, M. Interspecific genetic linkage map, segregation distortion and genetic conversion in coffee (*Coffea* sp.). Theoretical and Applied Genetics, v. 101, p. 669–676, 2000.

LESPINASSE, D., RODIER-GOUD, M., GRIVET, L., LECONTE, A., LEGNATE, H., SEGUIN, M. A saturated genetic linkage map of rubber tree (*Hevea* spp.) based on RFLP, AFLP, microsatellite, and isozyme markers. Theoretical and Applied Genetics, v. 100, p. 127-138, 2000.

LITT, M., LUTY, J.A. A hypervariable microsatellite revealed by *in vitro* amplification of a dinucleotide repeat within the cardiac muscle actin gene. American Journal Human Genetic, v. 44, p. 398-401, 1989.

LIU, B.H. Statistical genomics, linkage, mapping and QTL analysis. Boca Raton: CRC Press.1988. 611p.

MEKSEM, K., NJITI, V.N., BANZ, W.J., IQBAL, M.J., KASSEM, M.M., HYTEN, D.L., YUANG, J., WINTERS, T.A., LIGHTFOOT, D.A. Genomic regions that underlie soybean seed isoflavone content. Journal of Biomedicine and Biotechnology, 1:1, p. 38-44, 2001.

MEKSEM, K., LEISTER, D., PELEMAN, J., ZABEAU, M., SALAMINI, F., GEBHARDT, C. A high-resolution map of the vicinity of the R1 locus on chromosome V potato based on RFLP and AFLP markers. Molecular and General Genetics, v. 249, p. 74-81, 1995.

MIKLAS, P.N., JOHNSON, W.C., DELORME, R., GEPTS, P. QTL conditioning physiological resistance and avoidance to white mold in dry bean. Crop Science, v. 41, p. 309-315, 2001.

MULLIS, K., FALLONA, F. Specific synthesis of DNA *in vitro* via a polymerase catalyzed chain reaction. Methods in Enzymology. v. 55, p. 335-350, 1987.

NACHIT, M.M., ELOUAFI, I., PAGNOTTA, M.A., SALEH, A.E., IACONO, E., LABHILILI, M., ASBATI, A., AZRAK, M., HAZZAM, H., BENSCHER, D., KHAIRALLAH, M., RIBAUT, J.M., TANZARELLA, O.A., PORCEDDU, E., SORRELLS, M.E. Molecular linkage map for a intraspecific recombinant inbred population of durum wheat (*Triticum turgidum* L. var. durum). Theoretical and Applied Genetics, v. 102, p. 177-186, 2001.

NIKAIDO, A., YOSHIMARU, H., TSUMURA, Y., SUYAMA, Y., MURAI, M., NAGASAKA, K. Segregation distortion for AFLP markers in *Cryptomeria japonica*. Genes Genet Systems, v. 74, p. 55-59, 1999.

NJITI, V.N., MEKSEM, K., IQBAL, M.J., JOHNSON, J.E., KASSEM, M.A., ZOBRIST, K.F., KILO, V.Y., LIGHFOOT, D.A. Common loci underlie field resistance to soybean sudden death syndrome in Forrest, Pyramid, Essex, and Douglas. Theoretical and Applied Genetics, v. 104, p. 294-300, 2002.

OLSON, M., HOOD, L., CANTOR, C., DOSTESTEIN, D. A common language for physical mapping of the human genome. Science, v. 254, p. 1434-1435, 1989.

PARAN, I., MICHELMORE, R.W. Development of reliable PCR-based markers linked to downy mildew resistance genes in lettuce. Theoretical and Applied Genetics, v. 85, p. 985-993, 1993.

PRICE, A.H., STEELE, K.A., MOORE, B.J., BARRACLOUGH, P.B., CLARK, L.J. A combined RFLP and AFLP linkage map of upland rice (*Oryza sativa* L.) used to identify QTLs for root-penetration ability. Theoretical and Applied Genetics, v. 100, p. 49-56, 2000.

