SILVAN GOMES DE BRITO

OTIMIZAÇÃO DO MAPEAMENTO GENÉTICO VEGETAL VIA SIMULAÇÃO COMPUTACIONAL

RECIFE

2012

SILVAN GOMES DE BRITO

OTIMIZAÇÃO DO MAPEAMENTO GENÉTICO VEGETAL VIA SIMULAÇÃO COMPUTACIONAL

Dissertação apresentada ao Programa de Pós-Graduação em Agronomia "Melhoramento Genético de Plantas", da Universidade Federal Rural de Pernambuco, como parte dos requisitos para obtenção do grau de Mestre em Agronomia.

COMITÊ DE ORIENTAÇÃO:

Professor Dr. Péricles de Albuquerque Melo Filho, Orientador – UFRPE

Professor Dr. Diogo Gonçalves Neder, Coorientador – UEPB

RECIFE

2012

Ficha catalográfica

B862o Brito, Silvan Gomes de
Otimização do mapeamento genético vegetal via simulação computacional / Silvan Gomes de Brito. -- 2012.
78 f.: il.

Orientador: Péricles de Albuquerque Melo Filho.
Dissertação (Mestrado em Melhoramento Genético de Plantas) – Universidade Federal Rural de Pernambuco, Departamento de Agronomia, Recife, 2012.
Referências.

1. Melhoramento genético 2. F2 3. Duplo-haplóide 4. Marcadores genéticos 5. Mapa de ligação I. Melo Filho, Péricles de Albuquerque, orientador II. Título

CDD 581.15

OTIMIZAÇÃO DO MAPEAMENTO GENÉTICO VEGETAL VIA SIMULAÇÃO COMPUTACIONAL

SILVAN GOMES DE BRITO

Dissertação defendida e aprovada pela Banca Examinadora em:_____/____/_____.

ORIENTADOR: _____

Prof. Dr. Péricles de Albuquerque Melo Filho - UFRPE

EXAMINADORES:

_____

Prof. Dr. José Luiz Sandes de Carvalho Filho - UFRPE

_____

Prof. Dr. Diogo Gonçalves Neder - UEPB

_____

Dr. Roberto de Albuquerque Melo – UFRPE

Recife, PE
Julho, 2012

A Deus

Ofereço

Aos meus pais, Sebastião Gomes e Maria da Penha, a minha irmã, Silvana, a minha esposa, Paula Fernanda e aos amores de minha vida (minhas filhas), Janice, Samantha Gabrielle e Lavínia Fernanda e a meu avô Adalberto Gomes (Betinho) (*in memorian*)

Dedico

Agradecimentos

À Deus em primeiro lugar, por ter me dado a oportunidade de viver e a capacidade para enfrentar e vencer as dificuldades proporcionadas ao longo da vida. À minha família, pelo apoio e confiança no meu trabalho. Aos meus queridos pais por terem proporcionado toda condição para concluir este meu objetivo.

À minha irmã pela força e carinho. Ao meu cunhado Eduardo pelo apoio e grande amizade. À minha esposa Paula Fernanda e seus pais José Gomes e Anunciada Dutra e Berenice Eugênio por acreditar em mim, dando crédito a minha escolha profissional e a minha inserção em sua família.

Ao meu orientador Prof. Dr. Péricles de Albuquerque, pelas valiosas contribuições que melhoraram a qualidade deste trabalho.

Ao meu Co-orientador Prof. Dr. Diogo Neder, pelo convívio, paciência, amizade e ensinamentos transmitidos, contribuindo para minha formação acadêmica.

Aos Professores do mestrado: Clodoaldo Anunciação; Edson Silva; Francisco Oliveira; Gerson Quirino; José Luiz Sandes; Mario Lira, Péricles, Rosimar Musser e Vívian Loges, pelos ensinamentos e dedicação ao Programa de Melhoramento Genético de plantas da UFRPE.

À secretária Bernadete Pinto de Lemos, pela paciência e constantes ajudas fornecidas.

Aos colegas da UFRPE: Adriana, Alisson, Alisson Jales, Ana Luisa, Ana Rafaela, Claudia, Felipe, Guilherme, Gustavo, Hudsonkleio, Hudson, Horace, Ismael Gaião, Gustavo, Ivanildo, Jayne, José Carlos, João Filipi, Kessyana, Lucas, Lenivânia, Lindomar, Marciana, Marília, Natália, Paulo, Rafaela, Rebeca, Ramon, Rodolfo, Renata, Samy, Taciana e Thiago Prates. Aos alunos e companheiros de avaliações e análises Eliane Cristina e Mairykon.

Ao apoio institucional e financeiro da Universidade Federal Rural de Pernambuco - UFRPE. A Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - CAPES, pela concessão à bolsa.

Enfim, a todos que não pude citar e aqueles que contribuíram para a conclusão desta minha etapa profissional.

Muito obrigado!

Tua caminhada ainda não terminou... A realidade te acolhe dizendo que pela frente o horizonte da vida necessita de tuas palavras e do teu silêncio.

Se amanhã sentires saudades, lembra-te da fantasia e sonha com tua próxima vitória. Vitória que todas as armas do mundo jamais conseguirão obter, porque é uma vitória que surge da paz e não do ressentimento.

É certo que irás encontrar situações tempestuosas novamente, mas haverá de ver sempre o lado bom da chuva que cai e não a faceta do raio que destrói.

Tu és jovem. Atender a quem te chama é belo, lutar por quem te rejeita é quase chegar à perfeição.

A juventude precisa de sonhos e se nutrir de lembranças, assim como o leito dos rios precisa da água que rola e o coração necessita de afeto.

Não faças do amanhã o sinônimo de nunca, nem o ontem te seja o mesmo que nunca mais.

Teus passos ficaram. Olhes para trás... mas vá em frente pois há muitos que precisam que chegues para poderem seguir-te.

Charles Chaplin

RESUMO

O mapeamento genético baseia-se no ordenamento linear e estabelecimento da distância entre marcas associadas a genes responsáveis pelo controle de características qualitativas e quantitativas. A construção de mapas genéticos é considerada uma das aplicações de maior impacto da tecnologia de marcadores moleculares na análise genética de espécies, e potencialmente, no melhoramento de plantas. Um mapa genético de ligação pode ter baixa, média e alta resolução, de acordo com menor ou maior número de genes ou marcadores ordenados. Um fator de fundamental importância para se obter dados consistentes que resultem em mapas mais acurados é o tamanho da amostra ou da população, o nível de saturação nos grupos de ligação e tipo de marcador a ser utilizado. Desse modo, objetivou-se com este trabalho estimar o tamanho ideal de população e saturação do genoma para a obtenção de mapas de ligação confiáveis por meio de simulação de dados em computador. Foram gerados três genomas com níveis de saturação de 5, 10 e 20 cM, contendo 210, 110 e 60 marcas, respectivamente, para populações F_2 e populações duplo-haplóide. Cada genoma foi composto por 10 grupos de ligação, com um tamanho de 100 cM cada. Para cada nível de saturação do genoma foram geradas populações com 100, 200, 300, 500, 800 e 1000 indivíduos, com 100 repetições cada, sendo utilizado marcadores codominantes e dominantes quando as populações eram do tipo F_2 e apenas dominante para populações duplo-haplóide. Estas populações foram mapeadas utilizando um LODmín de 3 e frequência máxima de recombinação de 30%. Dos mapas obtidos foram extraídas informações referentes ao 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 obtida pela correlação de Spearman e grau de concordância das distâncias nos mapas com o genoma original obtida pelo estresse. Populações de mesmo tamanho tendem a produzir mapas com maior acurácia em níveis de saturação do genoma mais elevados. O tamanho ideal de populações F_2 para mapeamento genético é de no mínimo 200 indivíduos quando os marcadores forem do tipo codominante e de 300 quando os marcadores forem do tipo dominante, independente do nível de saturação do genoma. Enquanto que em populações duplo-haplóide o tamanho ideal é de 200, 500 e 1000 indivíduos quando os níveis de saturação do genoma forem de 5, 10 e 20 cM, respectivamente.

Termos para indexação: melhoramento genético, F_2, duplo-haplóide, marcadores genéticos, mapa de ligação.

ABSTRACT

Genetic mapping based on the planning and establishment of the linear distance between marks associated with genes responsible for controlling qualitative and quantitative characteristics. The construction of genetic maps is considered the most impact applications of the technology of molecular markers in genetic analysis of species, potentially in plant breeding. A genetic map of linkage may be low, medium and high resolution in accordance with the greater or lesser number of genes or ordered markers. A factor of considerable importance to obtain consistent data that result in more accurate maps is the sample size or population, level of saturation in the linkage groups and marker type to be used. Thus, the aim of this work was to estimate the optimum size of population and saturation of the genomes were generated with saturation levels of 5, 10 and 20 cM, containing 210, 110, and marks 60, respectively, and F2 populations double-haploid populations. Each genome was composed of 10 linkage groups, with a size of 100 cM each. For each level of saturation of the genome populations were generated at 100, 200, 300, 500, 800 and 1000 individuals, with 100 replicates, each codominant and dominant markers used when the type F_2 populations were dominant and only double-haploid populations. These populations were mapped using LODmín 3 and a maximum frequency of recombination of 30%. From the maps obtained were extracted information regarding the number of linkage groups and marks for group, size of linkage group, distance between adjacent marks, variance of the distances between adjacent marks, marks inversion obtained by Spearman correlation and degree of agreement of distances on maps with the original genome, obtained by the stress. Populations of the same size tend to produce maps with greater accuracy in higher levels of genome saturation. The optimum size of F_2 populations for genetic mapping must be of at least 200 individuals when codominant markers are of type and 300 when the markers are the dominant type, regardless of the saturation level of the genome. While double-haploid populations in the optimal size was 200, 500 and 1000 individuals when the saturation levels of the genome were 5, 10 and 20 cM, respectively.

Index terms: genetic improvement, F_2, double-haploid, genetic markers, linkage map.

LISTA DE TABELAS

CAPÍTULO II - OTIMIZAÇÃO DO MAPEAMENTO GENÉTICO VEGETAL DE POPULAÇÕES F_2 VIA SIMULAÇÃO COMPUTACIONAL

CAPÍTULO III - OTIMIZAÇÃO DO MAPEAMENTO GENÉTICO VEGETAL DE POPULAÇÕES DUPLO-HAPLÓIDE VIA SIMULAÇÃO COMPUTACIONAL

SUMÁRIO

CAPÍTULO I

INTRODUÇÃO GERAL

1 INTRODUÇÃO GERAL

Nos últimos anos, houve um aumento significativo de metodologias da genética molecular e suas aplicações visando conhecer o número de genes e, ou alelos envolvidos na expressão de determinado caráter, assim como posição e localização de regiões genômicas que controlam caracteres de importância e a quantificação dos efeitos dessas regiões nos cromossomos. Assim, estudos envolvendo o mapeamento podem ser úteis para aplicações no melhoramento genético, assumindo um papel importante na identificação de locos de características quantitativas (QTL - *Quantitative Trait Loci*).

O mapeamento genético é um passo necessário para entender a organização genômica e a relação entre genes e fenótipo. Para isto, é preciso encontrar a ordem e o espaçamento correto entre marcas no mapa de uma população. Um mapa genético é uma representação linear ou circular, de modo que marcadores estão ordenados e distanciados ao longo do cromossomo de um organismo (CRANE e CRANE, 2005).

No melhoramento de plantas e animais, os mapas genéticos são de fundamental importância, uma vez que permitem a visualização, mesmo que de forma relativa, da organização dos genes nos cromossomos. Possibilitam ainda 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 (SILVA, 2005).

Em estudos de plantas, um mapa genético é estimado a partir de um conjunto de dados derivado de uma população de mapeamento, que possui certas características que devem ser levados em conta no processo de estimação (CHEEMA e DICKS, 2009).

