CAIO CÉSIO SALGADO

INTEGRAÇÃO DE MAPAS GENÉTICOS

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

VIÇOSA
MINAS GERAIS – BRASIL
2008

CAIO CÉSIO SALGADO

INTEGRAÇÃO DE MAPAS GENÉTICOS

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

APROVADA: 28 de julho de 2008.

_____	_____
Prof. José Marcelo Soriano Viana (Co-orientador)	Prof. Pedro Crescêncio Souza Carneiro (Co-orientador)
_____	_____
Dr. Willian Silva Barros	Drª. Eveline Teixeira Caixeta

_____
Prof. Cosme Damião Cruz
(Orientador

AGRADECIMENTOS

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

Aos meus pais, Fernando Amir Salgado e Antonia Bressani Salgado, e meu irmão Rider, pelo apoio, amizade e compreensão em todos os momentos.

Aos meus tios Neomisa e Tenir por estarem sempre presente em minha vida.

Ao professor Cosme Damião Cruz, pela amizade, paciência, confiança e incentivo.

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

Aos amigos do laboratório de Bioinformática, em especial ao Márcio, Lívia, Leonardo, Adésio, Tatiana, Moysés, Carol, Felipe, Rafael, Danielle, Eliel, Willian e Tales pelo convívio agradável durante a realização deste curso.

Ao Lucas, um grande amigo, meu irmão de coração.

Aos amigos Fábio e Gustavo pela convivência, cumplicidade, companheirismo.

Aos amigos do curso de pós-graduação: Josi, Ana Paula, Alexandre, André, Juliana, Jaqueline, Marcelo, José Ângelo, Gustavo, Éder.

Aos amigos de outras datas que me apoiaram e me incentivaram nesta empreitada: Flávia, Patrícia, Lisete, Vander , Soraya, Fabiola e Zuy.

A FAPEMIG (Fundação de Amparo à Pesquisa do Estado de Minas Gerais), pela concessão da bolsa de estudos.

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

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

Aos amigos do GenMelhor: Cândida, Andréia, Gustavo, Janaína, Gilberto, Tiago.

A mãe Joaninha, aonde quer que esteje, pelo carinho, amor e amizade que sempre teve comigo.

SUMÁRIO

RESUMO

SALGADO, Caio Césio, M.Sc., Universidade Federal de Viçosa, julho de 2008. **Integração de Mapas Genéticos.** Orientador: Cosme Damião Cruz. Co-orientadores: José Marcelo Soriano Viana e Pedro Crescêncio Souza Carneiro.

O mapeamento genético facilita o trabalho de melhoramento uma vez que uma ou mais marcas do genótipo podem ser associadas a genes controladores de características qualitativas e quantitativas (QTL). Mapas genéticos para diversas espécies tem sido construídos por diferentes grupos de pesquisadores com diferentes tipos de marcadores e diferentes tipos de populações. Uma maneira de gerar mapas mais saturados para as diversas culturas seria a integração dos mapas genéticos já existentes. A chave para integração de mapas distintos é a presença de marcas comuns entre os mapas. Somente quando há um número mínimo de marcas comuns entre os diferentes mapas, estes podem ser integrados. Deste modo, o objetivo deste trabalho foi desenvolver um processo de integração de mapas genéticos e testar a eficiência do processo de integração. Para isso foram gerados e analisados dados a partir da simulação de genoma e de populações. Um dos fatores de fundamental importância para se obter dados consistentes em um trabalho de mapeamento é o tamanho da amostra ou população a ser trabalhada. Com base nestes dados simulados, avaliou-se o tamanho ótimo de populações para estudo de integração de mapas genéticos. Para estudo e obtenção dos mapas consenso foram simulados genomas parentais e amostras de populações F_2 co-dominantes, F_2 dominantes e de retrocruzamentos. As amostras geradas foram de tamanho 100, 200 e 400 indivíduos com três grupos de ligação cada e 11 marcas moleculares co-dominantes e dominantes espaçadas a 5, 10 e 15 centimorgans por grupo de ligação. Foram realizadas 10 repetições por amostra, sendo que destas, cinco foram utilizadas para construção de mapas consenso com análise multiloco e outras cinco sem análise multiloco. Concluiu-se que a obtenção dos mapas consenso se torna mais eficiente com aumento do tamanho da população. Um tamanho populacional de 200 indivíduos seria o suficiente para resgatar as informações originais. Para estudo da obtenção do mapa integrado efetivo foram simuladas F_2 co-dominante e retrocruzamento com tamanhos de 100, 200 e 400 indivíduos, com 21 marcas por grupo de ligação e

marcadores eqüidistantes a 5 cM, em um total de quatro simulações para F_2 co-dominante e quatro para retrocruzamentos. Cada genoma simulado foi fragmentado em quatro novos mapas de modo que, foram obtidos três mapas com oito marcadores e um com nove marcadores, sendo que cada um deles constando quatro marcadores que são âncoras entre os quatro mapas. Estes novos mapas foram alinhados, ordenados, integrados e em seguida comparados com o mapa de origem. Concluiu-se que os processos de ordenamento e integração são eficientes para obtenção do mapa integrado e, também, que o tamanho da população exerce influência sobre o mapa e as medidas de distâncias entre as marcas.

ABSTRACT

SALGADO, Caio Césio, M.Sc., Universidade Federal de Viçosa, July 2008. **Integrated Genetic Maps.** Adviser: Cosme Damião Cruz Co-Advisers: José Marcelo Soriano Viana and Pedro Crescêncio Souza Carneiro.

The genetic mapping facilitates the breeding work once one or more marks of the genotype can be associated to controlling genes of qualitative and quantitative characteristics (QTL). Genetic maps for several species have been built by different groups of researchers with different molecular markers and populations. A way to generate maps more saturated for those species would be the integration of the existent maps. The key to integrate different maps is the presence of common marks among them. Only when there are a minimum number of common marks among the different maps, these can be integrated. This way, the objective of this work was to develop a process of integration of genetic maps and to test the efficiency of this process. Data from the simulation of genome and populations were generated and analyzed. A important factor to obtain solid data in a mapping work is the sample or population size. Based in these simulated data it was evaluated the optimum population size to study the integration of genetic maps. To obtain and study the consensus maps, parental genomes and samples of co-dominant F2, dominant F2 and backcrosses populations were simulated. The generated samples had 100, 200, 400 individuals with 3 linkage groups each and 11 dominant and co-dominant molecular marks spaced by 5, 10 and 15 centiMorgans. 10 repetitions were accomplished by sample, five used to construct the consensus maps with analysis multilocus and other five without analysis multilocus. It was concluded that the obtention of the consensus maps becomes more efficient with the increase of the population size. A population size of 200 individuals would be enough to rescue the original information. To study the obtention of the effective integrated map samples of co-dominant F2 and backcrosses populations were simulated. The generated samples had 100, 200 and 400 individuals, 21 marks for linkage group and markers equidistant by 5 cM, in a total of four simulations for co-dominant F2 and four for backcrosses. Each simulated genome was fragmented in four new maps which three maps had eight markers and one had nine markers, each one of these maps containing four markers that are anchors among the four maps. These new maps were aligned, orderly and integrated and then compared with the original map. It was concluded that

the process of ordainment and integration are efficient to obtain the integrated map. The population size exercises influence on the map and the distances measures among the marks.

1. INTRODUÇÃO GERAL

A genômica pode ser entendida como a denominação da ciência que estuda o genoma de forma completa, integrando diversas áreas da genética como a genética mendeliana, citogenética, genética molecular, genética de populações, genética quantitativa; e outras áreas do conhecimento, como a ciência da computação e sistema automatizados.

Marcadores de DNA têm permitido a construção de mapas genéticos para várias espécies vegetais e animais. Estes mapas podem atingir alto grau de saturação devido à disponibilidade de grande número de marcas genéticas disponíveis. Para construção de mapas genéticos é utilizada a junção de técnicas que vão desde a obtenção e caracterização dos genitores, o desenvolvimento de populações segregantes, identificação de genótipos em locos por meio de técnicas de biologia molecular, estimativas das freqüências de recombinação e ordenamento das marcas ao longo dos grupos de ligação.

Mapas genéticos se tornaram ferramenta muito útil em vários campos da pesquisa genética, uma vez que permitem a visualização, mesmo que de forma relativa, da organização dos genes nos cromossomos.

No estudo genético de determinada característica é de interesse do pesquisador conhecer o número de genes e alelos envolvidos no controle da sua expressão, localização e posição relativa desses genes nos cromossomos, assim com a sua relação com outros genes.

Dentre as principais utilizações do mapa genético destacam-se a localização e o mapeamento de QTL (Quantitative Traits Loci), localização de genes ligados a doenças, estresse abiótico e mapeamento comparativo de genomas de diferentes espécies e também estudos de sintenia e clonagem de genes

O grande conteúdo de informações de ligação que está disponível, para vários organismos, tem criado a necessidade de integração dos mapas que foram obtidos independentemente. Mapas genéticos para diversos tipos de espécies tem sido

construídos por diferentes grupos de pesquisadores com diferentes tipos de marcadores e diferentes tipos de populações.

A chave para integração de mapas distintos é a presença de marcas comuns entre os mapas. Somente quando há um número mínimo de marcas comuns entre os mapas, estes podem ser integrados, e assim diferentes marcas podem ser localizadas em um grupo de ligação. Mas sempre que diferentes mapas são agrupados, alguns problemas podem surgir. A precisão das estimativas de freqüências de recombinação bem como o tipo de informação pode variar entre os conjuntos de dados. Por exemplo, os mapas a serem integrados podem ser baseados em diferentes tipos de população (F_2, retrocruzamento, entre outras) com variados tamanhos. Portanto, deve-se, pesar estas informações e obter um mapa representativo com um mínimo de tensão interna.

Por meio da incorporação da ciência da computação e de sistemas automatizados ao processo de melhoramento, as técnicas de simulação têm ajudado os melhoristas a reduzir o tempo e gastos com experimentos, pois a simulação abrange dois grandes aspectos que são a modelagem e a experimentação e têm como dois grandes objetivos compreender o sistema existente e prescrever recomendações que possam ser generalizadas para quaisquer situações. Além de apresentar menor custo e maior rapidez, garante certeza da direção e sentido das respostas

O presente trabalho visa o estudo e desenvolvimento de método apropriado para a integração de mapas genéticos entre populações de diferentes tipos e tamanhos a partir das estimativas de freqüências de recombinação. Este estudo foi realizado por meio de simulação de dados em computador. Os objetivos específicos do trabalho, para melhor compreensão do fenômeno estudado foram:

a) Avaliar a eficácia do processo da geração de mapas consensos, em que apenas marcadores comuns aos mapas estudados são considerados;

b) Subsidiar o entendimento de como as estimativas das distâncias são afetadas pela análise multiloco;

c) Avaliar a eficácia da geração de mapas integrados, a partir de mapas de diferentes populações. Incluir nestes estudos os conceitos de mapa ordenado, mapa alinhado e mapa integrado efetivo.

2. REVISÃO DE LITERATURA

2.1 Mapa Genético

Logo após as redescobertas dos trabalhos de Mendel, no final do século XlX, foram realizadas inúmeras pesquisas visando explicar a herança de vários caracteres nas mais diversas espécies de plantas e animais. Em 1902, W. Bateson, E.R. Saunders e R.C. Punnet demonstraram, em ervilha-doce, que as segregações dos caracteres cor da flor e formato do pólen não ocorriam de forma independente (Lander e Weinberg, 2000). T.H.Morgan e colaboradores trabalhando com *Drosophila melanogaster* ao analisarem o padrão de herança de um gene mutante ligado ao sexo, para cor dos olhos observaram que alguns caracteres não segregavam de acordo com a segunda lei de Mendel. Eles forneceram a primeira evidência de que os genes estão localizados em posições definidas nos cromossomos e que podem ser manipulados e avaliados experimentalmente. Morgan e colaboradores sugeriram que alguns genes estariam situados no mesmo cromossomo e que durante a meiose, ocasionalmente, ocorreriam, entre os homólogos, trocas de seguimentos denominados *crossing-over* ou permuta.

Em 1913, A.H.Sturtevant, interpretando dados oriundos da segregação de genes ligados, sugeriu o uso da porcentagem de recombinantes como indicador quantitativo da distância linear entre dois genes na construção de mapas genéticos. Os mapas mostravam que a posição dos genes correspondiam à sua ordem linear nos cromossomos. Assim, o conceito de localização dos genes em uma ordem linear passou a ser incorporado à teoria cromossômica da herança (Griffiths et al. 1998).

Os primeiros mapas genéticos foram fundamentados em marcadores morfológicos e citológicos. Estes se originaram de estoques cromossômicos com aberrações (como aneuploidias, translocações, deleções e inversões), principalmente nas culturas de milho, tomate, e ervilha (Coe et al, 1988; Rick e Yoder, 1988). Os estudos realizados com marcadores morfológicos em muito contribuíram para a elucidação do fundamento teórico da análise de ligação gênica e para construção das primeiras versões de mapas genéticos (Knapp, 1991). Atualmente, têm sido úteis nos casos em que os

grupos de ligação já estão associados a determinados cromossomos, podendo ser também utilizados para ancorar grupos de ligação construídos com base em marcadores moleculares a cromossomos específicos (Carneiro e Vieira, 2003)

Mapas genéticos se tornaram rapidamente uma poderosa ferramenta para os geneticistas uma vez que são de suma importância para o programa de melhoramento. O estudo de genomas inteiros, com a utilização de marcadores genéticos na análise de caracteres quantitativos, teve início a partir do século XX, quando foram estabelecidos os primeiros mapas genéticos intra-específcos em Drosophila melanogaster e em *Phaseolus vulgaris* (SAX, 1923).

Estoques de aneuplóidas foram utilizados como ferramentas para mapeamento de locos por muitas décadas (Lesley, 1932). Uma das maneiras de se obter plantas aneuplóides é por meio de radiação que induz mutações e podem gerar deficiências cromossômicas. Desta forma, populações desenvolvidas a partir de cruzamentos entre plantas irradiadas e cultivares mutantes foram excessivamente usadas para gerar mapas cromossômicos e conseqüente localização das mutações nos braços cromossômicos. Em 1990, com aumento do interesse em citogenética e o início do desenvolvimento de mapa de recombinação física, foi usado para quantificar a distribuição de crossing-over ao longo de cada cromossomo (Sherman e Stack 1995). Na mesma década, uma grande variedade de mapas de hibridização *in situ* também foram desenvolvidos, assim marcas genéticas e suas respectivas localizações físicas nos cromossomos foram usadas pelos citogeneticistas para construção de mapas genéticos (Funchs et al. 1996; Zhong et al. 1996). Estes são chamados de mapas citogenéticos, devido a sua conexão entre a ligação genética e a posição citológica (Harper e Candle. 2000). Muitas vezes os mapas citogenéticos são chamados de mapas genéticos na literatura e vale ressaltar que a ordem das marcas não é sempre conservada entre a posição genética e a citológica (Peterson et al. 1999).

Dentre as principais dificuldades encontradas para a confecção dos primeiros mapas genéticos, pode-se destacar a limitação em se obter marcadores genéticos consistentes e adequados para a análise de ligação. Embora os marcadores morfológicos tenham possibilitado o desenvolvimento de mapas genéticos no início do século passado, tais marcadores não apresentam qualidades esperadas para este tipo de abordagem, apresentando baixo nível de polimorfismo, pouca estabilidade ambiental e

número limitado de locos disponíveis para estudos de mapeamento (Ferreira e Grattapaglia, 1995).

O recente crescimento da genética molecular, por meio do grande número de marcadores gerados, tem causado grande interesse pelo mapeamento genético clássico. Como resultado, a análise de ligação e o mapeamento têm sido amplamente realizados por meio de processos computacionais.

As isoenzimas, denominadas marcadores bioquímicos, começaram a ser utilizadas como marcadores genéticos a partir da década de 60 (Lewontin e Hubby, 1966), permitindo a construção de mapas genéticos em várias espécies de plantas. A propriedade mais expressiva das isoenzimas como marcadores genéticos é a herança mendeliana simples com co-dominância entre alelos na maioria dos locos.

As isoenzimas apresentam limitações como: a reduzida cobertura dos genomas que são analisados em função do pequeno número de locos que podem ser identificados e o baixo nível de polimorfismo identificado por loco. Considerando que o número total de locos isoenzimáticos identificados seja superior a 100, são raros os trabalhos que utilizam mais de 30 locos (Murphy et al., 1990). Esse número é muito baixo para fins de mapeamento, principalmente quando se pretende obter ampla cobertura do genoma (Rick e Yoder, 1988; Ferreira e Grattapaglia, 1995).