PRICE, A.H., STEELE, K.A., MOORE, B.J., JONES, R.G.W. Upland rice grown in soil-filled chambers and exposed to contrasting water-deficit regimes II. Mapping quantitative trait loci for root morphology and distribution. Field Crops Research, v. 76, p. 25-43, 2002.

QUILLET, M.C. MADJIDAN, N., GRIVEAN, Y., SERIEYS, H., TERSAC, M., LORIEUX, M., BERVILLÉ, A. Mapping genetic factors controlling pollen viability in an interspecific cross in *Helianthus* sect*. Helianthus*. Theoretical and Applied Genetics, v. 91, p. 1195-1202, 1995.

RUBEENA, FORD, R., TAYLOR, P.W.J. Construction of an intraspecific linkage map of lentil (*Lens culinaris* ssp. culinaris). Theoretical and Applied Genetics, v. 107, p. 910-916, 2003.

SCHUSTER, I., CRUZ, C.D. Estatística genômica aplicada a populações derivadas de cruzamentos. Viçosa: UFV, 2004. 540p.

SÉNE, M., CAUSSE, M., DAMERVAI, C., THÉVENOT, C., PRIOUL, J-L. Quantitative trait loci affecting amylose, amylopectin and starch content in maize recombinant inbred lines. Plant Physiology Biochemistry, v. 38 (6), p. 459-472, 2000.

TAR'AN, B., MICHAELS, T.E., PAULS, K. P. Genetic mapping of agronomic traits in common bean. Crop Science, v. 42, p. 544–556. 2002.

TASMA, I.M., SHOEMAKER, R.C. Mapping flowering time gene homologs in soybean and their association with maturity (E) loci. Crop Science, v. 43, p.319–328, 2003.

TEIXEIRA, T.A.C. Padrões moleculares, diversidade genética e mapa parcial de ligação do cafeeiro. Viçosa, MG: UFV, 2001. 108p. (Dissertação Doutorado). Universidade Federal de Viçosa, 2001.

UGA, Y., FUKUTA, Y., CAI, H.W., IWATA, H., OHSAWA, R., MORISHIMA, H., FUJIMURA, T. Mapping QTLs influencing rice floral morphology using recombinat inbred lines derived from a cross between *Oryza sativa* L. and *Oryza rufipogon* Griff. Theoretical and Applied Genetics, v. 107, p. 218-226, 2003.

XING, Y.Z., TAN, Y.F., HUA, J.P., SUN, X.L., XU, C.G. Characterization of the main effects, epistatic effects and their environmental interactions of QTLs on the genetic basis of yield traits in rice. Theoretical and Applied Genetics, v. 105, p. 248-257, 2002.

XU, Y., ZHU, J., HUANG, N., McCOUCH, S.R. Chromosomal regions associated with segregation distortion of molecular markers in F2, backcross, double haploids, and recombinant inbred population in rice (*Oryza sativa* L.). Molecular and General Genetics, v. 253, p. 535-545, 1997.

XU, X.F., MEI, H.W., LUO, L.J., CHENG, X.N., LI, Z.K. RFLP-facilitated investigation of the quantitative resistance of rice to brown planthopper (*Nilaparvata lugens*). Theoretical and Applied Genetics, v. 104, p. 248-253, 2002.

YU, J-K., TANG, S., SLABAUGH, M.B., HEESACKER, A., COLE, G., HERRING, M., SOPER, J., HAN, F., CHU, W-C., WEBB, D.M., THOMPSON, L., EDWARDS, K.J., BERRY, S., LEON, A.J., GRONDONA, M., OLUNGU, C., MAES, N., KNAPP, S.J.. Towards a saturated molecular genetic linkage map for cultivated sunflower. Crop Science, v. 43, p.367–387, 2003.

YUAN, J., NJITI, V. N., MEKSEM, K., IQBAL, M. J., TRIWITAYAKORN, K., KASSEM, My. A., DAVIS, G. T., SCHMIDT, M. E., LIGHTFOOT, D. A. Quantitative trait loci in two soybean recombinant inbred line populations segregating for yield and disease resistance. Crop Science, v. 42, p. 271–277, 2002.