Com o avanço das técnicas de biologia molecular surgiram ferramentas que permitem a construção de mapas genéticos mais confiáveis, como o desenvolvimento de tipos de marcadores moleculares do DNA, que tem permitido a construção de mapas genéticos com elevado nível de saturação, caracterizado pela presença de marcas genéticas. Permitem também identificar polimorfismo genético diretamente no DNA. Entre as vantagens dos marcadores moleculares, pode-se citar

a obtenção de polimorfismo genético, identificação direta do genótipo sem influência do ambiente, a possibilidade de detecção de tais polimorfismos em qualquer estádio do desenvolvimento da planta (FALEIRO, 2007). Uma das limitações é a construção de populações de mapeamento. Dependendo do objetivo e da espécie em estudo, podem ser utilizados vários tipos de populações segregantes, tais como, populações F_2, progênies provenientes de retrocruzamentos (RC), progênies F_1 (*pseudo-testcross*), duplo-haplóides (DH) e linhas endogâmicas recombinantes (RILs – *Recombinant Inbred Lines*) (PEREIRA e PEREIRA, 2006), sendo que cada uma delas apresentam vantagens e desvantagens.

A construção de um mapa genético inicia-se com uma a análise de ligação entre pares de locos, para se verificar a existência de ligação entre eles através do teste de qui-quadrado (x^2), em seguida é necessário definir a frequência máxima de recombinação e o LOD mínimo, para inferir se dois locos estão ligados (BHERING et al., 2008).

Vários fatores afetam a disponibilidade de mapas genéticos fidedignos, como a quantidade de marcadores moleculares utilizados, tipo de população e número de indivíduos de que formam a população analisada. Com a quantidade de dados gerados para a construção de um mapa genético e a complexidade envolvida, torna-se necessário a utilização de um programa computacional, uma vez que as várias análises são realizadas simultaneamente. Com o auxílio de um Software, são estimadas as posições dos marcadores no mapa genético em determinada ordem.

Com isto, o presente estudo teve como objetivo principal identificar o tamanho da população, nível de saturação do genoma e o tipo de marcador para populações F_2 ideal para mapeamento genético em populações segregantes por meio de simulação de dados em computador. Assim, alguns aspectos foram determinados de acordo com os objetivos específicos que seguem:

a) a influência do número de indivíduos da população segregante sobre o mapeamento;
b) o efeito da saturação do genoma por marcadores moleculares no mapeamento;
c) o efeito do tipo de marcador a ser utilizado no mapeamento de diferentes populações;

d) o número adequado de indivíduos a ser utilizado no processo de mapeamento genético.

2 REFERENCIAL TEÓRICO

2.1 Construção de mapas genéticos

Mapa genético ou mapa de ligação é uma representação gráfica das distâncias entre genes e de suas posições relativas em um cromossomo. Um mapa genético é um diagrama onde são representados os genes com suas respectivas posições no cromossomo, ou seja, a distribuição sequencial dos genes ao longo dos cromossomos (NOMELINI et al., 2009). A partir desta ferramenta, é possível avançar em estudos que permitem compreender a ordem e espaçamento dos marcadores de genomas ainda desconhecidos, bem como comparar com de outras plantas, por meio de mapeamento comparativo e sequências do genoma de outras espécies vegetais.

Na maioria das situações, não sabemos como se comportam os genes e onde estão localizados no genoma. Com o rápido desenvolvimento da tecnologia molecular, grandes quantidades de dados moleculares estão agora disponíveis, fornecendo possibilidades para estimar os efeitos, bem como localização de locos de característica quantitativa (LUO e XU, 2003).

A influência do mapa genético não se limita a espécie de plantas cujos genomas não foram ainda sequenciados. Um mapa genético é estimado a partir de um conjunto de dados derivado de uma população de mapeamento, que possui certas características que devem ser levados em conta no processo de estimação (CHEEMA e DICKS, 2009).

A construção de um mapa genético compreende quatro etapas a seguir: a) identificação de marcadores genéticos; b) desenvolvimento de uma população segregante; c) análise da herança dos marcadores genéticos e d) estabelecimento da ordem dos marcadores genéticos e da distancias entre eles (VIEIRA et al., 2006).

2.2 Marcadores genéticos

Qualquer forma alélica originada de um genoma pode ser utilizada como um marcador genético, podendo esta ser um dado fenótipo, uma proteína ou um fragmento de DNA que codifica ou não um gene, possuindo uma sequência repetida

ou única no genoma (SOUZA, 2001). Um marcador genético pode ser uma característica qualquer de fácil percepção, expressa por um organismo, capaz de diferenciar geneticamente de outros organismos (FERREIRA e GRATTAPAGLIA, 1995).

Os marcadores mais antigos e amplamente difundidos são os que têm como base características morfológicas. Estes ainda continuam sendo aplicados com eficiência para certos tipos de germoplasma (BRETTING e WIDRLECHENER, 1995). Os marcadores morfológicos são de fácil identificação visual, como forma e resistência a patógenos, cor do hipocótilo, de pétalas, de brotações, de sementes, morfologia foliar, entre outros.

Dentre as principais dificuldades encontradas para a confecção de mapas oriundos de marcadores morfológicos, pode-se destacar a limitação em se obter marcadores genéticos consistentes e adequados para a análise de ligação (SALGADO, 2008). 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 devido ao baixo nível de polimorfismo, pouca estabilidade ambiental e número limitado de locos disponíveis para estudos de mapeamento (FERREIRA e GRATTAPAGLIA, 1995).

O uso desses marcadores é muitas vezes demorado, dispendioso e incerto, principalmente quando a espécie em estudo é resultado de cruzamentos de outras espécies morfologicamente semelhantes (KLASS, 1998). Com isso, a técnica de eletroforese com isoenzimas passou a ser utilizada com frequência em diferentes linhas de pesquisa, pois detecta, de forma indireta, polimorfismo em sequências de DNA e na carga elétrica de proteínas com função enzimática, gerado por mutações na sequência gênica (BOTREL e CARVALHO, 2004). Apresenta ainda natureza codominante e custos relativamente baixos tornando-a atraente e útil em pesquisas que necessitam de um grande número de indivíduos (CAIXETA et al., 2006), no entanto, apresenta um baixo número de variantes nas proteínas, limitando o desenvolvimento de mapas genéticos altamente saturados (SILVA, 2005).

Neste contexto, surgem os marcadores moleculares, também chamados marcadores genômicos, baseados na análise direta de sequências de DNA

polimórfico. Teoricamente qualquer fragmento de DNA pode ser utilizado como um marcador molecular, desde que ele revele polimorfismo entre indivíduos (SOUZA, 2001).

O método mais avançado é o sequenciamento, porém, existem métodos mais simples e mais baratos, suficientes para a maioria dos propósitos. Estes métodos incluem: a) a análise de polimorfismo de comprimento de fragmentos de restrição de DNA (RFLP - *Restriction Fragment Length Polymorphism*) - 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. Esses marcadores possuem expressão codominante, possibilitando assim, a identificação de genótipos homozigotos e heterozigotos; b) os baseados na reação em cadeia da polimerase (PCR – *Polymerase Chain Reaction*) (MULLIS e FALLONA, 1987) - 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, utilizado como iniciadores ou primers e que delimitam a sequência de DNA de fita dupla, alvo da amplificação. Esta técnica é utilizada na obtenção de marcadores (RAPD – *Randomly Amplified Polymorphic* DNA) (WILLIAMS et al., 1990), microssatélites (SSR – *Simple Sequence Repeats*) (MÉTAIS et al., 2002), polimorfismo de comprimento de fragmentos amplificados (AFLP – *Amplified Fragment Length Polymorphism*) (VOS et al., 1995) e regiões amplificadas caracterizadas por sequência (SCAR - *Sequence Characterized Amplified Regions*) (NEGI et al., 2000).

No melhoramento genético de plantas, tem-se buscado, cada vez mais, a seleção assistida por marcadores moleculares, visando obter maior eficiência na transferência de fatores genéticos. Assim, marcadores moleculares são ferramentas úteis para detectar variações no genoma, aumentando o poder da análise genética das plantas (CAIXETA et al., 2006).

Os marcadores moleculares apresentam algumas vantagens em relação aos marcadores morfológicos, como: maior número de alelos por loco com elevado nível de polimorfismo, o que facilita o mapeamento genético de populações segregantes,

não sofrem qualquer tipo de influência do ambiente e são codominantes (VIEIRA et al., 2006).

2.3 Populações segregantes para mapeamento genético

No mapeamento genético, diferentes tipos de populações segregantes podem ser empregados. As populações rotineiramente mais utilizadas são aquelas derivadas do cruzamento de linhagens homozigotas (puras) com diferenças contrastantes, o que origina uma geração F_1 a qual é autofecundada ou retrocruzada com um dos genitores para a produção de uma geração F_2 ou RC_1, respectivamente. Alternativamente, são utilizadas populações F_n (n= 3, 4,...,∞), duplo-haplóides e linhagens endogâmicas recombinantes (RILs) (LYNCH e WALSH, 1998).

Cada população apresenta suas próprias vantagens e desvantagens e os pesquisadores precisam decidir a população apropriada, devendo considerar o objetivo do projeto, a complexidade da característica em estudo, tempo disponível, e o tipo dos marcadores a serem utilizados, se dominantes ou codominantes.

A população mais indicada para construção de mapas moleculares é a população F_2, devido ao desequilíbrio de ligação atingir seu ponto máximo (YUONG, 1994; PEREIRA e PEREIRA, 2006). Estas populações são oriundas da autofecundação das plantas F_1 ou do intercruzamento destes indivíduos. Para marcadores codominantes a segregação ocorre na proporção 1:2:1, enquanto que nos dominantes, a proporção segue a ordem de 3:1. Dados provenientes da segregação de marcadores codominantes em populações F_2 são os que fornecem mais informação acerca do grau de ligação genética entre marcadores, enquanto que os marcadores dominantes apresentam pouca eficiência na detecção de ligação entre estes marcadores (LIU, 1998). Como vantagem, os marcadores codominantes apresentam 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) (SCHUSTER e CRUZ, 2008), os gametas de cada indivíduo são informativos. 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 (SILVA, 2005). Estas desvantagens podem ser contornadas por meio da propagação vegetativa ou com a utilização de populações $F_{2:n}$, geralmente $F_{2:3}$, de modo que os genótipos das plantas são determinados na geração F_2, enquanto as características fenotípicas são medidas em plantas F_3, com repetições (BARROS, 2007).

Outra população utilizada é a duplo-haplóide, obtida pela duplicação do número de cromossomos dos micrósporos das plantas da geração F_1, de modo que, ao final do processo, cada indivíduo apresentará dois cromossomos homólogos idênticos representando a variabilidade genética encontrada no indivíduo parental que produziu os micrósporos. Com a duplicação dos cromossomos, as plantas com genótipo AA e aa ocorrem na mesma proporção, ou seja, a segregação em locos individuais é de 1:1. Pelo exposto, a informação obtida com marcadores dominante neste tipo de população é igual à obtida com marcadores codominantes, de modo que todos os loci serão homozigóticos.

Dentre as populações utilizadas, as obtidas de linhagens homozigotas são ideais para a detecção e mapeamento de QTLs, pois todas as plantas da geração F_1 são geneticamente idênticas e mostram um completo desequilíbrio de ligação para os locos contrastantes nas linhagens (LYNCH e WALSH, 1998).

Em processos convencionais de melhoramento são necessários de sete a oito ciclos para se obter linhagens homozigotas enquanto que em duplo-haplóides em uma única geração a homozigose total é alcançada. Este apresenta algumas vantagens tais como, os indivíduos obtidos são homozigotos, diminuição no tempo necessário para o estabelecimento de estoques puros de novas cultivares, redução de custos, maior variância genética, melhor eficiência na seleção. Entretanto, quando comparadas com linhagens convencionais, aponta-se que em geral as plantas duplo-haplóides são inferiores em várias características, uma vez que não sofreram seleção natural para sua obtenção (BARBOSA, 2009).

Em populações de retrocruzamento, apenas dois genótipos são obtidos por meio do cruzamento de plantas F_1 com um dos genitores. Pelo exposto, um indivíduo heterozigoto (Aa) é cruzado com um homozigoto (aa) ou (AA). Assim, o indivíduo F_1 dará origem à metade de gametas A e metade a, enquanto que o outro genitor produzirá apenas um tipo de gameta (A ou a). Com isso, a segregação

esperada na população descendente será de 1:1, ou seja, apenas dois genótipos serão formados, de modo que metade será Aa e a outra metade AA ou aa.

As principais vantagens são: alta previsibilidade do método e o fato de que a segregação obtida representa a composição dos grãos de pólen do progenitor doador. A grande limitação é o tempo requerido e a quantidade de cruzamentos necessários para obtenção da população, o que pode inviabilizar sua utilização em espécies em que a hibridação é difícil (SCHUSTER e CRUZ, 2008) ou cujo ciclo de vida é demasiadamente longo. Além disso, 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 (SILVA, 2005), além de ser um método trabalhoso.