A partir da década de 80, marcadores genéticos baseados na análise direta de seqüências de DNA polimórficas foram implementadas, assim o mapeamento tornou-se ilimitado a todas as espécies. Diversos tipos de marcadores polimorficos de DNA foram desenvolvidos como RAPD[1], AFLP[2], RFLP[3], SSR[4] e SNP[5]. A utilização destes marcadores na construção de mapas genéticos difere em alguns aspectos como: número de locos que pode ser detectado, graus de polimorfismo entre e dentro de acessos e características de dominância (Maliepaard et al., 1997). Aspectos como custos, tempo necessário para a realização das avaliações e dificuldades práticas inerentes à execução de cada técnica variam de acordo com a técnica empregada

Fragmentos mapeados de grupos de ligação podem também servir como sondas para trabalhos com hibridização *in situ* em metáfases mitóticas, contribuindo para a

[1] Random Amplified Polymorphic DNA
[2] Amplified Fragment Length Polymorphic
[3] Restriction Fragment Length Polymorphicv
[4] Simple Sequence Repeat
[5] Single Nucleotide Polymorphisms

identificação grupos de ligação e cromossomo. É possível ainda associar marcas mapeadas a genes, simplesmente pelo sequenciamento dos fragmentos mapeados, os quais servem como ponto de partida para a comparação com seqüências depositadas nos bancos, facilitando a identificação e clonagem destes genes. Bancos de dados genômicos foram construídos a partir de sequenciamento de nucleotídeos. Atualmente inúmeros projetos genoma estão em andamento em todo o mundo gerando um enorme volume de seqüências de DNA, as quais estão depositadas em bancos de dados públicos como o National Center for Biotechnology Information (NCBI) (http://www.nci.nlm.nih.gov/) ou privados. Um grande número de mapas genéticos prontos esta disponível na internet no SOL Genomics Network (SGN) http://sgn.cornell.edu e http://www.nci.nlm.nih.gov/mapview ou em sites especializados em determinadas culturas como tomate em http://tomatomap.net (Van Deynze et al. 2006). Estes sites, além de apresentarem os mapas prontos, apresentam uma série de informações úteis como o número e freqüência dos mapas mais utilizados para determinadas culturas, cada mapa mostra todos os grupos de ligação e seus relativos tamanhos, estatísticas sobre o mapa, nome da sonda e acesso no GenBank, dentre outras informações.

A construção de um mapa genético não é uma tarefa simples. Várias premissas devem ser consideradas o que torna o trabalho laborioso. Em geral, o primeiro passo na construção de um mapa de ligação esta relacionacionado com a escolha dos genitores a serem cruzados, de forma que maximize o polimorfismo genético (Gratapaglia & Sederoff, 1994, Verhaegen et al., 1997). Uma vez selecionados os genitores, é necessário o desenvolvimento de uma progênie segregante, tais como F_2, RIL, retrocruzamento ou duplo haplóide. A etapa seguinte envolve a obtenção de marcas contrastantes entre os genitores que apresentem segregação mendeliana na população de mapeamento. A estratégia de busca pelas marcas polimórficas depende, principalmente, do tipo de marcadores utilizados e da diversidade genética da espécie estudada. Após a escolha dos marcadores polimórficos é necessária à análise do padrão de amplificação dos indivíduos do restante da população de mapeamento e a obtenção das estimativas de recombinação. O passo seguinte é a análise de segregação da ligação para pares de locos. Assim, inicialmente é necessário que todos os marcadores sejam analisados individualmente par a par, verificando se existe ligação entre eles. O teste apropriado para esta situação é o teste χ^2.

Entretanto, este teste é apenas qualitativo. Uma vez confirmada a existência de ligação entre duas marcas, é necessário adotar métodos quantitativos para estudar o grau de associação entre estas marcas. Para tanto, faz se uso da metodologia da Máxima Verossimilhança (Liu, 1998), que permite a obtenção de estimadores consistentes, de medida de freqüência de recombinação, denotada por r. O problema da não aditividade da unidade "r" é contornado com a utilização de funções de mapeamento, que semelhante a uma transformação de dados, resulta em uma alteração de escala. As funções de mapeamento mais comumente utilizadas são as de Haldane (1919) e de Kosambi (1944). A teoria matemática da função de mapeamento e como elas são relacionadas com a interferência é tratado com detalhe por Owen (1950) e Bailey (1961), e mais recentemente por Schuster e Cruz, (2004).

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

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

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

Dada a grande quantidade de informação necessária pra construção de mapas ligação, bem como a quantidade de informação contida nos bancos de dados necessárias para construção de mapas de ligação, torna-se imprescindível o desenvolvimento de ferramentas ou metodologias de análise adequadas. Tais ferramentas são desenvolvidas com base em modelos genéticos e fundamentadas a partir de pressuposições

matemático-estatísticas e computacionais. Assim, várias espécies em que o estudo de herança são restritos tiveram mapas genéticos desenvolvidos rapidamente como o maracujá (Carneiro et al. 2002) e Kiwi (Testolin et al. 2001).

Dentre as principais utilizações do mapa genético destacam-se a localização e o mapeamento de QTL (Quantitative Traits Loci), localização de genes ligados a doenças e estresse abiótico, o mapeamento comparativo de genomas de diferentes espécies, estudos de sintenia e clonagem genes.

A comparação das estruturas genômicas de diferentes espécies (estudos de sintenia), do ponto de vista de homologias dos genes, conservação das distâncias e ordem de ligação nos cromossomos, contribui para o entendimento sobre a evolução dos genomas (Tansksley et al, 1988; Kianian e Quiros, 1992; Ahn e Tanksley, 1993). O mapeamento comparativo é também uma estratégia para obtenção de mapa único de referência para a maioria das espécies vegetais cultivadas (Moore et al, 1995; Sewell et al, 1999), pelo menos ao nível de famílias taxonômicas.

Em resumo a utilização de mapas genéticos para estudo do genoma de plantas e suas inter-relações pode ser vista na Figura 1.

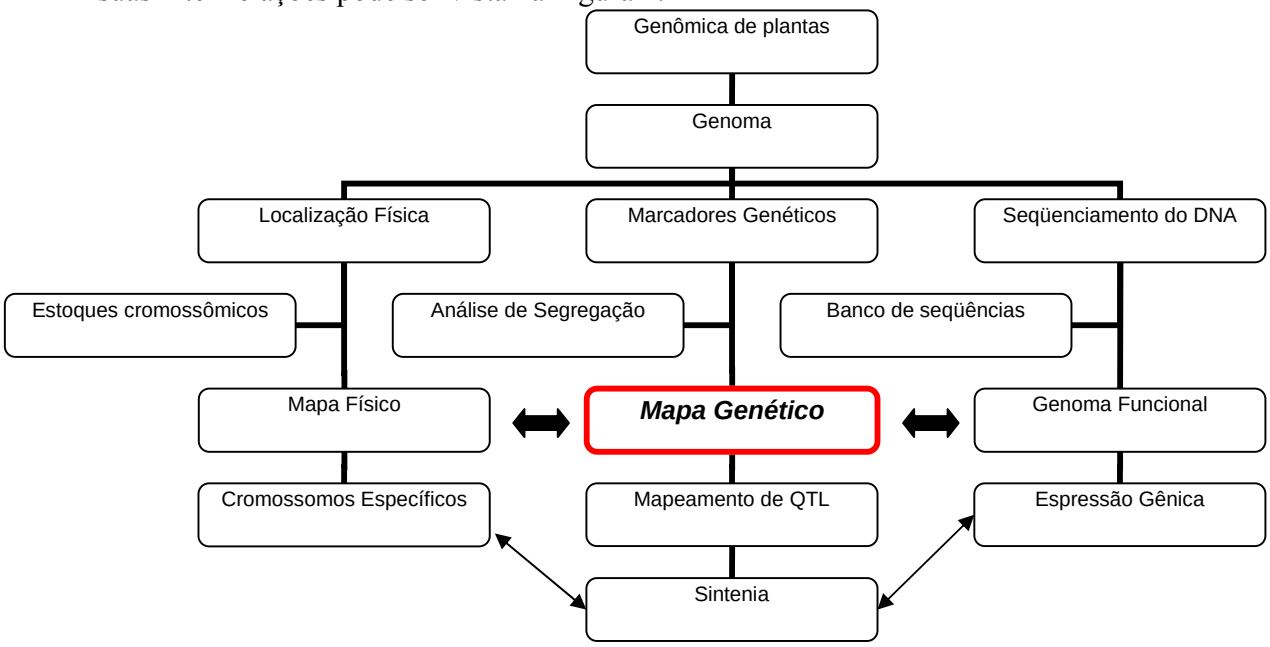


Figura 1 – Inter-relações entre as diferentes abordagens entre mapas genéticos e estudo de genomas de plantas. Fonte: Carneiro e Vieira, 2003

O ramo da genética que se ocupa com o desenvolvimento e a utilização de ferramentas estatísticas de análise e o processamento computacional dos bancos de dados genômicos é chamado de bioinformática. Mais especificamente, o ramo que se ocupa com a sondagem e a validação de conceitos estatísticos na análise genômica é conhecida como estatística genômica (LIU, 1998).

Em paralelo aos avanços na tecnologia de marcadores moleculares, o desenvolvimento de softwares analíticos – para a análise combinada de dados genotípicos, provenientes dos marcadores, e de dados fenotípicos – e a proliferação de metodologias genético-estatísticas possibilitaram o desenvolvimento de mapas genéticos voltados para o mapeamento de locos controladores de características de interesse.

O conceito base para o estabelecimento de mapas genéticos encontra-se na teoria de ligação ou não independência entre locos que se encontram proximamente ligados em um cromossomo (LIU, 1998). Estabelecer grupos de ligação com apenas dois locos tornou-se, a princípio, uma tarefa de relativa facilidade. No entanto, para o estabelecimento de mapas genéticos mais complexos, os procedimentos são computacionalmente dispendiosos e exigem um aparato estatístico mais avançado. Programas como "Mapmaker" (Lander et al,1987), "Linkage" 1 (Suitter et al, 1983), "Gmendel" (Liu e Knapp, 1990) foram desenvolvidos para ajudar na construção de mapas genéticos e estão disponíveis na internet em (http://linkage.rockefeller.edu/) e o programa GQMOL desenvolvido na Universidade Federal de Viçosa disponível em (http://ww.ufv.br/dbg/gqmol/gqmol.html)

2.2- Integração de mapas genéticos

O grande conteúdo de informações de ligação que estão atualmente disponíveis, em vários organismos com diferentes populações e marcadores, tem criado a necessidade de integração dos mapas que foram obtidos independentemente. A integração de mapas poderá ajudar e dar suporte aos esforços atuais de internacionalização de "projetos genoma" visando mapas genéticos e físicos de genomas completos

O fator chave na integração de mapas distintos é que estes possuam marcadores comuns, de forma a serem "âncoras" que possam interligá-los. Apenas quando um

número mínimo de âncoras é disponível, mapas distintos podem ser representados juntos. Porém, ainda que marcadores comuns estejam disponíveis, alguns problemas aparecem. Não apenas devido à precisão das estimativas da freqüência de recombinação que varia muito entre o conjunto de dados, mas também do tipo de informação utilizada. Por exemplo, um mapa poderia ser baseado em dados de populações F_2, com determinado tamanho e tipo de marcador, enquanto outro mapa pode ser baseado em dados de uma população de retrocruzamento. Outro mapa poderia ainda ser obtido por uma compilação "manual" dos dados disponíveis na literatura. O maior problema é definir os pesos destes tipos de informações que poderiam resultar em um mapa único, mais verossímil, contendo o mínimo de tensão interna, ou geralmente, satisfazendo critérios pré-estabelecidos de avaliação. Para isto, é necessário a utilização de ferramentas estatísticas.

A construção de um mapa de ligação é, essencialmente, a procura de um arranjo linear de marcadores e seus valores de recombinação. Várias estimativas da freqüência de recombinação entre um dado par de marcadores são obtidas, então, após uma criteriosa avaliação, deve ser transformada em apenas uma. Após, apontar os pesos dos dados disponíveis, uma pesquisa numérica para o melhor arranjo linear das marcas é formada. Se uma informação adicional não contém determinado par de estimativa, é possível que seja usada para ordenamento de marcas subseqüentes.

Um volumoso número de estimativas de freqüência de recombinação é obtido em diferentes experimentos, que implica na presunção de que verdadeiras freqüências de recombinações são iguais nestes experimentos. A distância de mapa entre dois marcadores é definida como um número significativo de eventos de recombinação, envolvendo determinada cromátide e as meioses daquela região. Esta é geralmente expressa em centimorgans (cM). A relação entre distância de mapa e freqüência de recombinação é expressa pela função de mapeamento genético. Diferentes funções de mapeamento podem assumir diferentes graus de interferência assumidos na permuta entre regiões adjacentes (Stam, 1993). Interferência é o fenômeno de eventos de recombinação em regiões adjacentes do cromossomo que não são independentes; em um teste clássico de três pontos, resulta num número de duplos recombinantes que diferem do que é esperado baseado na independência. No entanto, deve-se considerar que as várias funções de mapeamento não pressupõem a porcentagem de recombinação conjunta, ou seja, não são probabilidades de multilocos, para mais de três locos

(Goldstein et al., 1995). Modelos de análise de distância multilocos diferem dos modelos de dois locos porque toda a informação dos intervalos entre pares de locos de um grupo de ligação é considerada (Schuster e Cruz, 2004).

A integração de mapas tem sido extensivamente utilizada nas pesquisas científicas, como pode ser verificado pela quantidade de artigos publicados na última década com esta metodologia envolvendo diferentes espécies, tanto vegetais (Marcel et al., 2006; Lombard e Delourme, 2001; Ulloa et al., 2002 ; Qi et al., 2004; Somers et al., 2004; Saha et al., 2005; Lespinasse et al., 2000; Paran et al., 2004; Peleman et al., 2000; Sewell et al., 1999; Freyre et al., 1998; Klein et al., 2000; Chang et al., 2001; Song et al., 2004; Yan et al., 2005; Garcial et al., 2006; Doligez et al., 2006) como animais (Brown et al.,1998). Porém, a maioria dos relatos encontrados na literatura dizem respeito apenas à utilização de dois softwares que realizam a integração de mapas, sem uma explicação clara da metodologia utilizada por estes aplicativos. O software mais citado na literatura para a integração de mapas é o JoinMap (Stam, 1993).

Em espécies florestais, a grande maioria dos mapas genéticos é construído para genitores da população segregante, com pouca utilidade para os programas de melhoramento florestal. Isso se deve ao fato destes mapas serem construídos com marcadores dominantes RAPD e AFLP. Estes marcadores apresentam o problema de transferibilidade entre os genomas. Uma análise de 60 marcadores RAPD mapeados em 10 famílias de irmãos completos de *Eucalyptus urophylla* indicou que, em média, 64% dos marcadores RAPD poderiam ser transferíveis e informativos entre mapas de uma mesma população (Brondani et al., 1997). A transferência de marcadores informativos entre árvores dentro de uma mesma espécie cai rapidamente para menos de 15% se as árvores são oriundas de populações geograficamente distantes (Grattapaglia e Sederoff, 1994). Para contornar este problema são utilizados marcadores microssatélites. Cerca de 50 a 60% dos marcadores microssatélites são herdados em configurações totalmente informativas. A conservação dos locos de microssatélites entre espécies do mesmo gênero torna possível a obtenção de mapas "consenso", permitindo a comparação e integração de informação de mapeamento genético de marcadores, genes e QTLs entre experimentos independentes. Diversos trabalhos utilizando mapas consenso com microssatélites para espécies florestais podem ser encontrados na literatura: (Zhou et al., 2003; Yin et al., 2004; Achere et al., 2004; Brown et al., 2001; Sewel et al., 1999; Tani et al., 2003).

O processo de integração de mapas pode, também, ser requerido em análise de mapeamento avaliando famílias exogâmicas ou populações derivadas de pseudo-cruzamento teste para a obtenção das estimativas de freqüência de recombinação entre locos marcadores que não permitem tal obtenção de forma direta. Na análise deste tipo de progênie, diferentes tipos de cruzamentos podem ser observados. De forma resumida, podem-se classificar os seguintes tipos de acasalamentos quanto ao grau de informação da progênie (LYNCH & WALSH, 1998) (Tabela 1): (1) famílias derivadas de cruzamentos completamente informativos; (2) famílias derivadas de retrocruzamentos e (3) famílias derivadas de intercruzamentos.

Tabela 1. Tipos de famílias quanto ao grau de informação

Categoria	Cruzamento	Característica
Completamente Informativo	M_iM_j x M_kM_l	Alelos de cada genitor são distinguidos
Retrocruzamento	M_iM_j x M_kM_k	Alelos de apenas um genitor são distinguidos
Intercruzamento	M_iM_j x M_iM_j	Só descendentes homozigotos são informativos

Fonte: LYNCH & WALSH (1998)

Processo semelhante à integração de mapa será considerado para estimar valores de porcentagem de recombinação entre locos não informativos, a partir das informações das distâncias entre locos adjacentes às marcas, cuja distância não pode ser determinada por impossibilidade de detecção de classes recombinantes. Sendo assim para estes locos não informativos deve-se usar os marcadores âncoras que possibilitem que a estimação da freqüência de recombinação entre estes locos não informativos seja obtida de forma indireta, fazendo com que, permita a alocação de tais marcadores no mapa genético. Como exemplo, pode-se considerar o cruzamento $A_1A_2B_1B_2C_1C_1$ x $A_1A_1B_3B_4C_1C_2$. Neste caso a distância entre os locos A e C é indeterminada

.