WANG, S. Simulation study of the methods for mapping quantitative trait loci in inbred line crosses. Hangzhou, Zhejiang, China, 2000. 97p. (A Ph.D. Dissertation). Zhejiang University, 2000.

WANG, Y.X., WU, P., WU, Y.R., YAN, X.L. Molecular marker analysis of manganese toxicity tolerance in rice under greenhouse conditions. Plant and Soil, v. 238, p. 227-233, 2002.

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

WU, J., JENKINS, J.N., ZHU, J., McCARTY JR., J.C., WATSON, C.E. Comparations of quantitative trait locus mapping properties between two methods of recombinant inbred line development. Euphytica, v. 132, p. 159-166, 2003.

ZABEAU, M. Selective restriction fragment amplification: a general method for DNA fingerprinting. European Patent Application Nº 0534858 A1. 1993.

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

ZHANG, W.P., SHEN, X.Y., WU, P., HU, B., LIAO, C.Y. QTLs and epistasis for seminal root length under a different water supply in rice (*Oriza sativa* L.). Theoretical and Applied Genetics, v. 103, p. 118-123, 2001.

ZHENG, B.S., YANG, L., ZHANG, W.P. MAO, C.Z., WU, Y.R., YI, K.K., LIU, F.Y., WU, P. Mapping QTLs and candidate genes for rice root traits under different water-supply conditions and comparative analysis across three populations. Theoretical and Applied Genetics, v. 107, p. 1505-1515, 2003.

7 ANEXO

Trabalhos de mapeamento com populações RIL e algumas considerações

O tamanho de população e número de marcadores adequados a serem utilizados em mapeamento genético é uma questão ainda não muito bem definida, uma vez que ao se analisar diversos trabalhos de mapeamento em diversas espécies disponíveis na literatura não se nota a presença de um consenso.

Através da análise dos dados disponíveis no Quadro 3 (item 2.2 da Revisão de Literatura) é possível fazer as seguintes considerações:

1. A maioria dos trabalhos com populações RIL são com espécies autógamas (soja, arroz, trigo e feijão) e uma menor proporção com espécie alógama (milho). Isto se deve a maior facilidade de obtenção de linhagens endogâmicas em espécies autógamas;
2. Os tamanhos de população são bastante variados, incluindo populações com 48 linhagens até 330. Porém, a maioria dos trabalhos foram feitos com 150 linhagens endogâmicas ou menos;
3. O número de marcadores utilizados também é muito variável, há trabalhos com até 444 marcadores, porém a grande maioria utilizou menos de 200 marcadores. Quanto aos marcadores, há a utilização de diversos tipos, tanto dominantes com co-dominantes, mas para RIL a interação entre alelos de um mesmo loco não importa, uma vez que só há genótipos homozigotos dominantes ou homozigotos recessivos nas linhagens ocorrendo sempre segregação de 1:1;
4. As funções de mapeamento de Kosambi e Haldane foram utilizadas em alguns trabalhos e para outros não foi especificado. Talvez a não disponibilidade

pudesse ser julgada como sendo utilizada a própria porcentagem de recombinação como unidade de distância;

5. Os parâmetros LOD *score* mínimo e freqüência de recombinação máxima ($r_{máx}$) utilizados na determinação da ligação de locos gênicos não estão disponíveis em vários trabalhos. Para os trabalhos que apresentaram estas informações verifica-se a utilização de LOD mínimo de 2 ou 3 e $r_{máx}$ de 30 ou 40, sendo a maioria de 30;

Nos trabalhos clássicos de experimentação agrícola, usualmente, são fornecidos ao leitor informações do tipo de delineamento experimental, número de repetições e coeficiente de variação ambiental, possibilitando avaliar a qualidade do trabalho. Analogamente, as informações de LOD mínimo, $r_{máx}$ e variância das estimativas de freqüência de recombinação dos marcadores (pelo menos daqueles associados a QTLs) deveriam ser sempre disponibilizados ao leitor, de modo que este possa averiguar a qualidade do trabalho e fazer comparações de resultados.