As populações derivadas de linhagens endogâmicas recombinantes (RILs) são originadas por sucessivas gerações de autofecundação dos indivíduos de uma população F_2, até que elevado grau de homozigose seja alcançado. Podem ser obtidos pelo método da descendência de uma única semente (SSD-*Single seed descend*) (SCHUSTER e CRUZ, 2008) ou por cruzamentos entre irmãos, de maneira que ao final do processo todos os locos gênicos terão uma segregação de 1:1 (AA:aa). As sucessivas autofecundações fazem com que ao longo do processo ocorra a fixação gênica, ou seja, devido à redução da proporção de locos em heterozigose a metade em cada geração, tendendo a zero.

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 (SCHUSTER e CRUZ, 2008). 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. Dentre as vantagens, destacam-se o elevado índice de homozigose; os genótipos dos indivíduos podem ser perpetuados; possibilidade de cultivo em vários locais, devido a grande disponibilidade de sementes, o que aumenta a precisão na estimativa de QTLs, além de possibilitar mapeamento de locos considerando interação genótipo x ambiente para esses caracteres. As desvantagens da utilização

de RILs são o tempo requerido para a obtenção da população, de seis a oito gerações de autofecundações e, a impossibilidade da estimação de efeitos devido à dominância, uma vez que apresentam apenas dois genótipos na população segregante (SILVA, 2005).

2.4 Genotipagem da população segregante

Após a escolha do tipo de população é necessária à análise do padrão de amplificação dos indivíduos do restante da população de mapeamento e a obtenção das estimativas de recombinação. Na construção de um mapa genético, todos os marcadores são analisados dois a dois, para se verificar independência ou existência de ligação entre eles (LIU, 1998).

Uma matriz de dados é então gerada onde as linhas constituem os marcadores moleculares, e as colunas, os diferentes indivíduos genotipados. Aplica-se um teste estatístico (x^2) para cada marcador a fim de verificar se os marcadores moleculares segregam de acordo com as proporções mendelianas para um loco.

2.5 Estabelecimento da distância e ordenamento dos marcadores

Há vários métodos para estimar a distância em unidades de recombinação entre dois marcadores, expresso em porcentagem de recombinação entre dois locos. Assim, diversos métodos sofisticados utilizando a máxima verossimilhança, estimam a frequência de recombinação entre marcadores e algoritmos de ordenação rápida têm sido aplicados para a construção de mapas de maior precisão (SCHUSTER e CRUZ, 2008). De maneira geral, métodos de máxima verossimilhança estimam porcentagem de recombinação maximizando a probabilidade de se observar os dados genotípicos obtidos.

Uma vez estimadas as frequências de recombinação entre os pares de locos, o passo seguinte é formar os grupos de ligações (CRUZ e SILVA, 2006).

A ordem linear de locos marcadores dentro de cada grupo de ligação é deduzida a partir da distância genética entre eles (VIEIRA et al., 2006). De acordo com o número de genes ou marcadores, um mapa genético pode ter baixa, média ou alta resolução. Pelo exposto, em um grupo de ligação é possível encontrar a

ordem de seus marcadores que maximiza ou minimiza uma função de pontuação, de modo a avaliar a qualidade do ordenamento de um dado marcador, descrevendo se uma ordem é melhor ou mais adequada do que outra ordem de um marcador.

Agrupados os diversos marcadores em grupos de ligação, e de posse das distâncias genéticas duas a duas, um mapa pode ser montado manualmente em um procedimento clássico denominado mapeamento de três pontos, em que os marcadores são ordenados sequencialmente com base nas distâncias dois a dois (FERREIRA e GRATTAPAGLIA, 1998).

Há diversas maneiras de estimar a melhor ordem considerando três marcadores (SCHUSTER e CRUZ, 2008). Além disso, será necessária a utilização de uma função para identificar a melhor ordem. O principal método utilizado é conhecido como SARF (*Sum of Adjacent Recombination Fractions*). Esta técnica consiste em considerar, de cada vez, apenas as informações de dois locos, de modo que, a melhor ordem será aquela que quando combinada, proporcionar menor soma das frequências de recombinações adjacentes (SCHUSTER e CRUZ, 2008). Assim, neste estudo será abordado o método da frequência de recombinação entre dois locos.

2.6 Simulação computacional no mapeamento genético

Para a realização do mapeamento genético é necessário que se defina a população de mapeamento a ser utilizado, o tamanho da população de mapeamento, o número de marcas a ser obtido, o tipo de marcador a ser empregado, dentre outras variáveis. Para que seja possível a estimação das frequências de recombinação com boa acurácia, a subdivisão adequada das marcas nos grupos de ligação e o ordenamento adequado dentre dos grupos, gerando por fim, mapas genéticos fidedignos. Entretanto, a análise de estudos de mapeamento disponíveis na literatura não permite que se chegue a um consenso quanto a melhor estratégia para o mapeamento genético em plantas (SILVA, 2005).

Apesar da ausência de consenso na escolha do tamanho, tipos de populações, número de grupos de ligação, tipo de marcas, dentre outros, 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

RILs em três diferentes tamanhos de população. Este estudo foi simulado, por meio do procedimento de Monte Carlo, um genoma com quatro cromossomos e 64 marcadores, 16 marcadores por cromossomo não equidistantes, 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%. Os autores concluíram que em populações RILs derivadas pelo método SSD, os efeitos dos QTLs e suas posições foram mais próximos dos valores pré-fixados para populações com tamanho de 150 ou mais indivíduos.

Em seguida, Silva (2005) realizou em estudo de simulação procurando avaliar o efeito do tamanho da população e do nível de saturação do genoma no mapeamento de populações RIL, assim como definir o número de indivíduos e marcas a serem utilizados nestes estudos. 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. Após análise genômica, os “genomas analisados” foram comparados com os “genomas simulados” e desta forma, concluiu-se que tamanhos mínimos de 100, 154 e 500 indivíduos são 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.

Posteriormente, Good God (2008) estudou o mapeamento de famílias de meios irmãos via simulação computacional. Foram simuladas populações considerando níveis de saturação de genoma de 5, 10 e 20 cM, seis tamanhos populacionais 50, 100, 200, 300, 500 e 1000 indivíduos e cinco níveis de informação alélica (s = 2, 3, 4, 5 e 6 alelos). Após os procedimentos de análise genômica e comparação dos genomas, observou-se a necessidade de pelo menos 200 indivíduos para a recuperação de mapas genéticos fidedignos, desde que o número de alelos na população seja igual ou superior a quatro.

Em famílias de irmãos completos, Bhering e Cruz (2008) avaliaram o tamanho ótimo de indivíduos em populações de irmãos completos. Neste caso, foram simulados genomas parentais e amostras de populações de família de irmãos completos do tipo completamente informativas, e também não completamente

informativas com tamanho de população de 100, 200, 400 e 600 indivíduos, com três grupos de ligação cada, e 11 marcas moleculares codominantes e multialélicas, espaçadas a 10 cM por grupo de ligação. Os autores concluíram que, o tamanho populacional de 200 indivíduos é suficiente para recuperar as informações originais em populações completamente informativas, contudo, para a população não completamente informativa, é necessária a utilização de uma população maior, de 600 indivíduos.

REFERÊNCIAS

BARBOSA, M. P. M. **Avaliação do desequilíbrio de ligação e da origem genética em duplo-haplóides de milho.** 2009. 62p.Tese (Doutorado) - Faculdade de Ciências Agrárias e Veterinária da Universidade Estadual Paulista, Jaboticabal.

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

BHERING, L.; CRUZ, C. D. Tamanho de população ideal para mapeamento genético em famílias de irmãos completos. **Pesquisa Agropecuária Brasileira**, v.43, p.379-385, 2008.

BHERING, L.; CRUZ, C. D.; GOOD GOD, P. Estimativa de frequência de recombinação no mapeamento genético de famílias de irmãos completos. **Pesquisa Agropecuária Brasileira,** v.43, p.363-369, 2008.

BOTREL, M.C.G.; CARVALHO, D. Variabilidade isoenzimática em populações naturais de jacarandá paulista (*Machaerium villosum* Vog.). **Revista Brasileira de Botânica,** São Paulo, v.27, p.621-627, 2004.

BRETTING, P.K.; WIDRLECHENER, M.P. Genetic markers and horticultural germplasm management. **HortScience,** v. 30, p.1349-1355, 1995.

CAIXETA, E. T.; OLIVEIRA, A. C. B.; BRITO, G. G.; SAKIYAMA, N. S. Tipos de marcadores moleculares. In: BORÉM, A.; CAIXETA, E. T. **Marcadores moleculares.** Viçosa: UFV, 2006. p.9-78.

CHEEMA J.; DICKS, J. Computational approaches and software tools for genetic linkage map estimaton in plants. **Briefinds in Bioinformatics,** v.10, p.595-608, 2009.

CRANE, C. F.; CRANE, Y. A nearest-neighboring-end algorithm for genetic mapping. **Bioinformatics,** v.21, p.1579-1591, 2005.

CRUZ, C. D.; SILVA, L. C. Análise de marcadores moleculares. In: BORÉM, A.; CAIXETA, E. T. **Marcadores moleculares.** Viçosa: UFV, 2006. p.307-374.

FALEIRO, F. G. **Marcadores genético-moleculares: aplicados a programas de conservação e uso de recursos genéticos**. Planaltina-DF: Embrapa Cerrados, 2007. 102p.

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

GOOG-GOD, P. I. V. **Mapeamento genético em famílias de meios irmãos por simulação computacional.** 2008. 114p. Dissertação (Mestrado) - Universidade Federal de Viçosa, Viçosa.

KLASS, M. Applications and impact of molecular markers on evolutionary and diversity studies in *Allium*. **Plant Breeding,** v.117, p.297-308, 1998.

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

LUO, L.; XU, S. Mapping viability loci using molecular markers. **Heredity**, v. 90, p.459-67, 2003.

LYNCH, M.; WALSH, B. **Genetics and analysis of quantitative traits.** Sunderland, MA: Sinuauer Associaets, Inc. 980p. 1998.

MÉTAIS, I.; HAMON, B.; JALOUZOT, R.; PELTIER, D. Structure and level of genetic diversity in various bean types evidenced with microsatellite markers isolated form a genomic enriched library. **Theoretical and Applied Genetics,** v.104, p.1346-1352, 2002.

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.

NEGI, M.S.; DEVIC, M.; DELSENY, M.; LAKSHMIKUMARAM, M. Identification of AFLP fragments to seed colour in *Brassica juncea* and conversion to a SCAR marker for rapid selection. **Theoretical and Applied Genetics,** v.101, p.146-152, 2000.

NOMELINI, Q. S. S.; SILVA, H. D.; FARIA, T. C. Eficiência dos métodos de otimização simulatedannealin, delineação rápida em cadeia e ramos e conexões para construção de mapas genéticos. **Ciência e agrotecnologia**, v.33, p.1534-1537, 2009.

PEREIRA, M. G.; PEREIRA, T. N. S. Marcadores moleculares no pré-melhoramento de plantas. In: **Marcadores Moleculares.** Editores técnicos: Borém, A. e Caixeta, E. T. 2006. p.85-106.

SALGADO, C. C. **Integração de mapas genéticos.** 2008. 142f. Dissertação (Mestrado) - Universidade Federal de Viçosa, Viçosa.

SANTOS, E. K.; ZANETTINI, M. H. B. Androgênese: uma rota alternativa no desenvolvimento do pólen. **Ciência Rural**, v.32, p.165-173, 2002.

SCHUSTER, I.; CRUZ, C.D. **Estatística genômica aplicada a populações derivadas de cruzamentos controlados.** Viçosa: UFV, 2008. 568p.

SILVA, L. C. **Simulação do tamanho da população e da saturação do genoma para mapeamento genético de RILs.** 2005. 132f. Dissertação (Mestrado). Universidade Federal de Viçosa, Viçosa

SOUZA, A. P. Biologia molecular aplicada ao melhoramento. In: NASS, E.L.; VALOIS, A. C. C.; MELO, I. S.; VALADARES-INGLIS, M. C. (Ed). **Recursos Genéticos e Melhoramento-Plantas.** Rondonópolis: Fundação MT, p.938-965, 2001.