		½ A_1C_1	½ A_1C_2
½	A_1C_1	r = ?	
½	A_2C_1		

Porém, é possível obter valores de distância entre A e B e entre B e C. Desta forma, a âncora B possibilitará a integração das informações e posterior estimação da distância entre A e C. Desta forma, um estudo detalhado de diferentes metodologias para a integração de mapas se torna necessário e de grande importância para futuras pesquisas científicas.

2.3 Metodologias para integração de mapas genéticos

A grande maioria dos artigos publicados, nos últimos anos, na área de integração de mapas genéticos é baseada na metodologia desenvolvida por (Stam, 1993) utilizando o software JoinMap. São utilizados outros, como o descrito por Peleman et al., (2000) e Paran et al. (2004), com auxílio do programa Int-Map. Mais recentemente, Hu et al., (2004) apresentaram uma metodologia baseada em uma função de verossimilhança conjunta para auxiliar na construção de mapas genéticos integrados. Estudos de simulação mostram que a diferença desta técnica quando comparada com a metodologia proposta por Stam, (1993) é pequena quando marcas co-dominantes são usadas, mas a segunda se mostrou melhor, principalmente quando marcadores dominantes ou uma mistura de marcadores dominante e co-dominantes foram usados. Isso ocorre porque com a metodologia baseada na função de verossimilhança, encontra os pesos ótimos entre as diferentes classes de dados enquanto com o método proposto por Stam (1993) não prediz a informação de cruzamentos envolvendo marcadores dominantes. Muitas destas metodologias não estão claramente descritas na literatura, não apresentando as equações e modelos empregados, dificultando em muito a sua descrição detalhada. A seguir são apresentadas algumas destas metodologias.

Método proposto por Stam (1993)

Stam (1993) propôs uma metodologia, implementada em um programa computacional (JoinMap), para integrar mapas individuais resultantes de experimentos independentes que utilizam diferentes tipos de populações F_2, retrocruzamentos, RILs bem como diferentes tipos de marcadores. Este método tem sido amplamente utilizado por diferentes grupos de pesquisadores (Song et al. 2004, Doligez et al. 2006, Sewell et al. 1999, Yan et al. 2005).

O procedimento básico empregado pelo JoinMap é primeiramente calcular as freqüências de recombinações e os correspondentes valores de LOD (indicativo da probabilidade de ligação) para cada população. Neste caso, Stam (1993) adota a definição de LOD como sendo o logaritmo de base 10 ($\log_{10}$) da razão da probabilidade de dois locos estarem ligados, com uma considerada taxa de recombinação, sobre a probabilidade da ausência de ligação. Posteriormente, as estimativas de populações distintas e as estimativas independentes (se disponível) são combinadas e os valores de LOD correspondentes são registrados. Na metodologia de Stam (1993), ambos os tipos de estimativas (dados de população ou independentes) são tratadas como se tivessem sido obtidas de amostras binomiais. O tamanho hipotético da amostra binomial é calculado para produzir o mesmo valor de LOD (dados da população) ou o mesmo erro padrão (dados independentes). Este tamanho da amostra hipotética é então utilizado como peso para calcular as estimativas conjuntas e os valores de LOD. Assim, os pesos fixados para as distintas estimativas de uma correspondente freqüência de recombinação, reflete a quantidade de informação que está incluída nesta estimativa.

O objetivo principal do método é a obtenção de estimativas das distâncias para encontrar a ordem das marcas. Uma versão modificada do procedimento primeiramente descrito por Jensen e Jorgensen (1975) e depois por Lalouel (1977) e Weeks and Lange (1987) foi usada, essencialmente é o método dos mínimos quadrados. Um exemplo ilustra a idéia. Serão ordenados 4 marcas A, B, C, D, com distâncias entre intervalos iguais a x, y e z, respectivamente. Supondo que as estimativas da freqüência de recombinação r_{AB}, r_{CD}, r_{AC}, r_{BD} e r_{AD} são calculadas. Estas estimativas são transformadas em distâncias no mapa por função de mapeamento. As distância x, z, x+y, y+z, y+z+x são chamadas como d_{AB}, d_{CD}, d_{AC}, d_{BD} e d_{AD}. A distância de y só pode ser estimada indiretamente, esta situação é típica para combinação de dados de diversas fontes. Uma

medida da discrepância entre as distâncias obervadas e seus respectivos valores esperados e o quadrado da diferença.

$(x-d_{AB})^2$, $(z-d_{CD})^2$, $(x+y-d_{AC})^2$, $(y+z-d_{BD})^2$ e $(x+y+z-d_{AD})^2$

Como os valores de d não são exatos, são adotados pesos apropriados, os valores de LOD foram usados como peso. Chamando estes pesos de w_{AB}, $w_{CD,}$ etc. Assim o total de D é dado por:

$$D= w_{AB}(x-d_{AB})^2 + w_{CD}(z-d_{CD})^2 + w_{AC}(x+y-d_{AC})^2 + w_{BD}(y+z-d_{BD})^2 + w_{AD}(x+y+z-d_{AD})^2$$

Obtendo as derivadas em relação a x, y, z, tem-se:

$\delta D/\delta x = o$, $\delta D/\delta y=0$ e $\delta D/\delta z=0$

estas equações podem ser resolvidas pelos procedimentos normais, em resumo Stam (1993) calcula a fração de recombinação conjunta, r, entre T cruzamentos como a soma de pesos das estimativas individuais:

$$\hat{r} = \sum_{i=1}^{T} N_i \hat{r}_i \qquad (1)$$

onde N_i é o peso aplicado para a i-ésima estimativa de recombinação r_i.

Os pesos para cada r_i confiam ou no valor de LOD associado (derivado dos dados da população) ou no erro padrão (derivado dos dados independentes), e são obtidos para encontrar o tamanho da amostra binomial hipotética que produz o mesmo LOD ou erro padrão. Assim, por exemplo, para o caso de LOD com tamanho N, a seguinte expressão é utilizada:

$$\mathrm{LOD} = \ln\left(\frac{r^{rN}(1-r)^{(1-r)N}}{0{,}5^N}\right) \qquad (2)$$

A solução para N é dada por:

$$N = \frac{LOD}{\left(r\ln(r) - r\ln(1-r) + \ln(2-2r)\right)} \qquad (3)$$

Este N é usado como peso relativo para r na equação (1), tornando possível assim, calcular as estimativas conjuntas e os valores de LOD. Na metodologia de Stam (1993), a probabilidade binomial é utilizada independentemente do tipo de cruzamento (retrocruzamento, F_2) e marcador (dominante, co-dominante) utilizado.

A seguir é apresentado um esquema detalhado da metodologia de Stam (1993) para integração de mapas genéticos.

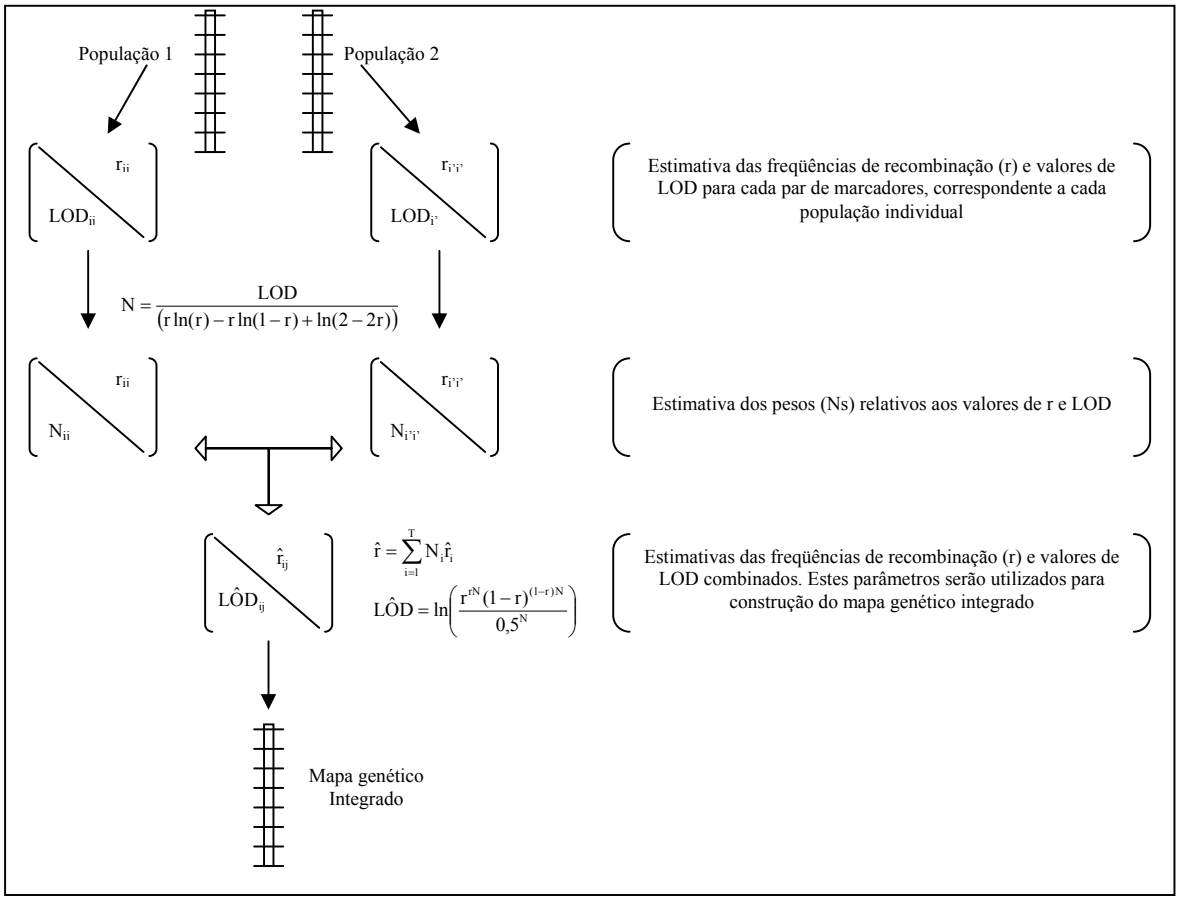


Figura 2. Esquema resumido da metodologia desenvolvida por Stam (1993) para integração de mapas genéticos.

Pela metodologia de Stam (1993), as estimativas da freqüência de recombinação dos dados iniciais são obtidas pela máxima verossimilhança.

Uma vez obtida as estimativas conjuntas dos pares de freqüência de recombinação (r) e dos valores de LOD, os mesmos são utilizados para construção do novo mapa genético integrado (Figura 2). Para o estabelecimento dos grupos de ligação Stam (1993) utiliza os valores de LOD combinado (conjunto) entre os pares de estimativas. Inicialmente um valor de LOD é fornecido, abaixo do qual a ligação não é considerada significante. Em qualquer estágio neste procedimento existe um conjunto de marcadores que tem sido fixado a um determinado grupo de ligação e outro conjunto de marcadores "livres", os quais ainda não estão fixados a nenhum grupo de ligação. Em cada etapa, a seguinte decisão é tomada: se nenhum dos marcadores livres é significativamente ligado, pelo valor de LOD, a um dos grupos existentes, um novo grupo de ligação é criado. Caso contrário, o primeiro marcador livre que se mostra ligado a um grupo existente é adicionado ao grupo. Um marcador livre é ligado a um grupo existente se ele estiver ligado, pelo valor de LOD, a pelo menos um marcador daquele grupo. Desta maneira verifica-se que o agrupamento dos marcadores depende do valor inicial de LOD crítico estabelecido. Valores maiores podem resultar em menores e mais grupos de ligação.

Método proposto por Peleman et al. (2000)

Esta metodologia foi desenvolvida primeiramente para integração de mapas genéticos do milho (Peleman et al. 2000). Posteriormente foi utilizada com sucesso para integração de seis mapas genéticos para pimenta (Paran et al. 2004).

Um grupo Holandês, denominado KeyGene, que presta serviços de pesquisa na área da genética e genômica, desenvolveram um software (Int-Map) para integração de mapas genéticos. O Int-Map (Peleman et al. 2000) integra mapas genéticos separados em um único mapa consenso, através de arquivos de dados que contenham as estimativas da freqüência de recombinação e a ordem dos marcadores nos mapas individuais.

A integração de mapas genéticos, pela metodologia descrita por Peleman et al. (2000) é realizada em três etapas. Primeiramente, os marcadores comuns entre pelo menos dois mapas são definidos como marcadores âncoras, e são usados para ligar correspondentes grupos de ligação nos mapas individuais. Posteriormente, a ordem consenso dos marcadores âncoras são calculadas das suas posições relativas em cada mapa individual. A seqüência dos marcadores no mapa consenso é iniciada com os marcadores mais comuns, adicionando-se seguidamente os mais próximos e mais comuns ao mesmo tempo. A posição consenso é calculada como o peso médio das distâncias do mapa individual entre novos marcadores e aqueles previamente posicionados. Antes da integração do mapa, devem ser fixadas duas variáveis importantes: a fração mínima de sobreposição entre os grupos (exemplo: pelo menos três dos 10 marcadores âncoras devem estar presentes no primeiro grupo de ligação) e a discordância máxima, em centimorgans (cM), entre dois marcadores âncoras adjacentes. Assim, os marcadores que desviam deste critério devem ser excluídos do mapa genético integrado. Por último, todos os marcadores únicos são posicionados no mapa integrado por meio de intra e extrapolação. A acurácia da posição estimada para um marcador é limitada ao intervalo entre os mais próximos marcadores âncoras dos respectivos mapas individuais.

Paran et al. (2004) utilizando esta metodologia, conseguiram integrar seis mapas individuais em um único mapa genético integrado para a cultura da pimenta (Capsicum spp.). O mapa integrado, contém 2262 marcadores, cobrindo aproximadamente 1832cM do genoma da pimenta, aumentando em aproximadamente 38% a densidade de marcadores (um marcador a cada 0,8cM, comparado com um a cada 2,1cM no mapa individual mais denso). Para a construção do mapa integrado, Paran et al. (2004) utilizaram 320 marcadores âncoras (marcadores comuns em pelo menos dois mapas individuais).

Esta metodologia é bastante simplista e pode acarretar em erros nas estimativas da distância entre os marcadores no mapa integrado.

Método proposto por Hu et al. (2004)

Ao contrário do método proposto por Stam (1993) que estima a informação da recombinação de determinado cruzamento pelo valor de LOD e então combina as

estimativas entre os cruzamentos assumindo uma amostragem de distribuição binomial, Hu et al. (2004) desenvolveram uma função de verossimilhança comum que combina informações entre todos os cruzamentos, para obter uma estimativa comum de recombinação.

Para melhor entendimento do método proposto por Hu et al. (2004), considere dois marcadores co-dominantes, A e B, ligados, cujos pais de um determinado cruzamento seja A_iB_k/A_jB_l e $A_{i'}B_{k'}/A_{j'}B_{l'}$ (i,i', j,j' = 1, 2, ...N_A; k,k', l,l' = 1, 2, ...N_B), onde N_A e N_B são os número de alelos por locos A e B, respectivamente. Similarmente, denote a configuração da descendência observada do cruzamento por A_yB_n/A_zB_m (y,z = i,i', j,j'; n,m = k,k', l,l'). Para o i-ésimo cruzamento do total de T cruzamentos, a probabilidade dos dados é:

$$\pi_i = \binom{N_i}{x_{il}.....x_{ik}} \sum_{f=1}^{N_{AB}} \sum_{m=1}^{N_{AB}} P(f,m) \prod_{k=1}^{k} \left[P(k \mid f,m)\right]^{x_{ik}}$$

(4)

Onde: P(f, m) é a probabilidade a priori do cruzamento parental (f x m); P(k|f, m) é a probabilidade esperada do fenótipo k (considerando o tipo de cruzamento parental de f x m, em que f e m referem-se aos genótipos da fêmea e macho, respectivamente); k é o número total de fenótipos que podem ser avaliados; x_{ik} é o número do fenótipo k-ésimo no i-ésimo cruzamento; N_{AB} (=N_AN_B (N_AN_B + 1)/2) é o número total do genótipo parental combinado para estes dois locus.

A equação anterior engloba todas as fases de ligação dos pais. Alternativamente, somente a fase mais provável pode ser usada. Entretanto, isto seria mais eficiente com tamanho maior de progênies, onde a fase incorreta pode ter probabilidade muito menor que a fase correta, e ter pequeno efeito na probabilidade total. Entretanto, quando o tamanho da progênie é pequeno, escolhendo a fase mais provável, pode-se introduzir um desvio (viés) devido aos efeitos do pequeno tamanho da amostra.