VIEIRA, E. A.; NODARI, R.O.; CARVALHO, F. I. F.; FIALHO, J. F. **Mapeamento genético de caracteres quantitativos e sua interação com o ambiente.** Planaltina-DF: Embrapa Cerrados, 2006. 26.p. (Embrapa Cerrados. Documentos, 170). Disponível em: < http://www.infoteca.cnptia.embrapa.br/handle/doc/570277>. Acesso em: 20 mai. 2012.

VOS, P.; BLEEKER, M.; REIJANS, M.; LEE, T. V.; HORNES, M.; FRIJTERS, A.; POT, J.; KUIPER, M.; ZABEAU, M. AFLP: A new technique for DNA fingerprinting. **Nucleic Acids Research,** v.23, p.4407-4414, 1995.

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.

YOUNG, N. D. Constructing a plant genetic linkage map with DNA markers. In: PHILLIPS, R.L.; VASIL, I.K. **DNA-based markers in plants.** Dordrecht, The Netherlands: Kluwer Academic Publisher, p. 39-57, 1994.

CAPÍTULO II

OTIMIZAÇÃO DO MAPEAMENTO GENÉTICO VEGETAL DE POPULAÇÕES F_2 VIA SIMULAÇÃO COMPUTACIONAL

Otimização do mapeamento genético vegetal de populações F_2 via simulação computacional

Silvan Gomes de Brito[(1)], Diogo Gonçalves Neder[(2)], Mairykon Coelho da Silva[(3)], Eliane Cristina Arcelino[(3)], Alisson Esdras Coutinho[(1)] e Péricles de Albuquerque Melo Filho[(4)]

[(1)]Mestrando vinculado ao Programa de Pós-Graduação em Melhoramento Genético de Plantas do Departamento de Agronomia, Universidade Federal Rural de Pernambuco, Rua Dom Manuel de Medeiros, s/n – Dois Irmãos – Recife-PE, CEP: 52171- 900. Email: gomesilvapb@hotmail.com, alissonesdras@yahoo.com.br; [(2)]Professor Doutor do Departamento de Agroecologia e Agropecuária/CCAA, Universidade Estadual da Paraíba/Campus de Lagoa Seca, Sítio Imbaúba, s/n, Zona Rural, Lagoa Seca-PB, Cep: 58.117-000. Email: dgneder@ccaa.uepb.edu.br; [(3)]Aluno de Graduação do Departamento de Agronomia, Universidade Federal Rural de Pernambuco, Rua Dom Manuel de Medeiros, s/n – Dois Irmãos – Recife-PE, CEP: 52171- 900. Email: mairicon@hotmail.com, elianearcelino@gmail.com; [(4)]Professor Associado do Departamento de Agronomia, Universidade Federal Rural de Pernambuco, Rua Dom Manuel de Medeiros, s/n – Dois Irmãos – Recife-PE, CEP: 52171- 900. Email: pericles@depa.ufrpe.br.

RESUMO

Este estudo objetivou avaliar o tipo de marcador molecular, o nível de saturação do genoma e o tamanho ideal de populações F_2 para a construção de mapas de ligação fidedignos por meio de simulação computacional. Foram simulados genomas parentais e populações F_2 considerando marcadores moleculares do tipo dominante e codominante, espaçados de forma equidistante a 5, 10 e 20 cM. Os tamanhos das populações geradas foram de 100, 200, 300, 500, 800 e 1000 indivíduos, com dez grupos de ligação cada e 100 repetições por amostra. Procedeu-se a análise de todas as populações geradas obtendo um genoma analisado o qual foi comparado com o genoma simulado inicialmente. Observou-se que o tamanho ideal de populações F_2 para mapeamento genético foi de no mínimo 200 indivíduos quando os marcadores são do tipo codominante, independente do nível de saturação. Para marcadores do tipo dominante, estes valores são de 300 indivíduos para 5 cM e 200 indivíduos para 10 e 20 cM. Populações de mesmo tamanho tendem a produzir mapas com maior acurácia em níveis de saturação do genoma mais elevados. Para todos os níveis de saturação o aumento do tamanho da população permitiu a obtenção de mapas de ligação com maior acurácia.

Termos para indexação: tamanho da população, tipo de marcador, nível de saturação, mapa de ligação.

Optimization of vegetal genetic mapping of F_2 populations via computational simulation

Silvan Gomes de Brito[(1)], Diogo Gonçalves Neder[(2)], Mairykon Coelho da Silva[(3)], Eliane Cristina Arcelino[(3)], Alisson Esdras Coutinho[(1)] e Péricles de Albuquerque Melo Filho[(4)]

[(1)]Mestrando vinculado ao Programa de Pós-Graduação em Melhoramento Genético de Plantas do Departamento de Agronomia, Universidade Federal Rural de Pernambuco, Rua Dom Manuel de Medeiros, s/n – Dois Irmãos – Recife-PE, CEP: 52171- 900. Email: gomesilvapb@hotmail.com, alissonesdras@yahoo.com.br; [(2)]Professor Doutor do Departamento de Agroecologia e Agropecuária/CCAA, Universidade Estadual da Paraíba/Campus de Lagoa Seca, Sítio Imbaúba, s/n, Zona Rural, Lagoa Seca-PB, Cep: 58.117-000. Email: dgneder@ccaa.uepb.edu.br; [(3)]Aluno de Graduação do Departamento de Agronomia, Universidade Federal Rural de Pernambuco, Rua Dom Manuel de Medeiros, s/n – Dois Irmãos – Recife-PE, CEP: 52171- 900. Email: mairicon@hotmail.com, elianearcelino@gmail.com; [(4)]Professor Associado do Departamento de Agronomia, Universidade Federal Rural de Pernambuco, Rua Dom Manuel de Medeiros, s/n – Dois Irmãos – Recife-PE, CEP: 52171- 900. Email: pericles@depa.ufrpe.br.

ABSTRACT

This study aimed to evaluate the type of molecular marker, the saturation level of the genome and the optimal size of F2 populations for the construction of linkage maps reliable by means of computational simulation. Parental genomes and F2 populations were simulated considering dominant and codominant molecular markers, spaced equidistantly at 5, 10 and 20 cM. The sizes of the generated populations were 100, 200, 300, 500, 800 and 1000 individuals, with ten linking groups and 100 replicates per sample. It was proceeded the analysis of all generated population obtaining a genome which was compared with the first simulated genome. It was observed that the ideal size of F_2 populations for genetic mapping has been at least 200 individuals when the markers are codominant type, regardless of the level of saturation. For dominant type of markers, these values are 300 individuals for 5 cM and 200 individuals for 10 and 20 cM. Populations of the same size tend to produce maps with greater accuracy in higher levels of genome saturation. For all levels of saturation the increase of population size has permitted to obtain linkage maps with greater accuracy.

Index terms: population size, marker type, saturation level, linkage map.

1 INTRODUÇÃO

Com o desenvolvimento da biologia molecular, tornou-se possível identificar marcas moleculares associadas a um ou mais genes, responsáveis pelo controle genético de características qualitativas e quantitativas (QTLs - *Quantitative Trait Loci*), assim como determinar o posicionamento relativo das mesmas através da construção de mapas de ligação.

Como os eventos de permutação ocorrem ao acaso ao longo do cromossomo, a probabilidade de recombinação é maior para locos que se encontram a uma maior distância do que para aqueles mais próximos. Essa é a idéia básica do mapeamento genético, ou seja, a taxa de recombinação entre locos é usada como referência para o cálculo de distância e ordenamento dos genes, ou marcadores, nos cromossomos (SCHUSTER e CRUZ, 2008).

Confirmada a existência de ligação entre duas marcas, é indispensável adotar métodos quantitativos para estudar o grau de associação entre marcas (ROCHA et al., 2003). O método de máxima verossimilhança é utilizado no mapeamento genético para a obtenção de várias estimativas, inclusive as da frequência de recombinação (LIU, 1998). Após a estimação, é necessário definir a frequência máxima de recombinação e o LODmínimo, para inferir se dois locos estão ligados. O objetivo é estabelecer critérios para a formação dos grupos de ligação. Depois de definidos os grupos de ligação e o ordenamento das marcas, a última abordagem consiste em estimar as frequências de recombinação, com base em informações multilocos entre pares de marcas, considerando todos os marcadores ligados em um grupo de ligação simultaneamente, o que resulta numa análise única para cada grupo de ligação (LYNCH e WALSH, 1998).

Dispondo destas informações, o estudo de melhoramento genético pode ser otimizado, tanto na eficiência da seleção dos melhores genótipos em uma população com variabilidade, quanto na velocidade que os ganhos genéticos serão obtidos. Isto se deve a possibilidade de seleção com base nos marcadores moleculares, excluindo-se o efeito ambiental sobre a expressão dos diferentes caracteres de interesse e dispensando etapas de avaliação e seleção.

Os mapas genéticos possibilitam a cobertura e análise completa de genomas, a decomposição de características genéticas complexas nos seus

componentes mendelianos, a localização de regiões genômicas que controlam caracteres de importância, a quantificação do efeito dessas regiões na característica estudada e a canalização de toda essa informação para uso em programas de melhoramento (FALEIRO et al., 2003).

Para isto é fundamental que se estime adequadamente as distâncias entre os genes, o estabelecimento correto do ordenamento e a formação de grupos de ligações que reflitam o número básico de cromossomos da espécie. Entretanto, a confiabilidade das informações contidas em um mapa de ligação depende do tipo de população, do tipo de marca molecular e, principalmente, do número de indivíduos considerados na obtenção da percentagem de recombinação entre pares de marcas (SALGADO et al., 2011).

Tanto o tamanho de população quanto o número de marcas para representação de cromossomos em grupos de ligação ainda não são bem definidos; há falta de padrão para a análise de dados de estudos sobre mapeamento (CRUZ, 2006).

Estas informações vêm sendo obtidas com facilidade por meio de análise computacional (FERREIRA et al., 2006; SILVA et al., 2007; BHERING e CRUZ, 2008; GOOD GOD, 2008), via modelos estatísticos com intuito de descrever sistemas genéticos complexos através de simulação de dados otimizando o mapeamento genético.

O objetivo do presente estudo foi determinar o tipo de marcador molecular, nível de saturação do genoma e tamanho da população ideal a ser empregado no mapeamento genético de populações F_2.

2 MATERIAL E MÉTODOS

2.1 Simulação dos dados

Para a geração dos dados nas duas populações foi considerada uma espécie vegetal diplóide com 2n = 2x = 20 cromossomos utilizando o módulo de simulação do aplicativo computacional GQMOL (CRUZ, 2004), o qual permite gerar informações sobre genomas, genótipos de genitores e indivíduos de diferentes tipos de populações.

2.2 Simulação do genoma, níveis de saturação e tipos de marcas

Foram gerados três genomas com três níveis de saturação: saturado, medianamente saturado e pouco saturado, com intervalos entre marcas adjacentes de 5, 10 e 20 cM, representando um número total de marcas por genoma de 210, 110 e 60, respectivamente. Cada genoma composto por 10 grupos de ligação, com 100 cM em cada grupo e, portanto, com comprimento total de 1000 cM. Foram considerados os tipos de marcadores moleculares dominantes e codominantes, devido a proporção de segregação esperada ser de 1:2:1, distribuídos de forma equidistante no genoma.

2.3 Simulação dos genitores

Para a simulação dos genitores na população F_2, para cada nível de saturação do genoma foi simulado uma situação em que os pais eram homozigotos, sendo um dominante e outro recessivo. Nesta ocasião, admitiu-se que na geração F_1 todos os locos se encontravam em fase de aproximação.

2.4 Tamanho da população

Foram simuladas 100 amostras com 100, 200, 300, 500, 800 e 1000 indivíduos com 11 marcas por grupo de ligação distribuídas de forma equidistante a

5, 10 e 20 cM, o que equivale a um total de 3600 simulações (100 amostras x 3 níveis de saturação x 6 tamanhos de população x 2 tipos de marcadores).

2.5 Procedimentos para simulação dos indivíduos da população F_2

A estratégia básica de simulação é percorrer ao longo dos cromossomos, realizando permutas em cada intervalo entre marcas adjacentes, de acordo com as distâncias dos marcadores, conforme descrito por Silva (2005). O processo de simulação da população F_2 seguiu os seguintes passos:

a) a partir do genoma simulado foram construídos os genótipos parentais homozigotos e contrastantes para os marcadores, de modo que a geração F_1 estivesse em aproximação para todos os marcadores;

b) a partir do genótipo da geração F_1, foram gerados gametas para a formação dos indivíduos das populações F_2. 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 em uma região entre marcadores adjacentes foi dada de acordo com a distância destes marcadores no genoma simulado. Ressalta-se que a maior distância implica maior possibilidade de ocorrência de recombinação.

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

2.6 Análise Genômica – Mapeamento

Após a obtenção dos dados, seguiram-se as demais etapas do processo de mapeamento da população F_2, como descrito a seguir:

2.6.1 Análise de segregação de locos individuais

Foram aplicados testes de qui-quadrado (x^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 1:2:1.

A estatística qui-quadrado é dada por:

$$x^2 = \sum_{i-1}^{n} \left[\frac{(\mathrm{Obs}_i - \mathrm{Esp}_i)^2}{\mathrm{Esp}_i} \right]$$

Onde:

x^2 é o 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 específica (1:2:1) para cada loco, foi testada a 5% de probabilidade (erro tipo I).

2.6.2 Estimação da porcentagem de recombinação, determinação dos grupos de ligação e ordenamento das marcas

Após a aplicação dos testes de segregação, seguiu-se a etapa da estimação da percentagem de recombinação entre pares de marcas utilizando o método da máxima verossimilhança como descrito por Schuster e Cruz (2008). No presente estudo foram utilizados, respectivamente, valores de 3 e de 30% para LOD_{min} e r_{max}, bem como utilização da propriedade transitiva.

Após a formação dos grupos de ligação empregou-se o método da SARF (*Sum of Adjacent Recombination Fractions*), para determinar a melhor ordem das marcas nos grupos. Após a realização de todos os passos descritos até então, foram formados os grupos de ligação para a população simulada, utilizando como medida de distância a porcentagem de frequência de recombinação.

2.7 Comparação dos genomas

O passo seguinte foi à comparação dos grupos de ligação formados para todas as populações simuladas com aqueles grupos de ligação estabelecidos no genoma simulado, como descrito a seguir. Foi analisado o número de grupos de ligação obtidos; o tamanho dos grupos de ligação; as distâncias médias entre marcas adjacentes nos grupos de ligação; as variâncias das distâncias entre marcas adjacentes nos grupos de ligação; o estresse (expressa o grau de concordância dos valores de distância entre cada par de marcas adjacentes nos grupos de ligação, simulados em relação às distâncias nos respectivos pares de marcas no genoma de referência); inversão de posição dos marcadores, verificada pela correlação de Spearman.

2.7.1 Números de grupos de ligação e marcas por grupo

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

2.7.2 Tamanho do grupo de ligação

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

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

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

2.7.3 Média das distâncias entre marcadores adjacentes no grupo de ligação

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

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

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

2.7.4 Variância das distâncias entre marcas adjacentes

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

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

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

2.7.5 Correlação de Spearman

A correlação de Spearman, também conhecida como correlação de *rank*, é utilizada quando não é possível mensurar variações contínuas, como variáveis x e y nos n membros de uma população. Contudo, 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. Assim, sua utilização na análise de genomas foi proposta por Ferreira et al., (2006), conforme descrito a seguir:

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

Em que:

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

2.7.6 Estresse

O coeficiente de estresse (S) é utilizado como medida de adequação das distâncias estimadas em relação às verdadeiras distâncias determinadas no genoma de referência. De acordo com Ferreira et al., (2006), o coeficiente de estresse é dado por:

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

Em que: S é o valor estimado do estresse, em percentagem, para o grupo de ligação do genoma analisado; d_{0k} é a distância entre marcas adjacentes m_k e m_{k+1}

no grupo de ligação do genoma analisado; 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).

Foram comparadas pelo teste de Scott Knott, a 5% de probabilidade (erro tipo I) as médias das variáveis: tamanho do grupo de ligação, distância média de marcas adjacentes, variância e estresse para cada grupo de ligação obtido para vários tamanhos de população. 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.

Todas as análises foram realizadas utilizando o programa ONEMAP (MARGARIDO et al., 2007) e as comparações por meio do programa computacional R (R DEVELOPMENT CORE TEAM, 2011).

3 RESULTADOS E DISCUSSÃO

Para determinar qual a melhor combinação de variáveis para a construção de mapas genéticos deve-se levar em consideração alguns critérios pré-estabelecidos que indiquem a confiabilidade do mapa resultante, como o número de indivíduos e nível de saturação.

De acordo com Bhering e Cruz (2008), um fator a desqualificar uma determinada população para análise é a junção de grupos de ligação, podendo ser total, em que um grupo inteiro se liga a outro grupo inteiro ou parcial, quando um grupo de ligação se liga à parte de outro grupo de ligação. Assim sendo, o número de grupos de ligação formados pode ser diferente do esperado para a espécie estudada.

Neste estudo, o número esperado era de 10 grupos de ligação, que foi o número que se tinha no genoma original usado para a simulação de genitores e populações.

A distância entre marcas medida pela frequência de recombinação deve ser determinada com a maior acurácia possível a fim de se determinar a posição relativa da marca no genoma e sua relação com outras marcas e genes ligados. No presente estudo a distância média variou de acordo com o nível de saturação do genoma simulado, sendo de 5, 10 e 20 cM. Por fim, a ocorrência ou não de inversões da ordem dos marcadores também pode ser considerada como indicativo de confiabilidade na construção de mapas genéticos.

Considerando o número de grupos de ligação formados para as 100 amostras analisadas, observou-se um número diferente do esperado, para as amostras com população de 100 indivíduos nos três níveis de saturação, sendo os valores obtidos de 82, 84 e 86 para os níveis de saturação de 5, 10 e 20 cM, respectivamente utilizando marcadores codominantes (Tabela 1). Este fato foi devido ao pequeno número de indivíduos existentes na população que não representam muito bem a diversidade de gametas produzidos pelos parentais alterando significativamente o número de grupos de ligação obtidos com a análise.

No entanto, para populações com tamanhos de 200, 300, 500, 800 e 1000 indivíduos, foi verificada a formação correta dos 10 grupos para todas as repetições, independente do nível de saturação considerado. Sendo assim, o número de grupos

de ligação recuperados tendeu a ser maior com o aumento no número de indivíduos. Estes resultados estão de acordo com os encontrados por Bhering e Cruz (2008). Estes autores verificaram que populações constituídas com 200 indivíduos foram suficientes para recuperar as informações originais do genoma de forma satisfatória, na construção de mapas moleculares.

Pode-se inferir que é necessário um tamanho populacional de no mínimo de 200 indivíduos para a obtenção de mapas de ligação com o número correto de grupos de ligação em populações $F_{2.}$

Colaborando com este resultado verificou-se que o tamanho médio do genoma analisado para todos os tratamentos foi bastante próximo do valor esperado de 100 cM, independente do tamanho da população e nível de saturação do genoma. Do mesmo modo, a distância média entre marcas adjacentes apresentaram valores praticamente iguais aos esperados em relação ao nível de saturação do genoma (5, 10 e 20 cM). Apesar de diferenças significativas terem sido observadas em ambos os casos, os resultados sugerem que estas se devem a questão de amostragem, especialmente ocasionado pela redução do número de marcadores nos genomas menos saturados.

Não se verifica uma tendência de aumento de precisão com o aumento do tamanho da população de mapeamento, mas sim uma variação ao acaso das médias. Silva et al., (2007) afirmam que populações com pequeno número de indivíduos não proporcionam uma boa amostra da diversidade gamética total dos genitores.

Analisando-se os níveis de saturação em conjunto para tamanho e distância, conclui-se que independente do nível de saturação, populações com tamanho igual ou superior a 200 indivíduos podem ser utilizadas no processo de mapeamento genético.

Por outro lado, a correlação de Spearman mostrou que uma vez definido o número de grupos de ligação a ordenação correta de todas as marcas moleculares independe do tamanho da população e de seu nível de saturação, apresentando como resultado uma correlação igual à unidade e consequentemente indicando a correta ordem de marcas para todos os casos analisados. Estes resultados diferem daqueles encontrados por Ferreira et al., (2006) em seu estudo sobre o tamanho e

tipo de população na precisão dos mapas genéticos, onde observaram inversões na ordem dos marcadores em todos os tipos de população.

Valores de correlação de Spearman iguais à unidade indicam que a ordem das marcas no grupo de ligação obtido no mapeamento das populações F_2 não foi alterada em relação à ordem previamente conhecida no genoma o qual foi utilizado para a geração das populações. Entretanto, se os valores da correlação de Spearman são menores que uma unidade, indica que a ordem das marcas no grupo de ligação obtido no mapeamento das populações segregantes foi alterada em relação à ordem previamente conhecida no genoma utilizado para a geração das populações. Assim, é possível afirmar que nenhuma das 100 repetições apresentou inversão de marcas nas populações com tamanhos de 100, 200, 300, 500, 800 e 1000 indivíduos.

Ao analisar o efeito do tamanho da população, verifica-se redução nos valores de variância e estresse médio em relação às distâncias entre marcas adjacentes à medida que se aumenta o tamanho da população segregante, em todos os níveis de saturação do genoma. Por fim, estes resultados indicam que uma população de 200 indivíduos é suficiente para obtenção de mapas de ligação com número de grupos de ligação esperado, ordenamento correto das marcas, tamanho médio de grupos de ligação e distância média entre marcas coerentes. O valor destas variáveis demonstra a existência de uma variação nas estimativas das distâncias entre marcas adjacentes, a qual tende a diminuir com o aumento do tamanho da população de mapeamento. Desta forma, para todos os níveis de saturação, o aumento do tamanho populacional permite a obtenção de mapas de ligação mais fidedignos.

Os resultados para os marcadores do tipo dominantes estão apresentados na Tabela 2. Primeiramente, considerando o número de grupos de ligação formados no nível de saturação de 5 cM, observa-se o número correto de 10 grupos apenas para populações com mais de 300 indivíduos. Enquanto que em populações com 100 e 200 indivíduos, apenas 36 e 92 repetições, respectivamente, apresentaram o número correto de grupos de ligação. No entanto, para populações com tamanho maior que 300 indivíduos, foi possível obter a formação dos 10 grupos de ligação em todas as repetições.

À medida que o número de indivíduos aumenta o número de grupos de ligação recuperado tende a ser igual ao número de grupos de ligação do genoma original. Já para os níveis de saturação de 10 e 20 cM, a recuperação dos grupos de ligação só foi possível em populações com tamanho maior que 200 indivíduos. Nesta ocasião, os valores das repetições em que houve a formação dos 10 grupos de ligação, foram respectivamente, 84 e 86 para os níveis de saturação de 10 e 20 cM.

Estes resultados podem ser explicados pelo aumento no nível da saturação. Era esperado que populações mais saturadas apresentassem um maior número de repetições com o número correto de grupos de ligação recuperados. No entanto, para o nível de saturação de 5 cM um tamanho populacional maior é necessário para obter o número correto de grupos de ligação, provavelmente devido a maior dificuldade no agrupamento das marcas quando utilizado marcadores dominantes. No entanto, verifica-se que com aumento do tamanho da população o número de grupos de ligação recuperados tende a ser igual ao número de grupos de ligação do genoma simulado.

Assim como para marcadores codominantes a correlação de Spearman mostrou o ordenamento correto de todas as marcas moleculares em todas as repetições que apresentaram o número correto de grupos independe do tamanho da população e de seu nível de saturação, apresentando como resultado uma correlação igual à unidade. Com isso, é possível afirmar que nenhuma das 100 repetições apresentou inversão de marcas nas populações com tamanhos de 100, 200, 300, 500, 800 e 1000 indivíduos nos três níveis de saturação. Resultados semelhantes foram encontrados por Salgado et al., (2011) quando estudavam o uso da variância como metodologia alternativa para integração de mapas genéticos.

Quanto ao tamanho médio dos grupos de ligação e a distância média entre marcas não houve diferença significativa em relação ao tamanho da população para o nível de saturação de 5 e 10 cM e para o nível de 20 cM esta não acompanhou o aumento do número de indivíduos. Este resultado indica não haver uma relação direta entre estas variáveis e o aumento do tamanho populacional, sendo este resultado provavelmente produto do erro de amostragem.