A probabilidade a priori do cruzamento parental P(f,m) pode ser as freqüências genotípicas da população, e incluem até mesmo a fase de ligação (desequilíbrio de ligação). Entretanto, usualmente as freqüências genotípicas não tem sido obtidas, e o nível natural de desequilíbrio de ligação são muito próximos de zero entre quaisquer marcadores proximamente ligados. Assim, utiliza-se uma distribuição uniforme a priori.

Além disso, com qualquer tamanho razoável de progênies (aproximadamente 20 ou mais), a probabilidade a priori carrega pouco peso comparado ao padrão de segregação das progênies.

Na equação, o número de todos os possíveis fenótipos (k) da descendência depende do número de alelos por loco e se os locos exibem dominância ou co-dominância. Com a suposição da distribuição uniforme a priori para a distribuição do genótipo parental em diferentes cruzamentos, a freqüência esperada de cruzamentos informativos também depende do tipo de marcador.

Atualmente as populações de mapeamento e melhoramento são as mesmas, e na maioria dos casos, os pais e os descendentes são genotipados, não sendo necessário a probabilidade de distribuição uniforme a priori. Desse modo, a equação anterior pode ser reduzida para:

$$\pi_i = \begin{pmatrix} N_i \\ x_{il}.....x_{ik} \end{pmatrix} \prod_{k=1}^{k} [P(k \mid f, m)]^{x_{ik}}$$

(5)

Em cada um dos dois casos acima, a verossimilhança comum de T cruzamentos é o produto de π_i $(i = 1,.....T)$:

$$L = \prod_{i=1}^{T} \pi_i \qquad (6)$$

que é maximizada com relação a um único r entre todos os T cruzamentos. Também é importante notar que nos casos em que os genótipos dos pais estão disponíveis e T=1, a equação acima é reduzida para os casos tradicionais envolvendo apenas um único cruzamento.

No método de Hu et al. (2004), as estimativas de máxima verossimilhança da taxa de recombinação pode ser obtida por meio do método de Newton-Raphson. Considerando a t-ésima estimativa da taxa de recombinação (r^t), a taxa de recombinação na etapa iterativa seguinte (r^{t+1}), pode ser calculada por:

$$\hat{r}^{t+1} = \hat{r}^t - \left(\frac{\partial^2 \ln L}{\partial r^2} \Big|_{r=\hat{r}^t} \right)^{-1} \frac{\partial \ln L}{\partial r} \Big|_{r=\hat{r}^t} \qquad (7)$$

O método de Hu et al. (2004), apresenta algumas vantagens sobre os demais, dentre elas a verdadeira fração de recombinação é constante entre todos os cruzamentos, ao contrário do método de Stam (1993) que combina linearmente as estimativas de recombinação entre cruzamentos em que as taxas de recombinação podem variar grandemente. O método de Hu et al. (2004), é adequado tanto para dados de múltiplos cruzamentos interespecíficos e seus variantes (retrocruzamentos e F_2), quanto para famílias de meio-irmãos (progênies de polinização livre, amostradas de populações naturais), não requerendo dados dos pais (eles podem ser homozigotos ou heterozigotos).

Este método estima a freqüência de recombinação de múltiplos experimentos, por meio da máxima verossimilhança, a qual é mais acurada e precisa, quando comparada com a metodologia de Stam (1993), principalmente quando marcadores dominantes são utilizados. Desse modo, com a utilização de marcadores dominantes, o tamanho da amostra pode ser significativamente reduzido utilizando o procedimento da máxima verossimilhança (Hu et al. 2004). Segundo Hu et al. (2004), o método da máxima verossimilhança supera o de Stam (1993) para marcadores dominantes e/ou populações F_2, devido a amostragem binomial ser implicitamente designada para marcadores co-dominantes e para quando eventos de recombinação são diretamente observados nos dados (no caso dos retrocruzamentos). Esta distribuição de amostragem para r não é apropriada para marcadores dominantes, nem para marcadores co-dominantes em uma população F_2 e nem para uma combinação de marcadores (dominantes e co-dominantes).

2.4 Aplicativos computacionais

Embora a metodologia descrita por Hu et al. (2004) apresenta, segundo os próprios autores, várias vantagens importantes sobre os demais métodos, atualmente, a metodologia de Stam (1993) é a mais utilizada para integração de mapas genéticos. Este fato deve-se principalmente ao sucesso do programa JoinMap (Figura 3), tanto para construção de mapas genéticos individuais como para integração de diferentes mapas. O software JoinMap (Van Ooijen e Voorrips, 2001) que emprega a metodologia de

integração descrita por Stam (1993), apresenta como vantagens a simplicidade, rapidez e eficiência na obtenção de mapas genéticos integrados.

Hu et al. (2004), por meio de simulações Monte Carlo (Wu et al., 2003), envolvendo vários cruzamentos independentes. Os resultados das simulações indicaram que estes dois métodos diferem na precisão das estimativas, apontando o método da função de verossimilhança conjunta apresenta superioridade na precisão da freqüência de recombinação, principalmente quando foram utilizados dados de marcadores dominantes.

A integração de mapas derivados de várias populações tem diversas vantagens, como o aumento da saturação do mapa e da cobertura do genoma, além de a ordem das marcas comuns tornam-se mais precisas. Mapas integrados também são usados em estudos comparativos entre espécies, (Beavis and Grant, 1991) e tem evideciado rearranjos cromossômicos entre grupos de ligação homólogos (Dirlewanger et al., 2004; Pelgas et al,. 2006; Nicolas et al., 2007). Mapas consenso integrados são também usados para detecção de QTL em múltiplas populações (Symonds et al., 2005; Blanc et al., 2006).

Recentemente, Van Os et al. (2006) construíram um mapa genético, não integrado, ultra denso para a cultura da batata. Este mapa contém 10.000 marcadores moleculares (AFLP, RFLP, SSR, CAPS e SCAR). Este tipo de mapa é de suma importância para o programa de melhoramento da batata, principalmente para detecção de QTLs, clonagem posicional de genes, para integração com mapas físicos e principalmente para alinhamento de mapas dentre a comunidade científica. Entretanto, Van Os et al. (2006) apontaram a falta de programas computacionais capazes de manipular grandes volumes de informações moleculares, como a principal dificuldade encontrada para construção de mapas genéticos altamente saturados de marcas.

Não existem dúvidas que os processos de integração ou construção de mapas genéticos com alta densidade de marcadores moleculares, são importantes para qualquer programa de melhoramento genético. Entretanto, há uma carência de métodos ou mesmo programas computacionais que atendam a demanda e o fluxo de informações geradas pelos laboratórios. Avanços na programação de softwares que são fundamentados podem ser de grande utilidade para integração de mapas genéticos, assim como vem sendo aplicada para aprimorar a detecção de QTL.

Na tabela 2 encontram-se alguns artigos de integração de mapas genéticos publicados nos últimos anos em diversas culturas, bem como o software utilizado, número de populações, comprimento do mapa, a distância em cM de locos adjancentes, o número de marcadores âncoras. Pela tabela pode observar que há não um padrão entre o tipo de marcador e população usada na integração de mapas, bem como o número mínimo de marcadores que seriam indicados pra fazer o processo de ancoragem com certo grau de confiabilidade nos resultados.

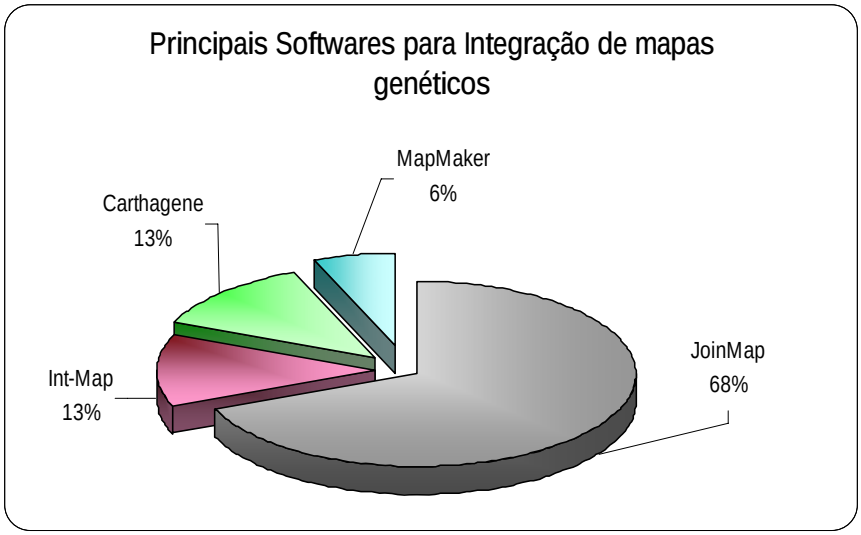


Figura 3. Principais softwares utilizados para a construção de mapas genéticos integrados.

Tabela 2. Principais programas, tamanho de população, número de marcadores, comprimento do mapa, distância média entre dois lócus adjacentes e número de marcadores âncoras utilizados para integração de mapas genéticos.

Cultura	Software	Nº populações	Nº de marcadores	Comprimento do mapa (cM)	Dist. média (cM) entre dois locus adjacentes	Nº de marcadores âncora*	Referência
Cevada	JoinMap	6 (3RIL;3DH)	3258	1081	0,33	502	Marcel et al. (2006)
Canola	Carthagene	3 (DH)	540	2429	4,50	253	Lombard e Delourme (2001)
Algodão	JoinMap	4 ($F_{2.3}$)	284	1502,6	5,30	-	Ulloa et al. (2002)
Milheto	JoinMap	4 (F2)	418	1123	2,60	116	Qi et al. (2004)
Trigo	JoinMap	4	1235	2569	2,20	-	Somers et al. (2004)
Azevém	JoinMap	2	922	1840,9	2,00	42	Saha et al. ((2005)
Seringueira	JoinMap	2	717	2144	3,00	-	Lespinasse et al. (2000)
Videira	Carthagene	5	515	1647	3,30	-	Doligez et al. (2006)
Cana-de-açúcar	JoinMap	Pseudo-testcross	357	2602,4	7,30	-	Garcia et al. (2006)
Pimenta	Int-Map	6	2262	1832	0,80	320	Paran et al. (2004)
Milho	Int-Map	23	5650	1707	0,30	2100	Peleman et al. (2000)
Citrus	JoinMap	2	217	527	2,40	61	Oliveira et al. (2005)
Rosa	JoinMap	2	542	977	0,55	141	Yan et al. (2005)
Soja	JoinMap	5	1849	2523,6	1,36	-	Song et al. (2004)
Pinus	JoinMap	4	357	1300	3,64	108	Sewell et al. (1999)
Feijão	MapMaker	3	563	1226	2,17	-	Freyre et al. (1998)
Média		5	1230	1546	2,60	404	

* marcadores comuns em pelo menos dois mapas individuais; RIL: Recombinante Inbreed Line; DH: Duplo haplóide.

2.5 Simulação

A busca de maior eficiência nos programas de melhoramento genético é constante, dada a pressão para se aumentar a produtividade e a estabilidade das espécies cultivadas. Há várias alternativas que podem ser trabalhadas, contudo, a maioria delas demanda a realização de experimentos em várias condições. Essa estratégia, além de difícil generalização, exige tempo e muito recurso. Com as facilidades computacionais atuais, a principal opção é o emprego de simulação computacional que, além de demandar menos recurso e tempo, pode ser generalizada com mais facilidade.

Para realizar a simulação computacional, alguns conceitos devem ser compreendidos. De acordo com Cruz (2001), é por meio da linguagem de programação que se pode estabelecer comunicação com o computador e fazer ser entendido por ele, de tal forma que os dados sejam convenientemente analisados. A linguagem é um conjunto de códigos, regras e vocabulários, que fará com que o computador entenda a instrução. Atualmente, dispõe-se de várias linguagens de programação orientadas a objetos como, o Delphi e Visual Basic.

Diversos aplicativos são utilizados para a simulação na área da genética e melhoramento de plantas utilizam diversas linguagens, como o Diallel (Burows & Coors, 1994), MENDEL (Euclydes, 1996), GENES (Cruz, 1998) e o GQMOL (UFV, 2008).

A simulação consiste em projetar um modelo de um sistema real e conduzir experimentos com este modelo, no intuito de compreender o comportamento ou de avaliar várias estratégias para a operação deste sistema. A simulação abrange dois grandes aspectos que são a modelagem e a experimentação e têm como dois grandes objetivos compreender o sistema existente e prescrever recomendações que possam ser generalizadas para quaisquer situações. Além de apresentar menor custo e maior rapidez, garante certeza da direção e sentido das respostas (Ferreira, 2001; Wagner, 2002).

A simulação computacional tem sido de grande utilidade em estudos genéticos sob vários contextos, como o de populações, do indivíduo ou do próprio genoma. Ela demanda dos geneticistas o desenvolvimento de modelos biológicos que retratem, da melhor maneira possível, os fenômenos de interesse, e dos programadores as rotinas

para o processamento adequado, apesar de impor restrições, para que a influência de certos fatores possa ser avaliada (Cruz, 2001).

Para que se possa simular o comportamento de fenômenos reais, deve-se utilizar modelos que representem um objeto, sistema ou idéia (Wagner, 2002). Portanto, a modelagem é um aspecto importante na simulação. O modelo deve ser suficientemente simples para ser operacionalizado e interpretado adequadamente, mas seu desempenho deve ser comparável ao modelo real e, se a defasagem for grande, ele deve ser eliminado ou refinado (Cruz, 2001).

Ao utilizar-se de uma técnica de simulação, o pesquisador deve precaver-se contra erros, seja estes devidos a problemas como levantamentos amostrais, escolha inadequada das distribuições de probabilidades nos eventos de natureza aleatória, simplificação inadequada da realidade e erros de implantação do sistema simulado. Para a garantia de sua eficiência pode-se lançar mão de processos de validação. Essa validação consiste em fazer o sistema simulado operar nas condições do sistema real e verificar através de testes de hipóteses e outras análises estatísticas ou através de comparação com situações reais já analisadas, se os resultados observados na simulação condizem com os observados no sistema real (Cruz, 2001).

Uma utilização desses processos de validação foi realizada por Mackay & Caligari (1999). Eles constataram que a presença de *outliers* pode usualmente ser detectada por meio de rotinas de validação de dados, mas alguns erros maiores escapam desta detecção porque caem numa região de aceitação. Em um programa de melhoramento, estes erros podem ser raros, quando ocorrem, podem reduzir a resposta à seleção, devido a uma quantia desproporcional em sua freqüência. Combinações de taxas de erro altas (1%) e baixas (0,1%) foram simuladas entre 1 e 10 indivíduos selecionados de populações de tamanho 100 ou 1000. Quatro diferentes tipos de erros foram simulados ajustando as médias e variâncias dos erros maiores. Os autores concluíram que os erros maiores causaram larga redução na resposta à seleção, especialmente quando presente em uma taxa de erro de 1% em uma população de tamanho 1.000.

De acordo com Ferreira (2001), no melhoramento de plantas, o uso da simulação é justificado quando: (a) as soluções analíticas não existem ou o grau de dificuldade e o número de variáveis envolvidas não permitirem a realização de inferências adequadas sobre o problema; (b) pretende-se comparar a eficiência de um novo procedimento ou

técnica em relação a outros já existentes e até mesmo consagrados; (c) os métodos que são rotineiramente empregados em algumas etapas de um programa de melhoramento propiciem qualquer melhoramento ou aumento de sua eficiência.

Os primeiros trabalhos de simulação, utilizando recursos computacionais relacionados à genética e melhoramento de plantas, foram apresentados por Fraser (1957a,b). Esses trabalhos avaliaram o efeito da ligação nas taxas de ganho com a seleção massal e nos avanços dos ganhos genéticos.

Darvasi et al. (1993) realizaram estudo de simulação, usando populações de retrocruzamento, para determinar o efeito do espaçamento de marcadores, efeito do gene e o tamanho da população no poder da ligação entre o marcador e o QTL, no erro-padrão dos estimadores de máxima verossimilhança e do efeito do QTL sobre a localização no mapa de ligação. Neste estudo, constatou-se que o poder de detecção do QTL foi virtualmente o mesmo, usando marcadores espaçados de 10 ou um número infinito de marcadores, com a ocorrência de pequeno decrécimo quando os marcadores foram espaçados de 20 ou até mesmo 50 cM. A vantagem de utilizar intervalo de mapeamento em vez de análise com apenas um marcador foi sem importância. O poder de resolução de uma ligação marcador-QTL foi definido como sendo de 95% do intervalo de confiança, para um mapa de localização de QTL que seria obtido através da utilização de infinitos marcadores. Descobriu-se que, reduzindo o espaçamento entre marcadores para menos do que o poder de resolução, não houve ganho apreciável estreitando o intervalo de confiança. É interessante observar que, nesse trabalho de simulação, foram geradas 1000 repetições para cada parâmetro de combinação.