Semelhante aos marcadores codominantes, a variância das distâncias entre marcas e o estresse mostraram que em populações com tamanho de 300 indivíduos

foi suficiente para se obter mapas de ligação fidedignos em população F_2 usando marcadores dominantes com nível de saturação de 5 cM. Entretanto, para os níveis de saturação de 10 e 20 cM, populações com 200 indivíduos são suficiente para obtenção de mapas mais confiáveis. Estes resultados podem ser explicados

O valor destas variáveis diminuiu significativamente com o aumento do tamanho populacional em todos os níveis de saturação do genoma. O aumento do número de indivíduos na população F_2 permite a obtenção de mapas de ligação com maior acurácia.

4 CONCLUSÕES

1 O tamanho ideal de populações F_2 para mapeamento genético deve ser de no mínimo 200 indivíduos quando os marcadores são do tipo codominante independente do nível de saturação.

2 Para marcadores do tipo dominante, estes valores variam de 200 indivíduos para 10 e 20 cM e de 300 indivíduos para 5 cM.

3 Populações de mesmo tamanho tendem a produzir mapas com maior acurácia em níveis de saturação mais elevados.

4 Para todos os níveis de saturação o aumento do tamanho da população permite a obtenção de mapas de ligação com maior acurácia.

REFERÊNCIAS

BHERING, L. L.; CRUZ, C. D. Tamanho da população ideal para mapeamento genético em famílias de irmãos completos. **Pesquisa Agropecuária Brasileira,** Brasília, v.43, p.379-385, 2008.

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

CRUZ, E. M. **Efeito da saturação e do tamanho de populações F_2 e de retrocruzamento sobre a acurácia do mapeamento genético.** 2006. 119p. Tese (Doutorado) - Universidade Federal de Viçosa, Viçosa.

FALEIRO, F. G.; SCHUSTER, I.; RAGAGNIN, V. A.; CRUZ, C. D.; CORRÊA, R. X.; MOREIRA, M. A.; BARROS, E. G. Caracterização de linhagens endogâmicas recombinantes e mapeamento de locos de características quantitativas associados a ciclo e produtividade do feijoeiro-comum. **Pesquisa Agropecuária Brasileira,** Brasília, v.38, p.1387-1397, 2003.

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

GOOD GOD, P. I. V. **Mapeamento genético em famílias de meio irmãos por simulação computacional.** 2008. 114p. Tese (Doutorado) – Universidade Federal de Viçosa, Viçosa.

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

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

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

R DEVELOPMENT CORE TEAM. **R: a language and environment for statistical computing.** Vienna, Austria: R Foundation for Statistical Computing. Disponível em: www.R-project.org. Acesso em: 10 nov. 2011.

ROCHA, R. B.; PEREIRA, J. F.; CRUZ, C. D.; QUEIROZ, M. V.; ARAÚJO, E. F. O mapeamento genético no melhoramento de plantas: teoria e aplicações inseridas em um programa de melhoramento de melhoramento de plantas. **Biotecnologia, Ciência e Desenvolvimento,** v. 30, p.27-32, 2003.

SALGADO, C. C.; CRUZ, C. D.; NASCIMENTO, M. BARRERA, C. F. S. O uso da variância como metodologia alternativa para integração de mapas genéticos. **Pesquisa Agropecuária Brasileira,** v.46, p.66-73, 2011.

SCHUSTER, I.; CRUZ, C.D. **Estatística genômica aplicada a populações derivadas de cruzamentos controlados.** Viçosa: UFV, 2008. 568p.

SILVA, L. C. **Simulação do tamanho da população e da saturação do genoma para mapeamento genético de RILs.** 2005. 132p. Dissertação (Mestrado) - Universidade Federal de Viçosa, Viçosa.

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

Tabela 1. Número de grupos de ligação e valores médios de correlação, tamanho, distância, variância e estresse dos grupos de ligação em função do tamanho de população e nível de saturação do genoma em populações F_2 utilizando marcadores codominantes.

SG[(1)]	TP[(2)]	NRA[(3)]	Tamanho	Distância	Correlação	Variância	Estresse
5	100	82	100,023 a	5,000 a	1	2,337 a	27,093 a
	200	100	99,508 b	4,975 b	1	1,178 b	18,672 b
	300	100	99,402 b	4,970 b	1	0,794 c	14,617 c
	500	100	99,335 b	4,966 b	1	0,466 d	11,915 d
	800	100	99,294 b	4,966 b	1	0,291 e	9,404 e
	1000	100	99,320 b	4,968 b	1	0,237 e	8,309 f
10	100	84	99,471 b	9,948 b	1	4,449 a	18,666 a
	200	100	99,766 a	9,977 a	1	2,320 b	13,418 b
	300	100	99,797 a	9,978 a	1	1,596 c	10,819 c
	500	100	99,689 b	9,971 b	1	0,926 d	8,045 d
	800	100	99,682 b	9,967 b	1	0,592 e	6,052 e
	1000	100	99,813 a	9,981 a	1	0,475 f	5,716 e
20	100	86	99,435 c	19,914 c	1	8,597 a	12,588 a
	200	100	100,290 a	20,058 a	1	4,049 b	9,507 b
	300	100	100,397 a	20,079 a	1	2,831 c	7,788 c
	500	100	100,013 b	20,003 b	1	1,618 d	5,878 d
	800	100	100,324 a	20,064 a	1	1,021 e	4,654 e
	1000	100	100,367 a	20,073 a	1	0,921 f	4,296 e

[(1)]SG = Saturação do genoma; [(2)]TP = Tamanho da população; [(3)]NRA = 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 Scott Knott a 5% de probabilidade.

Tabela 2. Número de grupos de ligação e valores médios de correlação, tamanho, distância, variância e estresse em função do tamanho de população e nível de saturação do genoma em populações F_2 utilizando marcadores dominantes.

SG[1]	TP[2]	NRA[3]	Correlação	Tamanho	Distância	Variância	Estresse
5	100	36	1	100,800 a	5,039 a	5,213 a	38,217 a
	200	92	1	99,611 a	4,981 a	2,475 b	26,111 b
	300	100	1	99,537 a	4,977 a	1,587 b	22,007 c
	500	100	1	99,564 a	4,977 a	0,955 c	16,773 d
	800	100	1	99,551 a	4,977 a	0,586 c	13,527 e
	1000	100	1	99,472 a	4,976 a	0,472 c	12,358 e
10	100	84	1	100,115 a	10,012 a	9,814 a	25,230 a
	200	100	1	99,667 a	9,968 a	4,865 b	18,748 b
	300	100	1	99,902 a	9,991 a	3,195 c	15,120 c
	500	100	1	100,224 a	10,022 a	1,866 d	12,042 d
	800	100	1	99,818 a	9,981 a	1,176 d	9,433 e
	1000	100	1	99,688 a	9,967 a	0,943 d	8,409 e
20	100	86	1	92,448 b	19,269 b	14,994 a	18,685 a
	200	100	1	99,939 a	20,013 a	8,003 b	13,032 b
	300	100	1	100,247 a	20,050 a	5,503 c	10,105 c
	500	100	1	100,561 a	20,113 a	3,333 d	8,417 d
	800	100	1	100,214 a	20,043 a	2,134 e	6,545 e
	1000	100	1	100,379 a	20,076 a	1,698 e	5,906 e

[1]SG = Saturação do genoma; [2]TP = Tamanho da população; [3]NRA = 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 Scott Knott a 5% de probabilidade.

CAPÍTULO lll

OTIMIZAÇÃO DO MAPEAMENTO GENÉTICO VEGETAL DE POPULAÇÕES DUPLO-HAPLÓIDE VIA SIMULAÇÃO COMPUTACIONAL

Otimização do mapeamento genético vegetal de populações duplo-haplóide via simulação computacional

Silvan Gomes de Brito[(1)], Diogo Gonçalves Neder[(2)], Mairykon Coelho da Silva[(3)], Eliane Cristina Arcelino[(3)], Alisson Esdras Coutinho[(1)] e Péricles de Albuquerque Melo Filho[(4)]

[(1)]Mestrando vinculado ao Programa de Pós-Graduação em Melhoramento Genético de Plantas do Departamento de Agronomia, Universidade Federal Rural de Pernambuco, Rua Dom Manuel de Medeiros, s/n – Dois Irmãos – Recife-PE, CEP: 52171- 900. Email: gomesilvapb@hotmail.com, alissonesdras@yahoo.com.br; [(2)]Professor Doutor do Departamento de Agroecologia e Agropecuária/CCAA, Universidade Estadual da Paraíba/Campus de Lagoa Seca, Sítio Imbaúba, s/n, Zona Rural, Lagoa Seca-PB, Cep: 58.117-000. Email: dgneder@ccaa.uepb.edu.br; [(3)]Aluno de Graduação do Departamento de Agronomia, Universidade Federal Rural de Pernambuco, Rua Dom Manuel de Medeiros, s/n – Dois Irmãos – Recife-PE, CEP: 52171- 900. Email: mairicon@hotmail.com, elianearcelino@gmail.com; [(4)]Professor Associado do Departamento de Agronomia, Universidade Federal Rural de Pernambuco, Rua Dom Manuel de Medeiros, s/n – Dois Irmãos – Recife-PE, CEP: 52171- 900. Email: pericles@depa.ufrpe.br.

RESUMO

O mapeamento genético é um passo necessário para entender a organização genômica e a relação entre genes e o fenótipo. Um dos principais problemas está em encontrar a ordem, o espaçamento correto dos marcadores em um mapa genético, assim como o número de indivíduos a compor uma população. Deste modo, o objetivo deste estudo foi avaliar o nível de saturação do genoma e o tamanho ideal de populações simulada duplo-haplóide para a construção de mapas de ligação mais confiáveis por meio de simulação computacional. Foram simulados genomas parentais e populações duplo-haplóide considerando marcadores moleculares do tipo dominante, espaçados de forma equidistante a 5, 10 e 20 cM. Os tamanhos das populações geradas foram de 100, 200, 300, 500, 800 e 1000 indivíduos, com dez grupos de ligação cada e 100 repetições por amostra. Procedeu-se a análise de todas as populações geradas obtendo um genoma analisado o qual foi comparado com o genoma simulado inicialmente. Observou-se que o tamanho ideal de populações duplo-haplóide para mapeamento genético foi de no mínimo 200, 500 e 1000 indivíduos para genomas saturados, medianamente saturados e com baixa saturação. Populações de mesmo tamanho tendem a produzir mapas com maior acurácia em níveis de saturação do genoma mais elevados.

Termos para indexação: grupos de ligação, nível de saturação, genoma, mapa de ligação, ordem de marcas.

Optimization of vegetal genetic mapping of double-haploid populations via computational simulation

Silvan Gomes de Brito[(1)], Diogo Gonçalves Neder[(2)], Mairykon Coelho da Silva[(3)], Eliane Cristina Arcelino[(3)], Alisson Esdras Coutinho[(1)] e Péricles de Albuquerque Melo Filho[(4)]

[(1)]Mestrando vinculado ao Programa de Pós-Graduação em Melhoramento Genético de Plantas do Departamento de Agronomia, Universidade Federal Rural de Pernambuco, Rua Dom Manuel de Medeiros, s/n – Dois Irmãos – Recife-PE, CEP: 52171- 900. Email: gomesilvapb@hotmail.com, alissonesdras@yahoo.com.br; [(2)]Professor Doutor do Departamento de Agroecologia e Agropecuária/CCAA, Universidade Estadual da Paraíba/Campus de Lagoa Seca, Sítio Imbaúba, s/n, Zona Rural, Lagoa Seca-PB, Cep: 58.117-000. Email: dgneder@ccaa.uepb.edu.br; [(3)]Aluno de Graduação do Departamento de Agronomia, Universidade Federal Rural de Pernambuco, Rua Dom Manuel de Medeiros, s/n – Dois Irmãos – Recife-PE, CEP: 52171- 900. Email: mairicon@hotmail.com, elianearcelino@gmail.com; [(4)]Professor Associado do Departamento de Agronomia, Universidade Federal Rural de Pernambuco, Rua Dom Manuel de Medeiros, s/n – Dois Irmãos – Recife-PE, CEP: 52171- 900. Email: pericles@depa.ufrpe.br.

ABSTRACT

Genetic mapping is a necessary step to understand the genomic organization and the relationship between genes and phenotypes. A major problem is to find the order, the correct spacing of the markers in a genetic map, and the number of individuals to compose a population. Thus, the objective of this study was to evaluate the saturation level of the genome and the optimal size of simulated double-haploid populations for the construction reliable linkage maps by means of computer simulation. Parental genomes and double-haploid populations were simulated considering dominant molecular markers, spaced equidistantly at 5, 10 and 20 cM. The sizes of the generated populations were 100, 200, 300, 500, 800 and 1000 individuals, with ten linking groups and 100 replicates per sample. It was proceeded the analysis of all generated population obtaining a genome which was compared with the first simulated genome. It was observed that the optimal size of double-haploid populations for genetic mapping has been at least 200, 500 and 1000 individuals for saturated genomes, medium unsaturated and low saturation. Populations of the same size tend to produce maps with greater accuracy in higher levels of genome saturation.

Index terms: linkage groups, saturation level, genome, linkage map, order marks.

1 INTRODUÇÃO

Estudos com enfoque molecular são praticados com objetivo de explorar a variabilidade genética no melhoramento de plantas através de técnicas biotecnológicas. Com o advento dos marcadores moleculares, tornou-se possível a construção de mapas de ligação para diversas espécies vegetais.

Na construção de mapas genéticos são utilizados diferentes tipos de populações segregantes derivada do cruzamento entre duas linhas puras (populações F_2, retrocruzamentos, F_1"pseudo-testcross"), assim como, linhagens endogâmicas recombinantes - RILs (*Recombinant Inbred Lines*) derivada de uma população F_2 por meio de sucessivas autofecundações, ou uma população de linhagens duplo-haplóide obtidas a partir da cultura de anteras ou de micrósporos de plantas F_1 (SCHUSTER e CRUZ, 2008) a fim de auxiliar o mapeamento de QTLs (*Quantitative Trait Loci* – Loco de Característica Quantitativa). Além disso, têm amplo emprego como ferramenta biotecnológica em estudos de variabilidade genética para o melhoramento das plantas cultivadas (TEIXEIRA et al., 2004; FRANCESCHINELLI et al., 2006; MELO et al., 2009), permitindo o mapeamento de genes de interesse agronômico.

Populações de duplo-haplóides têm sido utilizadas para o mapeamento de QTLs em várias espécies (FORSTER et al., 2000; RADOEV, et al., 2008; WANG et al., 2011;SEYMOUR et al., 2012).

O processo para se obter uma população de duplo-haplóide é bem mais rápido do que a obtenção de populações RILs (SEMAGN et al., 2010). Para duplo-haplóide é necessária apenas uma geração de recombinação para se alcançar a homozigose, enquanto que para linhas endogâmicas recombinantes são realizados sucessivos ciclos de autofecundação da planta F_1, no mínimo seis gerações.

A disponibilidade de mapas confiáveis depende de vários fatores tais como o tipo e tamanho da população, bem como tipo e número de marcadores (SILVA et al., 2007).

A resolução do mapa e a capacidade de se determinar a sequência de marcadores nele estão diretamente relacionadas ao tamanho da população, dada a importância de se estimar de forma adequada as distâncias entre os genes, a fim de estabelecer o ordenamento correto, bem como a formação de grupos de ligações

que reflitam o número básico de cromossomos da espécie (BHERING e CRUZ, 2008).

Ao definir o tipo de população para mapeamento, o conteúdo médio de informação associado à variância podem ser utilizados com critérios para estimar a frequência de recombinação (r) a fim de avaliar a confiabilidade do mapa genético (ROCHA et al., 2004).

Apesar da disponibilidade de vários estudos sobre mapeamento genético, há ainda a necessidade de estudos específicos relacionados com a determinação do número ideal de indivíduos numa dada população (CRUZ, 2006; BHERING e CRUZ, 2008).

O objetivo do presente estudo foi determinar o nível de saturação do genoma, bem como identificar o tamanho da população ideal a ser empregado no mapeamento genético de populações duplo-haplóides via simulação computacional.

2 MATERIAL E MÉTODOS

2.1 Simulação dos dados

Para a geração dos dados nas duas populações foi considerada uma espécie vegetal diplóide com 2n = 2x = 20 cromossomos utilizando o módulo de simulação do aplicativo computacional GQMOL (CRUZ, 2004), o qual permite gerar informações sobre genomas, genótipos de genitores e indivíduos de diferentes tipos de populações.

2.2 Simulação do genoma, níveis de saturação e tipos de marcas

Foram gerados três genomas com três níveis de saturação: saturado, medianamente saturado e pouco saturado, com intervalos entre marcas adjacentes de 5, 10 e 20 cM, representando um número total de marcas por genoma de 210, 110 e 60, respectivamente. Cada genoma composto por 10 grupos de ligação, com 100 cM em cada grupo e, portanto, com comprimento total de 1000 cM. Foram considerados os tipos de marcadores moleculares dominantes para populações duplo-haplóide devido a proporção de segregação esperada ser de 1:1, distribuídos de forma equidistante no genoma.

2.3 Simulação dos genitores

Para a simulação dos genitores na população duplo-haplóide, para cada nível de saturação do genoma foi simulado uma situação em que os pais eram homozigotos, sendo um dominante e outro recessivo.

2.4 Tamanho da população

Foram simuladas 100 amostras com 100, 200, 300, 500, 800 e 1000 indivíduos com 11 marcas por grupo de ligação distribuídas de forma equidistante a 5, 10 e 20 cM, o que equivale a um total de 1800 simulações (100 amostras x 3 níveis de saturação x 6 tamanhos de população).

2.5 Procedimentos para simulação dos indivíduos da população duplo-haplóide

A estratégia básica de simulação é percorrer ao longo dos cromossomos, realizando permutas em cada intervalo entre marcas adjacentes, de acordo com as distâncias dos marcadores, conforme descrito por Silva (2005). O processo de simulação da população duplo-haplóide seguiu os seguintes passos:

a) a partir do genoma simulado foram construídos os genótipos parentais homozigotos e contrastantes para os marcadores, de modo que a geração F_1 estivesse em aproximação para todos os marcadores;
b) a partir do genótipo da geração F_1, foram gerados gametas para a formação dos indivíduos das populações duplo-haplóide. 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 em uma região entre marcadores adjacentes foi dada de acordo com a distância destes marcadores no genoma simulado. Ressalta-se que a maior distância implica maior possibilidade de ocorrência de recombinação.

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

2.6 Análise Genômica – Mapeamento

Após a obtenção dos dados, seguiram-se as demais etapas do processo de mapeamento da população duplo-haplóide, como descrito a seguir:

2.6.1 Análise de segregação de locos individuais

Foram aplicados testes de qui-quadrado (x^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.

A estatística qui-quadrado é dada por:

$$x^2 = \sum_{i-1}^{n} \left[\frac{(\mathrm{Obs}_i - \mathrm{Esp}_i)^2}{\mathrm{Esp}_i} \right]$$

Onde:

x^2 é o 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 específica (1:1) para cada loco, foi testada a 5% de probabilidade (erro tipo I).

2.6.2 Estimação da porcentagem de recombinação, determinação dos grupos de ligação e ordenamento das marcas

Após a aplicação dos testes de segregação, seguiu-se a etapa da estimação da percentagem de recombinação entre pares de marcas utilizando o método da máxima verossimilhança como descrito por Schuster e Cruz (2008). No presente estudo foram utilizados, respectivamente, valores de 3 e de 30% para LOD_{min} e r_{max}, bem como utilização da propriedade transitiva.

Após a formação dos grupos de ligação empregou-se o método da SARF (*Sum of Adjacent Recombination Fractions*), para determinar a melhor ordem das marcas nos grupos. Após a realização de todos os passos descritos até então, foram formados os grupos de ligação para a população simulada, utilizando como medida de distância a porcentagem de frequência de recombinação.

2.7 Comparação dos genomas

O passo seguinte foi à comparação dos grupos de ligação formados para todas as populações simuladas com aqueles grupos de ligação estabelecidos no genoma simulado. Foi analisado o número de grupos de ligação obtidos; o tamanho dos grupos de ligação; as distâncias médias entre marcas adjacentes nos grupos de ligação; as variâncias das distâncias entre marcas adjacentes nos grupos de ligação;

o estresse (expressa o grau de concordância dos valores de distância entre cada par de marcas adjacentes nos grupos de ligação, simulados em relação às distâncias nos respectivos pares de marcas no genoma de referência); inversão de posição dos marcadores, verificada pela correlação de Spearman.

2.7.1 Números de grupos de ligação e marcas por grupo

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

2.7.2 Tamanho do grupo de ligação

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

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

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

2.7.3 Média das distâncias entre marcadores adjacentes no grupo de ligação

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

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

Em que: $\overline{d}$ é a distância média de dois marcadores adjacentes no grupo de ligação do genoma analisado, L é o tamanho do grupo de ligação do genoma

analisado e I é o número de intervalos entre marcas adjacentes, dado por m-1, onde m é o número de marcas no grupo de ligação.

2.7.4 Variância das distâncias entre marcas adjacentes

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

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

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

2.7.5 Correlação de Spearman

A correlação de Spearman, também conhecida como correlação de *rank*, é utilizada quando não é possível mensurar variações contínuas, como variáveis x e y nos n membros de uma população. Contudo, 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. Assim, sua utilização na análise de genomas foi proposta por Ferreira et al., (2006), conforme descrito a seguir:

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

Em que:

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

2.7.6 Estresse

O coeficiente de estresse (S) é utilizado como medida de adequação das distâncias estimadas em relação às verdadeiras distâncias determinadas no genoma de referência. De acordo com Ferreira et al.,(2006), o coeficiente de estresse é dado por:

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

Em que: S é o valor estimado do estresse, em percentagem, para o grupo de ligação do genoma analisado; d_{0k} é a distância entre marcas adjacentes m_k e m_{k+1} no grupo de ligação do genoma analisado; 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).

Foram comparadas pelo teste de Scott Knott, a 5% de probabilidade (erro tipo I) as médias das variáveis: tamanho do grupo de ligação, distância média de marcas adjacentes, variância e estresse para cada grupo de ligação obtido para vários tamanhos de população. 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.

Todas as análises foram realizadas utilizando o programa ONEMAP (MARGARIDO et al., 2007) e as comparações por meio do programa computacional R (R DEVELOPMENT CORE TEAM, 2011).

3 RESULTADOS E DISCUSSÃO

Vários tipos de populações controladas vêm sendo utilizado na construção de mapas de ligação cada vez mais saturados em busca de uma maior confiança no ordenamento de marcadores moleculares.

A veracidade das informações do mapa depende do tipo de população, do tipo de marcador molecular e principalmente do número de indivíduos considerados na obtenção da percentagem de recombinação entre pares de marcas (SALGADO et al., 2011) além de técnicas de análises estatísticas e computacionais para estimativa de ligação e de distâncias entre marcadores, bem como seu ordenamento nos cromossomos.

A qualidade de um mapa pode ser avaliada pela confiança no agrupamento e ordenamento dos marcadores, assim como a proporção do genoma representado pelo comprimento e a densidade total do mapa, tamanho dos intervalos entre os marcadores. Por outro lado, à medida que aumenta a cobertura de um mapa genético, o número de grupos de ligação aproxima-se do número de cromossomos.

Era esperada a formação de 10 grupos de ligação para os três níveis de saturação. Considerando o número de grupos de ligação formados para as 100 amostras analisadas, observou-se um número diferente do esperado para tamanhos de população de 100, 200, 300, 500 e 800 indivíduos nos três níveis de saturação (Tabela 1). É possível visualizar que as populações com 100 indivíduos apresentaram menor número de repetições com recuperação dos 10 grupos de ligação, com valores de 78, 42 e 4 para os níveis de saturação de 5, 10 e 20 cM, respectivamente.

Este fato pode ser explicado pelo pequeno número de indivíduos utilizado na constituição da população de duplo-haplóide não ter sido eficiente na localização de ligação entre marcas, o que acabou não representando muito bem a diversidade de gametas produzidos pelos parentais. Silva (2005), estudando o mapeamento em populações RILs, detectou número de grupos de ligação acima do esperado em populações com tamanho inferior a 100 indivíduos e atribuiu à divisão de grupos de ligação em consequência da não detecção de ligação entre marcas onde certamente existia. Deste modo, independente do nível de saturação, o uso de tamanho da

população com 100 indivíduos é inadequado para mapeamento em populações duplo-haplóide.

Conforme verificado, o número de grupos de ligação recuperados tende a ser igual ao número de grupos de ligação do genoma simulado à medida que o tamanho da população e a densidade do mapa genético aumentam. Assim, sugere-se que são necessárias populações com tamanho mínimo de 200, 500 e 1000 indivíduos, para construção de mapas de ligação em populações duplo-haplóide nos níveis de saturação de 5, 10 e 20 cM, respectivamente.