Ferreira (1995) avaliou a eficiência do mapeamento de QTLs e da seleção assistida por marcadores. Para isso foi gerada uma população F_2 com 400 plantas. Considerou-se o caráter sendo controlado por genes distribuídos em todos os 10 cromossomos, mapeando-se os QTLs em todos eles e considerando uma distribuição aleatória. Os valores genotípicos e fenotípicos foram gerados considerando ausência e presença de dominância e ausência de epistasia e quatro níveis de herdabilidade (0,750; 0,500; 0,250 e 0,125). Foi constatado que, com o aumento do número de genes controlando a característica quantitativa, houve um aumento de probabilidade de se detectar significativamente marcadores ligados a QTLs. O tamanho populacional de 400 plantas F_2 foi adequado, pois foram obtidos, mesmo com baixos níveis de herdabilidade

(h2 = 0,125), marcadores ligados aos QTLs com ligação menor que 10% de recombinação.

Um estudo da regressão como medida de adaptabilidade e estabilidade para seleção de cultivares adaptadas a ambientes bons e ruins foi o objetivo da simulação proposta por Simmonds (1991). As simulações mostraram que a seleção sistemática para um ambiente ruim é necessária, propondo-se a exploração das interações genótipos x ambientes.

Frish et al. (1999), para verificar a confiabilidade do software utilizado, compararam o mapa de ligação original utilizado em seu trabalho, no qual foi baseado em dados experimentais de F_2, com um mapa de ligação construído a partir de dados simulado de indivíduos F_2, pelo programa MAPMAKER (Lander et al., 1987). Pela comparação foi visto que os mapas estavam em excelente concordância, confirmando que os modelos usados nos dois programas eram similares.

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

Martinez et al. (1999) desenvolveram procedimento de mapeamento baseado em modelo randômico utilizado para se investigar sua robustez e adequação para mapeamento de QTL em populações onde prevalecem estruturas de famílias de meios irmãos. Sob o modelo randômico, a localização do QTL e componentes de variância foram estimadas usando técnicas de máxima verossimilhança. A estimação de parâmetros é feita baseando-se na abordagem do modelo de pares de irmãos. A proporção de genes idênticos por descendência (IBD) no QTL foi estimada através de dois marcadores flanqueadores. Já as estimativas para os parâmetros e poder de QTL foram obtidas usando dados simulados, variando o número de famílias analisadas, a herdabilidade da característica, a variância, o número de marcadores e o número de alelos no QTL. Os fatores mais importantes que influenciaram o poder e os parâmetros do QTL foram a herdabilidade e a variância genética da característica. Segundo os autores, o número de alelos do QTL não influenciou as estimativas dos parâmetros avaliados, nem o poder de detecção do QTL. Com uma herdabilidade mais alta, foi observado um confundimento entre QTL e componentes poligênicos.

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

O que se verifica é que profissionais da área de informática têm pouco conhecimento dos problemas da área de genética, que apresentam certa complexidade por tratar de fenômenos biológicos e envolver princípios e, de certa forma, complexas distribuições probabilísticas na sua análise. Por outro lado, são raros os geneticistas com conhecimento e aptidão para atuarem na área da informática (Cruz, 2001). A integração entre essas áreas será fundamental para tornar a simulação computacional muito mais eficaz e utilizada entre melhoristas.

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

O programa GQMOL (Cruz, 2005) foi desenvolvimento com o propósito de analisar dados obtidos de estudos moleculares. O programa pode ser usado para análise de segregação de locos individuais, estimação de porcentagem de recombinação, agrupamento de marcas moleculares e mapeamento, incluindo estudos de QTL com populações controladas, populações exogâmicas, simulação, análise de imagem e agora mapa consenso. Além disso, conta com um módulo de ensino, apresentando vários procedimentos para entendimento de princípios estatísticos e genéticos envolvidos na análise genômica. Possui também o módulo “simulação”, onde diferentes genomas

podem ser simulados, levando em conta diferentes tamanhos amostrais, tipos de populações, variáveis quantitativas entre outras variáveis.

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

O programa GENES, amplamente utilizado em análises de modelos aplicados ao melhoramento de plantas e animais, é um software destinado à análise e processamento de dados por meio de diferentes modelos biométricos, contando com procedimentos uni e multivariados, enfatizando estimação de parâmetros genéticos. Também estão disponíveis procedimentos para análise de dados binários, geralmente obtidos de estudos moleculares, permitindo a análise e interpretação de fenômenos particulares desta área (Cruz, 2003). Possui ainda o modo "simulação", onde o usuário poderá avaliar tamanhos de amostras, número de famílias, plantas, repetições, em diferentes estudos.

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

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

3. MATERIAL E MÉTODOS

3.1 – Terminologias utilizadas

Para a definição e estudo do processo de integração de mapas, foi necessário identificar e designar os diferentes genomas e mapas gerados pelo processo de simulação e integração. Assim, algumas definições são indispensáveis para o melhor entendimento deste trabalho:

Genoma original – Genoma gerado com os marcadores, saturação e tamanho pré-estabelecidos. Este genoma é utilizado para se obter os genomas simulados.

Genoma simulado – Genoma simulado a partir do genoma original, cujas distâncias entre as marcas são estabelecidas em função do tipo de população e tamanho da população.

Marcador âncora – Marcador presente em pelo menos dois mapas genéticos individuais sujeito ao processo de integração.

Mapa consenso – Mapa construído pela junção ou integração de mapas que somente apresentam marcadores âncoras.

Mapa alinhado – Constituído pela junção de mapas que possuem pelo menos um marcador âncora por grupo de ligação. Os mapas são submetidos a apenas um alinhamento em que há preservação do posicionamento dos marcadores. Este mapa contêm todos os marcadores incluindo os âncoras repetidas vezes.

Mapa ordenado – Mapa integrado com o melhor ordenamento da marcas, os valores das distâncias entre marcadores não leva em consideração as particularidades de cada mapa analisado. Os marcadores âncoras são representados uma única vez.

Mapa Integrado Efetivo – Mapa integrado com todos os marcadores com a melhor ordem, cujas distâncias entre os marcares âncora e não âncoras foram estimadas por análise multiloco.

3.2 Simulação de dados

Para gerar os dados foi utilizado o módulo de simulação do aplicativo computacional GQMOL (CRUZ, 2008), que permite gerar informações sobre genomas, genótipos, genitores, indivíduos de diferentes tipos de populações e dados de características quantitativas. Foram simulados genomas parentais e amostras de populações F_2 e de retrocruzamento, construída com base em marcadores dominantes e co-dominantes. Para as análises dos mapas consenso as populações geradas foram de 100, 200 e 400 indivíduos, com 3 grupos de ligação cada. Para mapas integrados foram simulados mapas com apenas um grupo de ligação e populações com 100, 150, 200 e 400 indivíduos. Para cada população foram geradas matrizes de distâncias entre os pares de locos. Depois de feita a simulação e gerada a matriz de distância foram obtidos os mapas genéticos a partir dos diferentes tamanhos e tipos de populações. Os diferentes mapas gerados foram integrados de acordo com as metodologias propostas.

3.2.1 Simulação do genoma

Para a construção dos mapas consenso com marcadores âncoras, foi tomada como referência uma espécie diplóide fictícia com 2n= 2x = 6 cromossomos, cujo comprimento total do genoma, por grupo de ligação, foi estipulado em 50, 100 e 150 cM. Foi gerado um genoma com nível de saturação de 11 marcas moleculares eqüidistantes (com 5, 10 e 15 cM de intervalo em cada grupo de ligação), (figura 4). Cada genoma foi composto por 3 grupos de ligação: 50, 100, 150 com um comprimento total de 150, 300 e 450 cM, respectivamente. Neste estudo foram usadas simulações de marcas co-dominantes e dominantes para F_2 e somente co-dominantes para retrocruzamento.

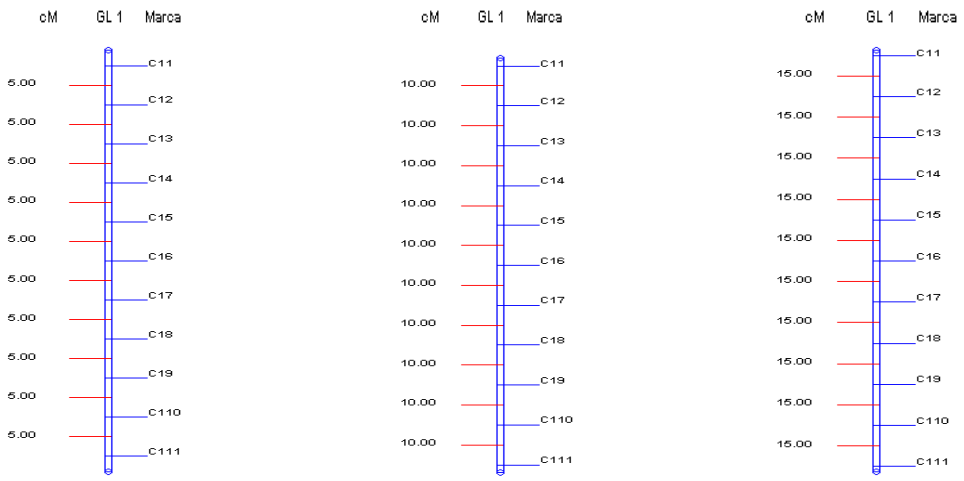


Figura 4 – Grupo de ligação 1 dos 3 genomas analisados com saturação de 11 marcas por grupo de ligação eqüidistando 5, 10 e 15 cM.

Para estudo do processo de integração foi tomado como referência uma espécie diplóide fictícia com 2n = 2x =2 cromossomos, cujo comprimento total do genoma, por grupo de ligação, foi estipulado em 100 cM. Foi gerado o genoma com nível de saturação de 21 marcas moleculares (ou 5 cM de intervalo) por grupo de ligação. O genoma foi composto por apenas um grupo de ligação, com comprimento total de 100 cM (Figura 5).

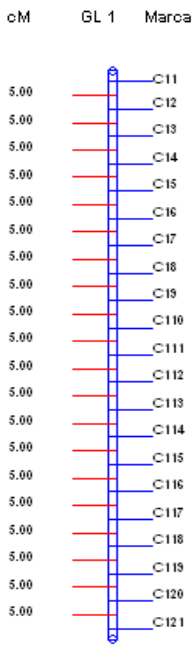


Figura 5 – Grupo de ligação 1 do genoma analisado com saturação de 21 eqüidistantes 5cM.

3.2.2 Simulação dos genitores

Para as populações utilizadas no estudo de mapas consenso e integrado, na análise de ligação gênica com dados de populações F_2 e de retrocruzamento, para cada grau de saturação do genoma estudado foi simulado a situação em que os pais são homozigotos, sendo um dominante e outro recessivo. Também foi admitido que todos os locos estivessem em aproximação. Na figura 6 é exemplificado o genótipo dos pais, P_1 e $P_{2,}$ em diferentes graus de saturação, referentes ao primeiro grupo de ligação (GL1).

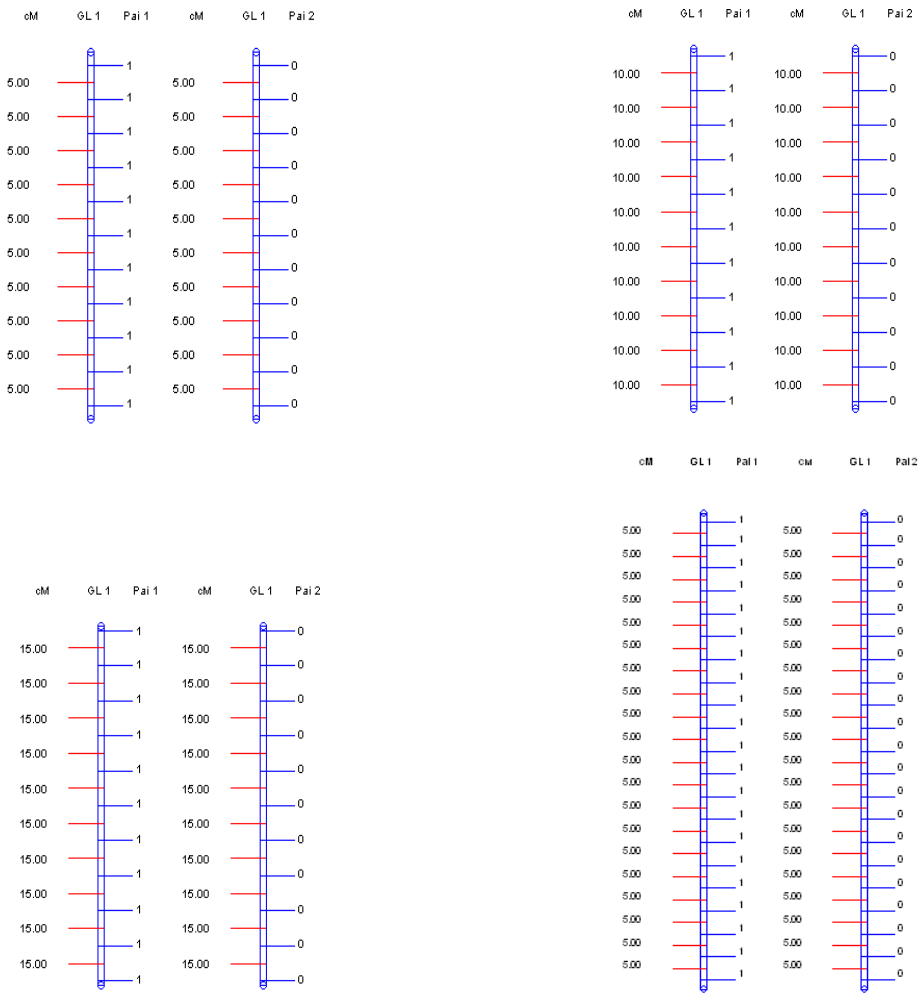


Figura 6 – Genoma dos genitores (pai 1 e pai 2), Correspondendo o grupo de ligação 1(GL1). Saturação dos genomas com 11 marcas eqüidistantes a 5cM(a), a 10cM(b) e a 15 cM(c) genomas utilizados para a obtenção de mapas consenso e saturação do genoma de 21 marcas equistantes a 5 cM para obtenção dos mapas integrados. Valor 1 identifica a presença da marca e 0 ausência.

3.2.3 Tamanho da população

Para obtenção dos mapas consenso tanto para população de retrocruzamento quanto para as F_2 foram geradas amostras de população com 100, 200 e 400 indivíduos, com 11 marcas por grupo de ligação eqüidistantes a 5, a 10 e a 15 cM e 10 repetições por grupo de ligação, Dessa forma um total de 270 simulações, foram realizadas sendo 90 para retrocruzamentos e 180 para população F_2.

Para os mapas integrados foram utilizadas populações F_2 co-dominante e retrocruzamento com tamanhos de 100, 150, 200 e 400 indivíduos, com 21 marcas por

grupo de ligação e marcadores eqüidistantes a 5 cM, em um total de 4 simulações para F_2 co-dominante e 4 para retrocruzamentos.

3.3 Procedimentos para simulação dos indivíduos das populações F_2 e de retrocruzamento

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

O processo de simulação das populações F_2 e de retrocruzamentos seguiram os seguintes passos:

1- A partir do genoma simulado, foram construídos os genótipos parentais homozigotos e contrastantes para os marcadores, de forma que a geração F_1 estivesse em aproximação para todos os marcadores.

2 – A partir do genótipo da geração F_1, foram gerados gametas para a formação dos indivíduos das populações F_2 e de retrocruzamento. A produção de gametas foi feita simulando-se o pareamento dos homólogos e realizando-se permutas ao longo dos cromossomos e considerando a não-existência de interferência nas regiões delimitadas por dois marcadores adjacentes. A probabilidade de ocorrência de recombinação numa região entre marcadores adjacentes foi dada de acordo com a distância destes marcadores no genoma simulado. 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.

3.4 Processo de mapeamento

3.4.1 Análise de segregação de locos individuais

Foram aplicados testes de qui-quadrado (χ^2) para verificar a razão de segregação de cada marca em todas as populações geradas, com suas particularidades.

A estatística qui-quadrado é dada por:

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

onde:

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

- Obs_i e Esp_i, são os valores observado e esperado, para a i-ésima classe fenotípica

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

A hipótese (H_0) de segregação específica para cada loco foi testada a 5% de probabilidade (erro tipo I). Se o valor de probabilidade calculado foi inferior ao pré-estabelecido, a hipótese H_0 era rejeitada, ou seja, havia a evidência de que a segregação não ocorreu de acordo com o esperado.

3.4.2 Estimação da porcentagem de recombinação

Após a aplicação dos testes de segregação, foi efetuada a etapa de estimação da porcentagem de recombinação entre pares de marcas, utilizando o método da máxima verossimilhança, conforme descrito por (Schuster e Cruz, 2004).