Colaborando com esses resultados, os valores da correlação de Spearman iguais à unidade nos 10 grupos de ligação obtidos no mapeamento das populações segregantes para os 3 níveis de saturação, indicam que a ordem das marcas não foi alterada em relação à ordem previamente estabelecida no genoma o qual foi utilizado para a geração das populações (Tabela 1). Nota-se que apesar das diferenças no número de repetições analisadas para cada tamanho de população, não foi verificada inversão das marcas, o que evidencia boa qualidade dos dados analisados. Entretanto, caso os valores de correlação de Spearman obtidos fossem menores que a unidade à indicação seria que a ordem das marcas nos grupos de ligação obtidos no mapeamento da população segregante teria sido alterada em relação à ordem do genoma de referência.

O tamanho de cada grupo de ligação esperado era de 100 cM, no entanto, se observa uma pequena variação no comprimento do genoma para os três níveis de saturação (Tabela 1). Quando comparado com o tamanho do grupo de ligação simulado, as diferenças são de 1,422 e 4,056 cM para o menor e maior valor, respectivamente. Apesar de todos os valores estarem situados acima do inicialmente esperado (100 cM), as médias para os tamanhos populacionais avaliados se aproximam do valor simulado independente do nível de saturação, não diferindo estatisticamente entre si pelo teste de Scott-Knott realizado com $P<0,05$. Estes resultados estão de acordo com os encontrados por Bhering e Cruz (2008). Os autores ressaltaram que não se pode fazer referência de um determinado tamanho da população, sendo necessária uma maneira adicional para observar o comportamento do comprimento médio dos grupos de ligação.

Comportamento semelhante é apresentado pela distância média entre marcas para o genoma da população simulada. Os valores esperados eram de 5, 10

e 20 cM, contudo, verifica-se que para este último, as distâncias médias foram variantes, de modo que para 5 e 10 cM, estatisticamente não houve diferença significativa entre as distâncias (Tabela 1). De modo geral, as distâncias médias entre os marcadores se aproximam dos valores esperados na medida em que o tamanho da população aumenta, assim como o tamanho médio dos grupos de ligação.

Esta aproximação e a inalterabilidade das estimativas nos níveis de saturação mais elevados podem ser explicadas em termos de precisão das estimativas de recombinação em populações com maiores tamanhos amostrais e do nível de resolução para alta saturação (GOOD GOD, 2008). Isso sugere que populações com maior número de indivíduos permitem uma melhor amostragem dos eventos de permuta dos gametas, contribuindo para estimativas mais precisas e acuradas.

A variância das estimativas das distâncias entre os marcadores também pode ser tomada como medida de análise da acurácia de mapas mais fidedignos.

Ainda na Tabela 1, é possível visualizar o comportamento da variância da distância média entre marcadores nos três níveis de saturação do genoma simulado. Como era de se esperar, os valores referentes à variância estão de acordo com as estimativas de tamanho e distâncias médias entre os marcadores.

Ao analisar os níveis de saturação em separado, observa-se uma diminuição da variância na medida em que o tamanho da amostra é ampliado. Esta redução é evidenciada pelas diferenças significativas dadas pelo teste de médias, em que as menores médias estão associadas às maiores populações, portanto, levando à maior precisão no mapeamento genético. Seja por exemplo, o nível de saturação de 5 cM, os valores da variância correspondem à 5,066 e 0,506 para os tamanhos de população com 100 e 1000 indivíduos, respectivamente.

Quando se considera um mesmo tamanho de população nos diferentes níveis de saturação, nota-se que o valor da variância será menor tanto quanto menor for o valor do nível de saturação do genoma (Tabela 1). Isto pode ser melhor entendido, quando se analisa, por exemplo, o tamanho da população de 100 indivíduos nos níveis de saturação de 5, 10 e 20 cM, onde os valores da variância são 5,066, 9,816 e 17,914, respectivamente. Alguns autores também observaram diminuição nos valores da variância na medida em que o tamanho da população e o

nível de saturação aumentaram, em estudos relacionados ao mapeamento de populações via simulação (SILVA, 2005; BARROS, 2007; BHERING e CRUZ, 2008; SALGADO, 2008). Portanto, quanto menor o valor de variância mais ajustados estarão os valores do real, indicando uma boa recuperação do genoma, bem como maior uniformidade na distribuição dos marcadores dentro dos grupos de ligação.

Colaborando com estes resultados, verifica-se que o estresse expressa o grau de concordância dos valores de variância da distância média entre cada par de marcadores nos grupos de ligação simulados em relação às distâncias nos respectivos pares de marcadores no genoma de referência.

De modo semelhante ao que ocorre com a variância, os valores do estresse tendem a diminuir à medida que o tamanho da população avaliada aumenta independente do nível de saturação do genoma (Tabela 1). A redução dos valores de estresse com o aumento do tamanho das populações dentro do nível de saturação do genoma é evidenciada pelas diferenças significativas quando comparadas pelo teste de Scott-Knot ($P<0,05$), de modo que, os maiores valores associados às menores populações, diferem significativamente das menores médias associadas às maiores populações para cada nível de saturação. Com essa diminuição, pode-se inferir que a qualidade dos dados analisados será mais confiável quanto maior for o número de indivíduos na população.

Ferreira et al., (2006) verificaram que os valores referentes ao estresse também diminuíam com o aumento no tamanho da população. Por fim, estes resultados indicam que populações com tamanho de mínimo 200, 500 e 1000 indivíduos, é suficiente para se obter mapas de ligação com número de grupos de ligação esperado, ordenamento correto das marcas, tamanho médio de grupos de ligação e distância média entre marcas coerentes .

5 CONCLUSÕES

1 O tamanho ideal de populações duplo-haplóide para mapeamento genético deve ser de 200, 500 e 1000 indivíduos quando os níveis de saturação do genoma são de 5, 10 e 20 cM, respectivamente.

2 Níveis de saturação mais elevados proporcionam mapas mais precisos, mais confiáveis em populações de mesmo tamanho.

3 O aumento no tamanho da população permitiu a obtenção de mapas de ligação com maior acurácia independente do nível de saturação.

REFERÊNCIAS

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

BHERING, L. L.; CRUZ, C. D. Tamanho da população ideal para mapeamento genético em famílias de irmãos completos. **Pesquisa Agropecuária Brasileira,** v.43, p.379-385, 2008.

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

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

FORSTER, B. P.; ELLIS, R. P.; THOMAS, W. T. B.; NEWTON, A. C.; TUBEROSA, R.; THIS, D.; EL-ENEIN, R. A.; BAHRI, M. H.; SALEM, M. B. The development and application of molecular markers for abiotic stress tolerance in barley. **Journal of Experimental Botany,** v.51, p.19-27, 2000.

FRANCESCHINELLI, E. V.; JACOBI, C. M.; DRUMMOND, M. G.; RESENDE, M. F. S. The genetic diversity of two Brazilian *Vellozia* (Velloziaceae) with different patterns of spatial distribution and pollination biology. **Annals of Botany,** v.97, p.585-592, 2006.

GOOD GOD, P. I. V. **Mapeamento genético em famílias de meio irmãos por simulação computacional.** 2008. 114p. Tese (Doutorado) – Universidade Federal de Viçosa, Viçosa.

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

MELO, R. A.; MENEZES, D.; RESENDE, L. V.; WANDERLEY JÚNIOR, L. J. G.; SANTOS, V. F.; MESQUITA, J. C. P; MAGALHÃES, A. G. Variabilidade genética em progênies de meios-irmãos de coentro. **Horticultura Brasileira,** v.27, p.325-329, 2009.

R DEVELOPMENT CORE TEAM. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Disponível em: www.R-project.org. Acesso em: 10 nov. 2011.

RADOEV, M.; BECKER, H. C.; ECKER, W. Genetic analysis of heterosis for yield and yield components in rapeseed (*brassica napus* l.) by quantitative trait locus mapping. **Genetics,** v.179, p.1547-1558, 2008.

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

SALGADO, C. C. **Integração de mapas genéticos.** 2008. 142p. Dissertação (Mestrado) - Universidade Federal de Viçosa, Viçosa.

SALGADO, C. C.; CRUZ, C. D.; NASCIMENTO, M. BARRERA, C. F. S. O uso da variância como metodologia alternativa para integração de mapas genéticos. **Pesquisa Agropecuária Brasileira,** v. 46, p.66-73, 2011.

SCHUSTER, I.; CRUZ, C.D. Estatística genômica aplicada a populações derivadas de cruzamentos controlados. Viçosa: UFV, 2008. 568p.

SEMAGN, K.; BJORNSTAD, A.; XU, Y. The genetic dissection of quantitative traits in crops. **Electronic Journal of Biotechnology,** v. 13, 2010. Disponível em:<http://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0717-45820100005000016&lng=es&nrm=iso>. Acesso em: 20 mai. 2012.

SEYMOUR, D. K. l. FILIAULTA, D. L.; HENRYA, I. M.; MONSON-MILLERA, J.; RAVIA, M.; Andy PANGA, COMAIA, L.; W. L. CHANA, S. W. L.; MALOOF, J. N. **Rapid creation of Arabidopsis doubled haploid lines for quantitative trait locus mapping.** v. 109, p. 4227-4232, 2012. Disponível em: < http://www.pnas.org/content/early/2012/02/23/1117277109.short>. Acesso em: 03 mai. 2012.

SILVA, L. C. **Simulação do tamanho da população e da saturação do genoma para mapeamento genético de RILs.** 2005. 132p. Dissertação (Mestrado) - Universidade Federal de Viçosa, Viçosa.

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

TEIXEIRA, H.; VIEIRA, MARIA DAS GRAÇAS, G. C.; MACHADO, J. C. Marcadores RAPD na análise da diversidade genética de isolados de *Acremonium strictum*. **Fitopatologia Brasileira,** v.29, p.651-655, 2004.

WANG, N.; QIAN, W.; SUPPANZ, I.; WEI, L.; MAO, B.; LONG, Y.; MENG, J.; MULLER, A. E.; JUNG, C. Flowering time variation in oilseed rape (*Brassica napus* L.) is associated with allelic variation in the FRIGIDA homologue BnaA.FRI.a. **Journal of Experimental Botany,** v.62, p. 5641–5658, 2011.

Tabela 1. Número de grupos de ligação e valores médios de correlação, tamanho, distância, variância e estresse dos grupos de ligação em função do tamanho de população e nível de saturação do genoma em populações duplo-haplóide utilizando marcadores dominantes.

SG	TP	NRA	Correlação	Tamanho	Distância	Variância	Estresse
5	100	78	1	104,056 a	5,203 a	5,066 a	46,060 a
	200	100	1	102,949 a	5,147 a	2,353 b	36,258 b
	300	100	1	103,490 a	5,174 a	1,571 b	32,500 c
	500	100	1	102,794 a	5,140 a	0,963 b	29,821 d
	800	100	1	103,255 a	5,163 a	0,623 b	27,795 e
	1000	100	1	103,112 a	5,156 a	0,506 b	26,722 e
10	100	42	1	103,879 a	10,388 a	9,816 a	28,248 a
	200	93	1	102,721 a	10,272 a	4,931 b	19,705 b
	300	99	1	103,145 a	10,314 a	3,233 c	18,138 b
	500	100	1	103,459 a	10,346 a	2,011 d	15,191 c
	800	100	1	103,268 a	10,327 a	1,248 d	13,374 c
	1000	100	1	103,211 a	10,321 a	1,017 d	12,835 c
20	100	4	1	103,828 a	21,021 a	17,914 a	23,995 a
	200	70	1	101,422 a	20,300 b	9,386 b	12,926 b
	300	83	1	101,503 a	20,311 b	7,328 c	12,330 b
	500	95	1	101,967 a	20,393 b	5,030 d	10,455 b
	800	99	1	101,824 a	20,365 b	3,030 e	8,417 c
	1000	100	1	101,560 a	20,312 b	2,483 e	7,187 c

(1)SG = Saturação do genoma; (2)TP = Tamanho da população; (3)NRA = Número de repetições avaliadas.

Médias seguidas pela mesma letra nas colunas, dentro de cada nível de saturação do genoma, não diferem entre si pelo teste de Scott Knott a 5% de probabilidade.