3.4.3 Determinação dos grupos de ligação

Na formação do grupo de ligação utilizou-se a propriedade transitiva, ou seja, se o loco A está ligado ao loco B, e o loco B está ligado ao loco C, logo o loco A está ligado ao loco C, independente da freqüência de recombinação estimada entre A e C e, portanto, A, B e C pertencem ao mesmo grupo de ligação. Os critérios utilizados no agrupamento foram a freqüência máxima de recombinação (r_{max}) e o LOD mínimo (LOD_{min}), para inferir se dois locos estão ligados. Foram utilizados os valores de 30% e 3, respectivamente, para r_{max} e LOD_{min}. As marcas mais próximas em relação às marcas já consideradas que atenderam aos dois critérios adotados, foram incorporadas ao grupo de ligação. Assim, o processo continua, investigando a porcentagem de recombinação e LOD entre marcas e possíveis vizinhos a serem incorporados às extremidades do grupo de ligação em construção.

3.4.4 Ordenamento das marcas no grupo de ligação

Após a formação dos grupos de ligação, foi determinada a melhor ordem das *n* marcas nos grupos.

Um dos processos é gerar todas as ordens possíveis ($n!/2$) e adotar como critério a identificação da melhor ordem como sendo aquela que proporciona menor soma de distâncias. Contudo, na medida em que o número de locos marcadores aumenta, aumenta-se grandemente o número de ordens possíveis, consumindo, portanto, um grande tempo de processamento. Então, foi utilizado, alternativamente, o método da soma das frações de recombinação adjacentes (SARF - *Sum of Adjacent Recombination Fractions*) para os mapas construídos sem análise multiloco. Neste processo , a melhor ordem é aquela que apresenta a menor soma das recombinações adjacentes. Considera-se a ordem original estabelecida pelo processo de agrupamento e aplica-se o algoritmo RCD (*Rapid Chain Delineation*), que consiste em realizar permutas entre dois marcadores vizinhos ou distantes envolvendo três ou quatro marcadores. A ordem é alterada se, após a permuta, a soma das distâncias adjacentes for reduzida. Após todas as permutas conclui-se que a melhor ordem é aquela de menor soma de distâncias adjacentes (Schuster e Cruz, 2004).

Para metade dos mapas simulados para a obtenção dos mapas consenso integrados efetivo utilizou-se da análise multiloco. Nesta análise considera-se que as informações baseadas em pares de locos, devem ser menos precisas que aquela que deveria considerar, simultaneamente, as informações de todas as classes disponíveis para todos locos do grupo de ligação. Desta forma novas distâncias serão calculadas pelo método gráfico da superfície de resposta ou método dos mínimos quadrados (Stam 1993).

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

3.4.5 – Mapa consenso

Para a construção dos mapas consenso foi utilizado o módulo mapa consenso do GQMOL, sendo que os mapas provenientes dos diferentes tamanhos e tipos de população simulados, foram usados para integração. Logo após integração dos mapas, eles foram comparados com o mapa original quanto ao tamanho do grupo de ligação, distância média dos marcadores adjacentes, variância média dos marcadores adjacentes, correlações de spearman e estresse. Na Figura 7 é apresentado um exemplo de mapa consenso.

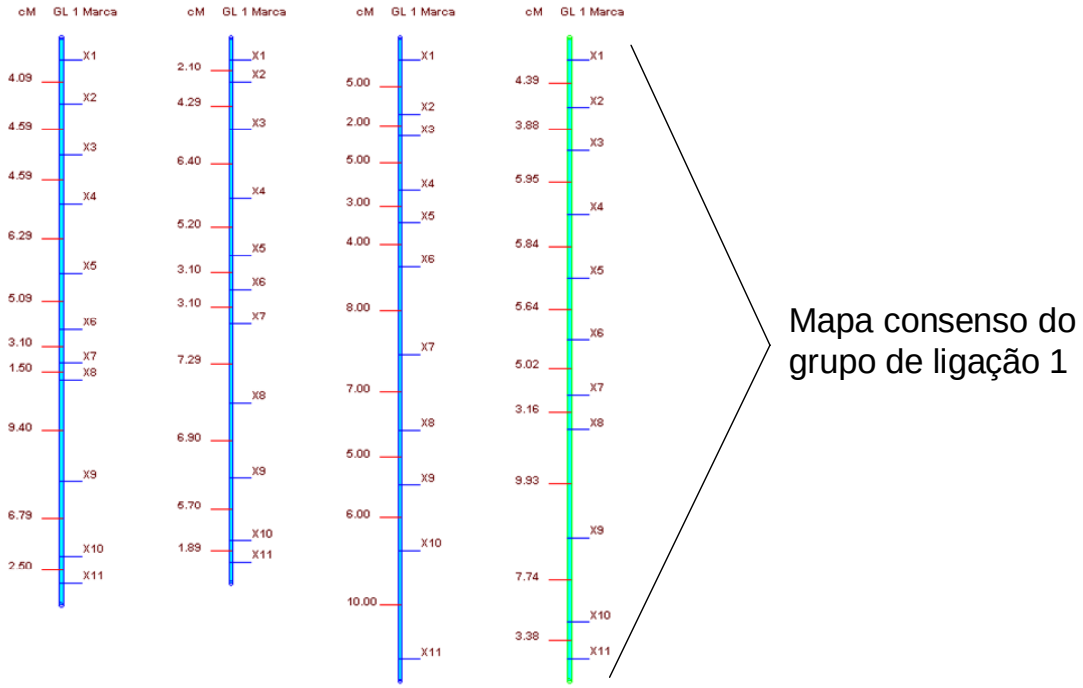
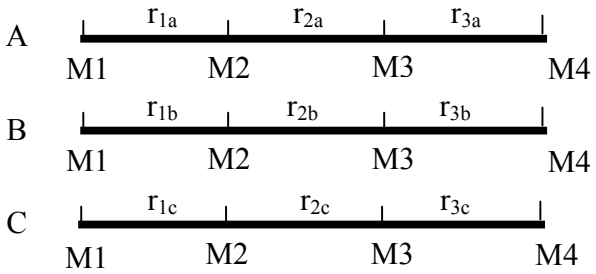


Figura 7 – Mapa consenso gerado pela integração de três diferentes mapas genéticos. São ilustrados onze ancoras com o mesmo ordenamento nos diferentes mapas

3.5 - Metodologia para obtenção do mapa consenso

Neste trabalho definiu-se mapa consenso como àquele resultante da análise conjunta de vários outros mapas genéticos que apresentam marcadores consenso, denominados de âncora.

Assim, será considerado como ilustração, a obtenção do mapa consenso a partir de três mapas designados por A, B e C, contendo quatro marcadores consenso M_1, M_2, M_3, M_4 como apresentado a seguir:

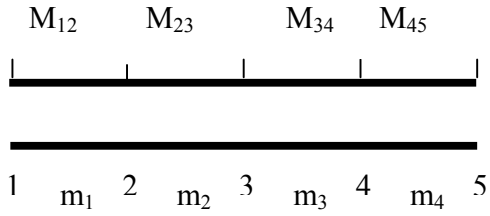


Como estes mapas podem ser gerados a partir de diferentes populações de diferentes tamanhos, o ordenamento poderá variar de uma população para outra para certas marcas, principalmente para aquelas bastante próximas. Nesta ilustração o ordenamento foi o mesmo.

Também é considerado a disponibilidade das matrizes de distância, em porcentagem de recombinação, entre cada par de marcadores em cada mapa, que serão denominadas de D_1, D_2 e D_3. Deve-se ressaltar a possibilidade de reconstituir D1, D2, D3 a partir das informações dos mapas, porém grande quantidade de erro deverá ser incorporada a matriz em razão do desconhecimento da verdadeira taxa de interferência entre regiões adjacentes.

Sabe-se que, para cada mapa de ligação, seria possível obter estimativas multiloco da distância entre os pares de âncoras, para aquele ordenamento.

Assim para o mapa 1, pode-se ter :



Sendo Mij a distância entre as marcas, i e j e mi o valor da distância obtido pela análise multiloco. Distância entre as marcas dada pela análise multiloco.

Para obtenção do vetor m é utilizado o sistema de equações:

Pm=Q

Em que

$$P = \begin{pmatrix} & a_{LC} \\ b_{LC} & \end{pmatrix}_{n-1xn-1} \qquad Q = \begin{pmatrix} q_1 \\ \cdots \\ q_{n-1} \end{pmatrix}_{n-1xi} \qquad m = \begin{pmatrix} m_1 \\ \\ m_{n-1} \end{pmatrix}$$

sendo :

n: número de marcas moleculares

a_{LC}: elemento da diagonal ou acima da diagonal da matriz P, dado por:

$$a_{LC} = \sum_{i=1}^{L} \sum_{j>c}^{n} L_{ij} \quad (L \leq C)$$

b_{LC}: elemento abaixo da diagonal da matriz P, dado por:

$$b_{LC} = \sum_{i>L}^{n} \sum_{j=1}^{c} L_{ij} \quad (L>C)$$

q_i=elemento da i-ésima linha do vetor Q

$$q_i = \sum_{i=1}^{L} \sum_{j>L}^{n} L_{ij} M_{ij}$$

L_{ij} é o peso devido a informação do valor da distância m_{ij}, admitido neste estudo como sendo o inverso da variância de m_{ij}. A variância depende do tipo de população, do tipo de marcador, da fase de ligação e próprio valor de m_{ij}.

Se a matriz D tiver sido reconstituída a partir do mapa genético esta análise será inócua, pois reconstituirá exatamente o mapa fornecido, para a função de mapeamento utilizado.

Entretanto, para obtenção do mapa consenso o valor de distância entre par de marcadores poderá ser melhor estimado não só pela informação dos elementos de uma particular matriz D, mas de todas elas tomadas em conjunto.

Assim, tem-se:

$P_1 m = Q_1$ para o mapa 1

$P_2 m = Q_2$ para o mapa 2

.

.

$P_c m = Q_c$ para o c-ésimo mapa de ligação

De forma conjunta, pode-se escrever:

$[P_1+P_2+.....Pc]m = [Q_1+Q_2+....Q_c]$

$P_c m=Q1_c$ e $m=P_c^{-1}Q_c$

3.6 - Comparação dos genomas

Para facilitar a compreensão, o termo "genoma simulado" deve ser referenciado como o verdadeiro (original). O termo "genoma analisado" refere-se, aos genomas construídos a partir das populações simuladas, o qual apresentará distorção em razão do processo de estimação, amostragem, critérios de agrupamento, dentre outros. O termo "genoma consenso" será utilizado para designar os genomas obtidos pela integração dos mapas provenientes das diferentes populações.

Em cada situação foram estudados os mapas consenso construídos a partir dos genomas analisados, e realizadas comparações quanto aos números de grupos de ligação obtidos, o número de marcas por grupo, os tamanhos dos grupos de ligação, as distâncias médias entre marcadores adjacentes nos grupos de ligação, as variâncias das distâncias entre marcas adjacentes nos grupos de ligação, o estresse, e se ocorrerá ou não inversão da ordem dos marcadores, verificada pela correlação de Spearman. Todas estas comparações foram efetuadas com utilização do Módulo: "Comparação de genomas", do aplicativo computacional GQMOL (Cruz, 2008).

Nas análises apresentadas a seguir foram utilizadas apenas as repetições (populações) em que houve formação de onze grupos de ligação no mapeamento genético.

3.6.1 Número de grupos de ligação e de marcas por grupo

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

3.6.2 Tamanho do grupo de ligação

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

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

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

3.6.3 Distância média de dois marcadores adjacentes no grupo de ligação

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

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

3.6.4 Variâncias das distâncias entre marcas adjacentes

Esta medida faz-se útil uma vez que o genoma original (simulado) apresentará marcas eqüidistantes tendo-se variância nula. Então foi avaliada a variância do genoma analisado partindo do pressuposto que o ideal é ter variância nula. A estimativa é dada por:

$\hat{\sigma}^2 = \frac{\sum_{k=1}^{m-1}\left(d_k - \bar{d}\right)^2}{I-1}$, I é o número de intervalos dado por m-1.

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

3.6.5 Correlação de Spearman

Foi empregada para avaliar o grau de concordância do ordenamento das marcas nos genomas analisados e o simulado, não é afetada pelas distâncias entre marcadores. É também conhecida como correlação de *rank*, sua aplicação é importante quando não é possível mensurar variação contínua das variáveis *x* e *y* nos *n* membros de uma população. Mas, sendo possível mensurar um *x* e um *y*, em forma de nota (*rank*), onde cada nota pode ser colocada em ordem para os *n* membros. Esta correlação expressa o grau de concordância nas notas das duas variáveis, desta forma a sua utilização na análise de genomas é proposta conforme descrito a seguir.

Adaptando-se a correlação de Spearman, para análise de genomas, tem-se:

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

em que: r_s é o valor estimado da correlação de Spearman para um grupo de ligação do genoma analisado $(-1 \leq r_s \geq 1)$, Δ_k é a diferença da nota do marcador M_k (k = 1,..., m) na posição k do grupo de ligação do genoma simulado e a nota do marcador M_k na posição k do grupo de ligação do genoma analisado. Onde, m é o número de marcadores no grupo de ligação do genoma simulado e a nota do marcador m_k, tanto no grupo de ligação do genoma simulado quanto no grupo de ligação do analisado, é o valor do índice k do referido marcador.

3.6.6 Estresse

O coeficiente de estresse (S) é utilizado como medida de adequação das distâncias estimadas em representar as verdadeiras distâncias estabelecidas no genoma simulado. Será estabelecido com o propósito similar ao apresentado em estudos de divergência genética (Cruz e Carneiro, 2003). É dado por:

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

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

3.6.7 Testes de comparação múltipla de médias

As médias das variáveis tamanho de grupo de ligação, distância média de marcas adjacentes, variância e estresse em cada população, para cada grupo de ligação, foram obtidas para cada situação anteriormente simulada, e foram comparados pelo teste de comparação múltipla de médias de Tukey a 1% de probabilidade (erro tipo I), com o auxílio do aplicativo computacional GENES (Cruz, 2001). Também foram comparadas as médias gerais (médias de todos os grupos de ligação), nas diferentes populações, para cada tipo de sub-genomas.

3.7 – Fluxograma ilustrativo

Para facilitar o entendimento da logística utilizada no trabalho de simulação é apresentado o fluxograma na (Figura 8). Os passos seguidos no processo de simulação foram:

1º) simulação dos genomas com nível de saturação de 5, 10 e 15 cM e 11 marcas moleculares;

2º) a partir dos genomas simulados foram construídos os genótipos dos genitores;

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

4º) as populações segregantes foram mapeadas;

5º) Integração dos mapas entre populações F_2 e retrocruzamento de diferentes tamanhos.

6°) os mapas obtidos foram comparados com o genoma simulado. Para a comparação foram utilizados os critérios apresentados no quadro mostrado no interior do fluxograma.

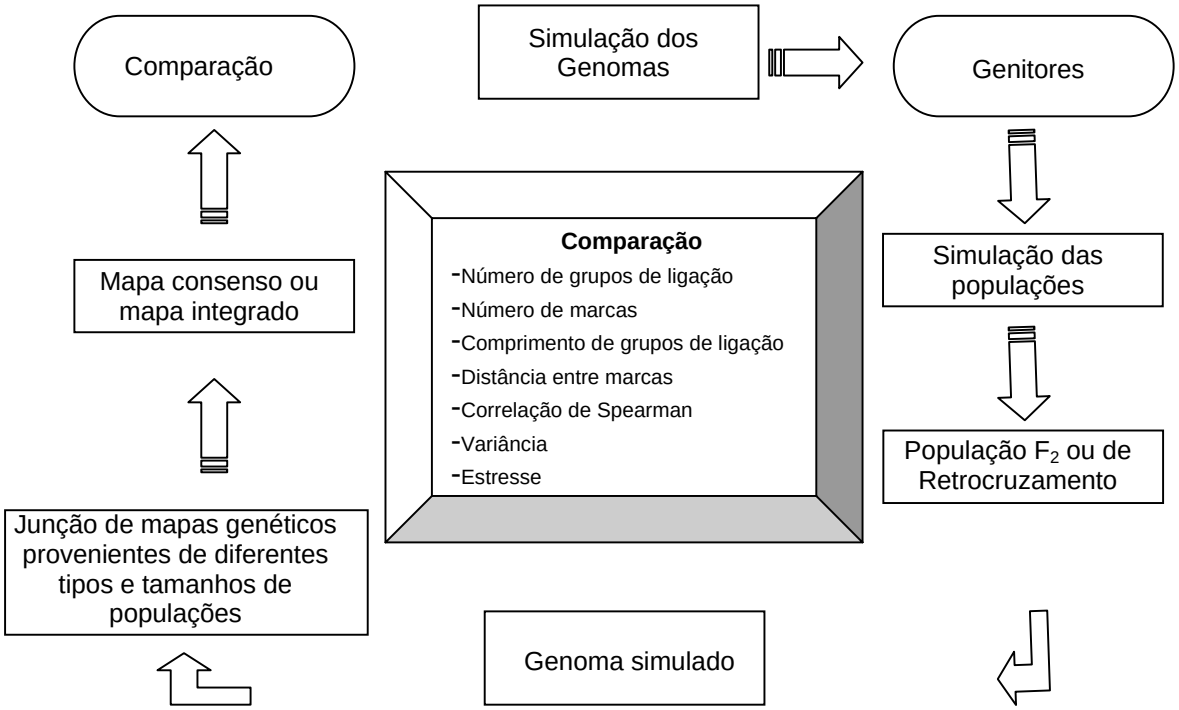


Figura 8 – Fluxograma ilustrativo do processo de simulação realizado neste trabalho.

4 Resultados e discussão

4.1 Qualidade dos mapas consenso individuais

Procurou-se avaliar a qualidade dos mapas consenso obtidos a partir da junção de três outros, gerados com diferentes tamanhos de população, tipos de população e tipo de marcador. Foram consideradas duas abordagens:

1 – A qualidade de um mapa consenso resultante da junção de mapas oriundos de diferentes tipos de populações de mesmo tamanho.

2 – A qualidade de um mapa consenso resultante da junção de mapas oriundos de um mesmo tipo de população com tamanhos diferentes.

Diferentes trabalhos com integração de mapas genéticos de diversas espécies cultivadas (Tabela 3), têm sido publicados nos últimos anos, utilizando informações de diferentes mapas construídos com diferentes tipos de marcadores e populações. Entretanto, pouco tem sido discutido sobre a resolução destes mapas integrados. A resolução do mapa pode ser entendida como a capacidade de se determinar a seqüência de marcadores, nele, estão diretamente relacionadas ao tamanho da amostra ou população. Respostas a estes questionamentos não são totalmente encontradas na literatura.

Tabela 3. Cultura, tamanho de população, número de marcadores, comprimento do mapa e número de marcadores âncoras utilizados para integração de mapas genéticos

Cultura	Nº populações	Nº de marcadores	Nº de marcadores âncora*	Referência
Cevada	6 (3RIL;3DH)	3258	502	Marcel et al. (2006)
Canola	3 (DH)	540	253	Lombard e Delourme (2001)
Algodão	4 ($F_{2,3}$)	284	-	Ulloa et al. (2002)
Milheto	4 (F2)	418	116	Qi et al. (2004)
Trigo	4	1235	-	Somers et al. (2004)
Azevém	2	922	42	Saha et al. (2005)
Seringueira	2	717	-	Lespinasse et al. (2000)
Videira	5	515	-	Doligez et al. (2006)
Cana-de-açúcar	Pseudo-testcross	357	-	Garcia et al. (2006)
Pimenta	6	2262	320	Paran et al. (2004)
Milho	23	5650	2100	Peleman et al. (2000)
Citrus	2	217	61	Oliveira et al. (2005)
Soja	5	1849	-	Song et al. (2004)
Feijão	3	563	-	Freyre et al. (1998)

* marcadores comuns em pelo menos dois mapas individuais; RIL: Recombinante Inbreed Line; DH: Duplo haplóide

4.1.1 Obtenção dos mapas consenso

Foram simuladas populações de retrocruzamento, F_2 dominante e co-dominante, com 3 grupos de ligação originalmente estabelecidos com 11 marcas em cada grupo, em que se variou a distâncias entre as marcas (eqüidistantes 5, 10 e 15 cM) e o tamanho da população (100, 200 e 400 indivíduos), com 10 repetições para cada população.

A qualidade dos mapas para estudos genéticos já é discutida na literatura. Segundo Soller e Beckmann (1983), quando marcadores co-dominantes estão disponíveis, análises baseadas em gerações F_2 são mais úteis que aquelas com base em gerações de retrocruzamento, por fornecerem informações tanto em relação à dominância quanto ao efeito maior do QTL identificado. Isso também pode ser uma explicação para a necessidade de maior tamanho de população de retrocruzamento para obtenção de mapas mais confiáveis.

Ooijen (1992) ressaltou que o procedimento de mapeamento de QTL de Lander e Botstein (1989) permite determinar, com limitações, a posição do QTLs em um mapa. QTLs com pequeno efeito aditivo ($\sigma^2_{explicada}$ = 1%), sendo necessária a análise de pelo menos 200 indivíduos, para que efeitos aditivos dessa magnitude sejam detectados. Considerando uma situação prática, 400 indivíduos parecem o menor número viável para trabalhos com RFLP, a menos que se esteja interessado apenas em genes de efeito muito grande ($\sigma^2_{explicada}$ >1%)

Previamente a realização do mapeamento genético de cada população segregante simulada, foram aplicados testes de qui-quadrado (χ^2) para verificar se a razão de segregação do locos individuais estava de acordo com o esperado, pela primeira lei de Mendel. Não foram observadas ocorrência de marcadores com razão de segregação diferente da esperada (P<0,05, dados não apresentados), 3:1 F_2 dominante, 1:1 retrocruzamento e 1:2:1 F_2 co-dominante. Estes resultados serviram como indicativo da boa qualidade dos dados gerados no processo de simulação

O número de grupos de ligação esperado no processo de mapeamento nas populações simuladas era igual a 3, independente da saturação do genoma utilizado para geração das populações segregantes. Porém, foi observada a ocorrência de mapas onde não houve a recuperação dos 3 grupos de ligação (Tabela 4) ocorrendo a formação de número maior de indivíduos ou menor. Segundo Cruz (2006), em populações F_2, tamanhos próximos de 100 e 200 indivíduos podem ser utilizados para uma recuperação completa do genoma, no caso de saturação de 5 e 10 cM, respectivamente. No caso de populações de retrocruzamento, as amostras devem ter o tamanho mínimo de 200 indivíduos nos níveis de saturação de 5 e 10 cM. As repetições que apresentaram número de grupos de ligação diferente de 3 foram descartadas devida a falta de informação ou informações errôneas geradas a partir dessas amostras. Exemplos de formação de grupos de ligação com número diferente do desejado pode ser visto na figura 9.

Observando-se a Figura 9 tem-se idéia do que acontece quando não se tem número de indivíduos e de marcas adequados. Os grupos que originalmente comportariam as 11 marcas não foram identificados, e no lugar deles foram identificados 2 grupos com 11 marcas e 2 grupos com número diferente de marcas, um com 6 marcas e o outro com 5 marcas.

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

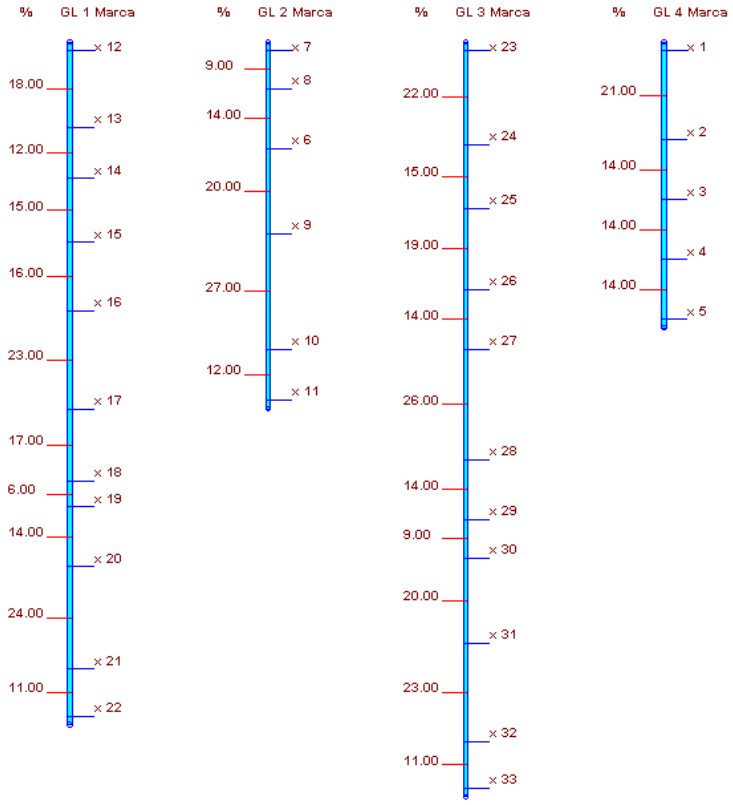


Figura 9: Representação da fragmentação dos 3 grupos de ligação originais em 4 grupos de ligação. Observe que o grupo de ligação 1 foi dividido em dois grupos, tendo, cada um, 5 e 6 marcadores.

A inversão na ordem dos marcadores também é um fator importante a ser verificado, corroborando com a confiabilidade dos dados obtidos. As inversões podem se dar de várias formas, havendo casos em que o grupo de ligação é recuperado, mas com alterações da ordem de uma ou mais marcas, (Figura 10). Além disso, as inversões podem acontecer, também, dentro de grupos de ligação já invertidos, dificultando a decisão de considerar, para fins de análise, este grupo de ligação.

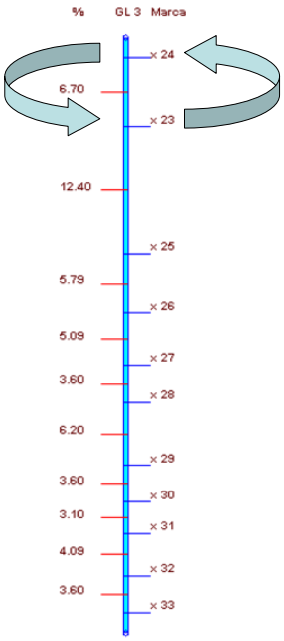


Figura 10: Representação da inversão de marcadores em um grupo de ligação. A seta indica o par de marcadores invertidos no grupo de ligação.

Um fator a desqualificar um determinado tamanho da população para análise é a junção de grupos de ligação. Essa junção pode ser total, em que um grupo inteiro se liga a outro grupo inteiro, ou parcial, quando um grupo de ligação se liga à parte de outro grupo de ligação. Na Figura 11 é apresentada uma situação em que um grupo de ligação formado possui parte dos 3 grupos previamente estabelecidos.

Em trabalhos de simulação é possível inferir sobre a qualidade do mapa, considerando o ordenamento e as distâncias entre marcadores. Entretanto, na prática os mapas são estabelecidos sem que se possa avaliar a sua verdadeira adequação e capacidade de representação do fenômeno genético estudado. Certamente que fazer a junção de mapas de alta qualidade, permitirá obter maior acurácia no mapa consenso. Mas quanto que um mapa de baixa qualidade poderá reduzir a acurácia, e se o mapa de alta qualidade aumentar esta acurácia é digno de investigação.

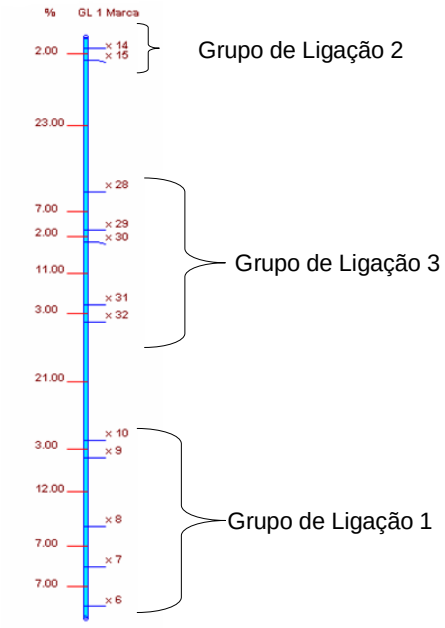


Figura 11: Representação de um grupo de ligação oriundo da junção de três grupos de ligação.

4.1.2 Estratégias para obtenção de mapas consenso

No procedimento de construção de mapas consenso foram descartados os mapas que apresentaram número de grupos de ligação diferente de 3, que é o número de grupos de ligação que se tem no genoma original, e os que apresentavam junção de grupos de ligação. Na tabela 4 estão apresentados o número de repetições que recuperou os 3 grupos de ligação para tamanhos de população de 100, 200 e 400 indivíduos, assim como número de inversões apresentados em cada um destes tamanhos de população. Segundo Young (1994), amostras com menos de 50 indivíduos, provavelmente, terão baixa resolução de mapeamento. Assim espera-se que à medida que se aumenta o número de indivíduos, a qualidade dos mapas aumente, como pôde ser constatado que o maior número de ocorrências para populações que não recuperaram os 3 grupos de ligação foram as populações com 100 indivíduos.

Foram construídos dois tipos de mapas consenso:

1 – Entre populações de mesmo tamanho – por meio da união dos mapas provenientes de 3 tipos de populações (F_2 co-dominante, F_2 dominante, retrocruzamento) com a mesma saturação (5, 10 ou 15 cM) e mesmo tamanho população (100, 200 ou 400).

2 - Dentro de populações de mesmo tamanho – pela junção de mapas provenientes da mesma população com a mesma saturação e diferentes tamanhos populacionais.

Os mapas consenso foram construídos como indicado na Figura 12, que ilustra os modos de obtenção dos mapas consenso obtidos a partir da saturação de 5 cM com os diferentes tipos de população (F_2 dominante, F_2 co-dominante e retrocruzamento) e diferentes tamanhos populacionais. O mapa consenso 1 entre populações foi obtido pela união de todos os mapas construídos a partir dos dados indicados pelo número um, os mapas consenso 2 e 3 foi feito o mesmo e assim subseqüentemente para todas as repetições. Para o mapa consenso 1, dentro de populações efetuou-se a união dos mapas construídos com os dados indicados por 1, 2 e 3, para os mapas consenso 2 e 3 repetiu-se o mesmo processo e também para todas as repetições. O processo repetiu para as saturações de 10 e 15 cM.

Para cada mapa consenso foram realizados 10 repetições, sendo que 5 delas foram provenientes de mapas que em sua construção utilizou-se a análise multiloco e 5 que não utilizou análise multiloco. O número de repetições de cada mapa consenso obtido, como descrito no processo (Figura 12), variou de acordo com o numero mínimo de recuperação dos 3 grupos de ligação dentro de cada grupo de mapas utilizados na integração.

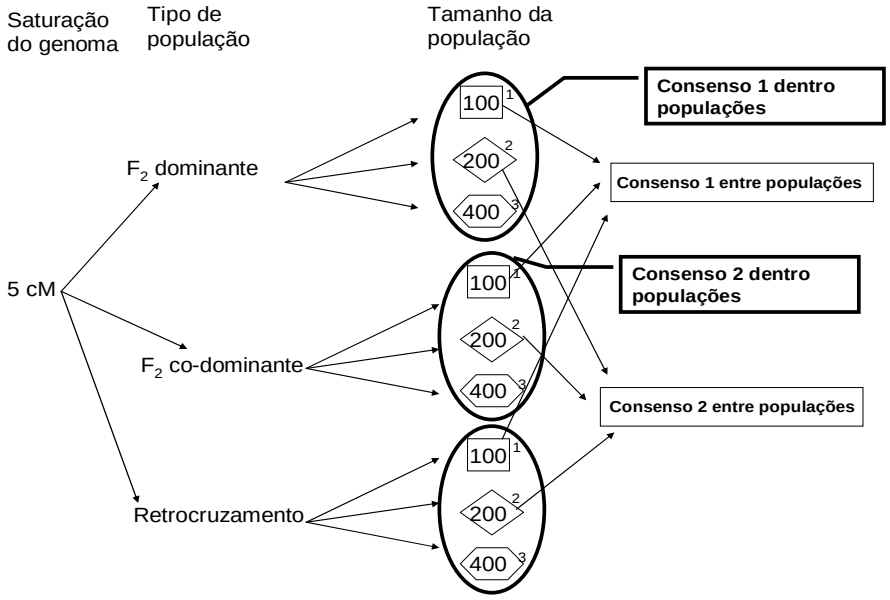
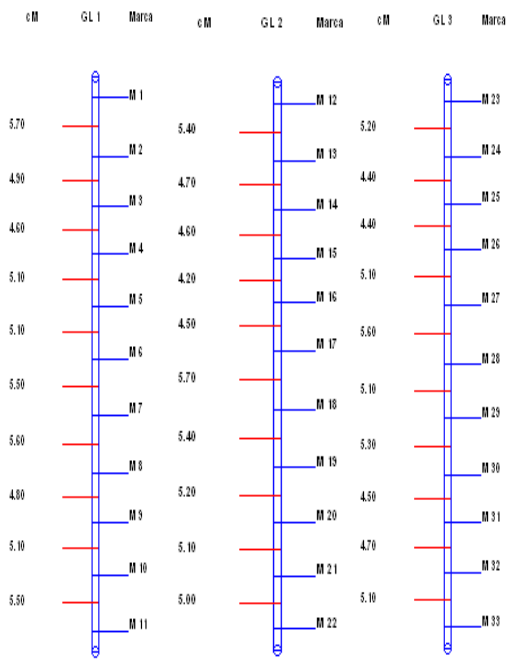


Figura 12 –Integração entre populações – união dos mapas com tamanho populacional 100 das diferentes populações gerando o consenso 1 entre populações, o mesmo processo se repetiu para os mapas construídos com população de 200 e 400 indivíduos. Integração dentro de populações – junção dos mapas construídos com população F_2 dominante de tamanhos 100, 200 e 400 indivíduos dentro de uma mesma população gerando o mapa consenso 1 dentro de populações, o mesmo se repetiu para as populações F_2 co-dominante e de retrocruzamento.

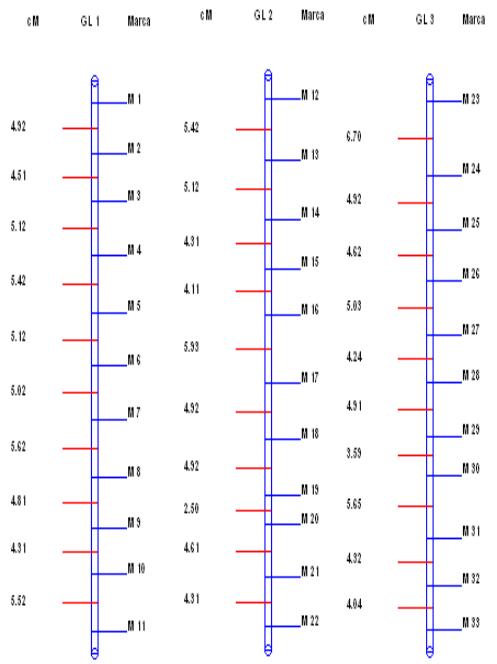
4.1.3 Genomas utilizados para comparação

Os genomas utilizados para comparação com os mapas consenso foram gerados a partir de saturações de 5, 10 e 15 cM e populações de 1000 indivíduos. Para cada um destes níveis de saturação foi gerado mapas com análise multiloco e sem análise multiloco. Estes mapas foram considerados ideais e utilizados como referência ("genoma ideal"). As comparações foram feitas entre os mapas consenso construídos a partir de um "genoma original" com o mesmo nível de saturação que foi gerado o "genoma ideal" (Figura 13). Somente para as correlações de Spearman que foram apresentadas as comparações entre o "genoma simulado" e "genoma original".

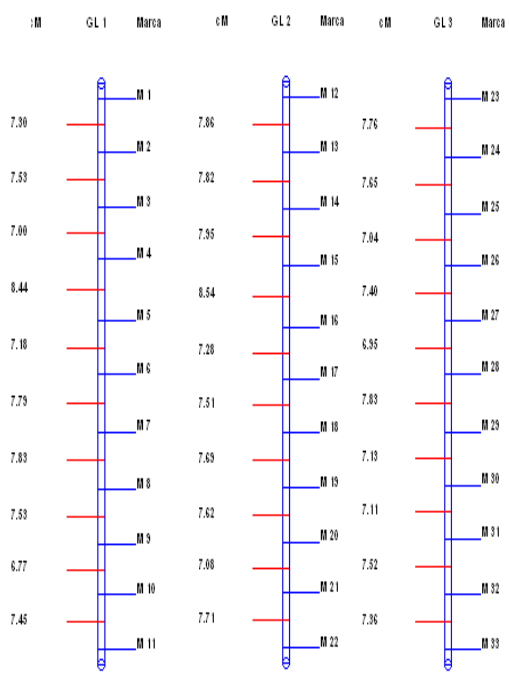
1a


cM
GL 1
Marca
M 1
M 2
M 3
M 4
M 5
M 6
M 7
M 8
M 9
M 10
M 11
5.70
4.90
4.60
5.10
5.10
5.50
5.60
4.80
5.10
5.50
cM
GL 2
Marca
M 12
M 13
M 14
M 15
M 16
M 17
M 18
M 19
M 20
M 21
M 22
5.40
4.70
4.60
4.20
4.50
5.70
5.40
5.20
5.10
5.00
cM
GL 3
Marca
M 23
M 24
M 25
M 26
M 27
M 28
M 29
M 30
M 31
M 32
M 33
5.20
4.40
4.40
5.10
5.60
5.10
5.30
4.50
4.70
5.10

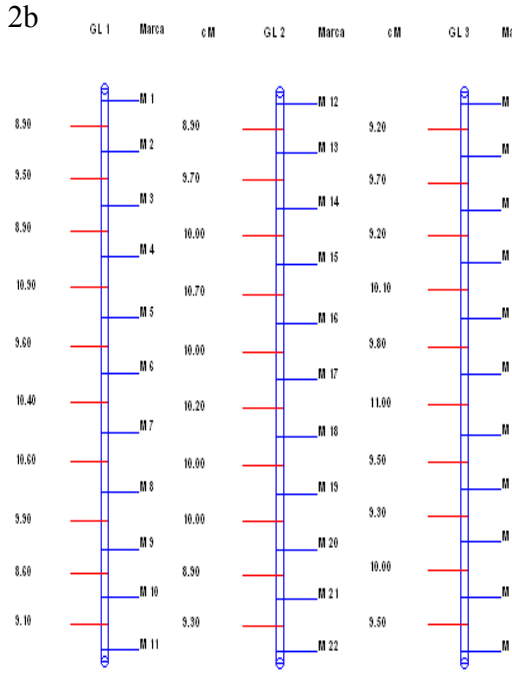
1b


cM
GL 1
Marca
M 1
M 2
M 3
M 4
M 5
M 6
M 7
M 8
M 9
M 10
M 11
4.92
4.51
5.12
5.42
5.12
5.02
5.62
4.81
4.31
5.52
cM
GL 2
Marca
M 12
M 13
M 14
M 15
M 16
M 17
M 18
M 19
M 20
M 21
M 22
5.42
5.12
4.31
4.11
5.33
4.92
4.52
2.50
4.61
4.31
cM
GL 3
Marca
M 23
M 24
M 25
M 26
M 27
M 28
M 29
M 30
M 31
M 32
M 33
6.70
4.92
4.62
5.03
4.24
4.31
3.59
5.65
4.32
4.84

2a


:M
GL 1
Marca
M 1
M 2
M 3
M 4
M 5
M 6
M 7
M 8
M 9
M 10
M 11
7.30
7.53
7.00
8.44
7.18
7.79
7.83
7.53
6.77
7.45
cM
GL 2
Marca
M 12
M 13
M 14
M 15
M 16
M 17
M 18
M 19
M 20
M 21
M 22
7.86
7.82
7.95
8.54
7.28
7.51
7.69
7.62
7.08
7.71
cM
GL 3
Marca
M 23
M 24
M 25
M 26
M 27
M 28
M 29
M 30
M 31
M 32
M 33
7.76
7.65
7.04
7.40
6.95
7.83
7.13
7.11
7.52
7.36

2b


GL 1
Marca
M 1
M 2
M 3
M 4
M 5
M 6
M 7
M 8
M 9
M 10
M 11
8.90
9.50
8.90
10.90
9.60
10.40
10.60
9.90
8.60
9.10
cM
GL 2
Marca
M 12
M 13
M 14
M 15
M 16
M 17
M 18
M 19
M 20
M 21
M 22
8.90
9.70
10.00
10.70
10.00
10.20
10.00
10.00
8.90
9.30
cM
GL 3
Marca
M 23
M 24
M 25
M 26
M 27
M 28
M 29
M 30
M 31
M 32
M 33
9.20
9.70
9.20
10.10
9.80
11.00
9.50
9.30
10.00
9.50

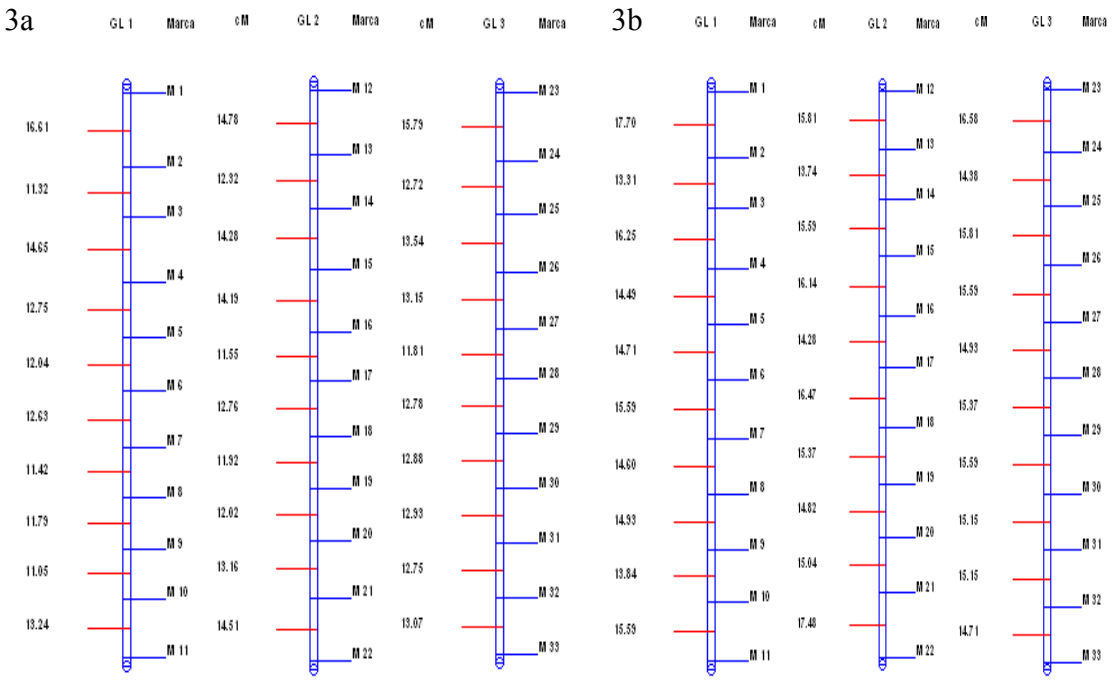


Figura 13 – 1a(com análise multiloco) e 1b(sem análise multiloco) mapa ideal usado para comparação, saturação 5 cM; 2a(com análise multiloco) e 2b(sem análise multiloco) mapa ideal usado para comparação, saturação 10 cM; 3a(com análise multiloco) e 3b(sem análise multiloco) mapa ideal usado para comparação, saturação 15 cM.

4.2 Comparação de Genomas

4.2.1 Correlação de Spearman entre medidas de distância

A correlação de Spearman foi utilizada para a identificação de inversão de posição de marcas dentro de cada um dos três grupos de ligação formados no mapeamento. A avaliação das inversões foram realizadas em repetições em que foram recuperados os três grupos de ligação. Na Tabela 4 está o número de inversões identificadas pela correlação de Spearman dos "genomas analisados" em relação ao "genoma original" e na Tabela 5a e 5b estão as correlações de Spearman do "genoma consenso" em relação ao "genoma ideal" para os mapas consenso entre e dentro de populações, respectivamente.

Valores de correlação de Spearman iguais à unidade indicam que a ordem das marcas no grupo de ligação obtido no mapeamento das populações segregantes não foi alterado em relação à ordem previamente estabelecida no genoma o qual foi utilizado para a geração das populações. Valores de correlação de Spearman diferentes da unidade, ou seja, menores do que 1, 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 o qual foi utilizado para a geração das populações.

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

Para o nível de saturação de 5 cM, para o "genoma analisado", os tamanhos de população de 100 indivíduos apresentam repetições com valores de correlação de Spearman diferente da unidade para F_2 co-dominante e retrocruzamentos em mapas construídos com análise multiloco. Para populações de tamanho 200 e 400 indivíduos não foram observados valores diferentes de um. Sendo que a presença de marcas invertidas no grupo de ligação somente nas populações de menor tamanho 100 e 200 indivíduos, o mesmo fato foi observado quando obteve-se os mapas consenso, o qual se justifica pela presença de maior número de marcas invertidas nos "genomas analisados".

Para o nível de saturação de 10 cM, para o "genoma analisado", o tamanho de população de 200 indivíduos apresentou inversão na ordem das marcas para população co-dominante e retrocruzamentos, para mapas construídos com e sem análise multiloco. Um resultado inesperado foi observado para as populações geradas a partir do genoma de saturação de 15 cM, na qual não foram observadas inversões nos genomas analisados.

A inversão de ordem nas marcas dentro de grupos de ligação foi maior tanto quanto menores foram os tamanhos de populações utilizadas na construção dos mapas. Portanto, a utilização de populações de tamanhos inadequados nos trabalhos de mapeamento pode levar a problemas relacionados a posicionamento de marcas nos grupos de ligação encontrados, podendo por exemplo, levar a resultados equivocados nos processos de mapeamento de QTL e seleção assistida por marcadores moleculares (Silva, 2005).

Em uma avaliação rápida poderia se afirmar que o nível de saturação de 15 cM seria mais indicado para construção de mapas consenso quando comparados com os níveis de saturação de 5 e 10 cM, uma vez que os mapas consenso construídos com este nível de saturação apresentaram menor número de marcas invertidas. Mas a avaliação na qualidade da saturação e do tamanho populacional em formar bons mapas não deve ser apenas em função da correlação de Spearman. Deve-se considerar, também, que este nível de saturação foi o que apresentou maior número de repetições que não recuperaram os 3 grupos de ligação.

Para os mapas consenso gerados pela integração dos diferentes mapas analisados entre populações de mesmo tamanho, foram observadas apenas 3 inversões entre todos os mapas analisados, sendo duas delas nas populações com tamanho populacional de 100 indivíduos e saturação de 10 cM e a terceira no tamanho populacional de 200 indivíduos e saturação de 5 cM, as quais não podem ser consideradas como indicativo de pior qualidade dos mapas consenso gerado, pois os genomas analisados já apresentavam inversões nestes níveis estudados. Fato interessante de ressaltar, para população F_2 co-dominante com saturação 5 e 10 cM e tamanho populacional de 100 e 200 indivíduos e para população de retrocruzamento com saturação de 10 cM e tamanho populacional 200 estas apresentavam inversões no genoma analisado e após a integração entre as populações nas respectivas saturações os mapas consenso gerados não apresentaram inversões recuperando a qualidade dos mapas (Tabela 5a).

Para os mapas consenso gerados dentro de populações de diferentes tamanhos, foi observado um número maior de inversões quando comparado com os mapas consenso entre populações, quatro inversões nas populações F_2 dominante e duas nas populações de retrocruzamento. Quando se compara o número de inversões dos mapas consenso com os genomas analisados observa-se que, nos mapas consenso, o número de inversões para populações de retrocruzamento se manteve e para F_2 dominante o número foi maior, apresentando três inversões a mais que nos genomas analisados. Quanto às populações F_2 co-dominantes não foram observadas inversões nos mapas consenso, recuperando a qualidade de alguns mapas.(Tabela 5b)

Tabela 4. Dados de população F_2 co-dominante, dominante e retrocruzamento. Saturação do genoma 5, 10 e 15 cM, tamanho da população, número de repetições com 3 grupos de ligação e número de inversões nas populações

População	Saturação *do genoma*	*Tamanho da população*	*Repetições*		*Correlação de Spearman*	
			*Multiloco**	*S/multiloco***	*Multiloco**	*S/multiloco***
F_2 co-dominante	5	100	5	5	1	0
		200	5	5	0	0
		400	5	5	0	0
	10	100	5	5	0	1
		200	5	5	1	0
		400	5	5	0	0
	15	100	4	5	0	0
		200	5	4	0	0
		400	5	5	0	0
F_2 dominante	5	100	5	5	0	0
		200	5	5	1	0
		400	5	5	0	0
	10	100	4	5	0	0
		200	5	5	0	0
		400	5	5	0	0
	15	100	4	4	0	0
		200	5	5	0	0
		400	5	5	0	0
População de Retrocruzamento	5	100	5	5	1	0
		200	5	5	0	0
		400	5	5	0	0
	10	100	4	5	0	0
		200	5	5	1	0
		400	5	5	0	0
	15	100	4	3	0	0
		200	4	4	0	0
		400	5	5	0	0

*Repetições em que foi utilizada análise multiloco para construção dos mapas

**Repetições em que não foi utilizada análise multiloco para construção dos mapas

Tabela 5a. Mapas consenso entre populações. Saturação do genoma, tamanho da população, número de repetições com 3 grupos de ligação e número de inversões nas populações.

Saturação	*Tamanho da*	*Repetições*		*Correlação de Spearman*	
do genoma	*população*	*Multiloco**	*S/multiloco***	*Multiloco**	*S/multiloco***
	100	5	5	0	0
5	200	5	5	1	0
	400	5	5	0	0
	100	4	5	1	1
10	200	5	5	0	0
	400	5	5	0	0
	100	4	3	0	0
15	200	4	4	0	0
	400	5	5	0	0

*Repetições em que foi utilizada análise multiloco para construção dos mapas

**Repetições em que não foi utilizado análise multiloco para construção dos mapas

Tabela 5b. Mapas consenso dentro de populações. Saturação do genoma, tamanho da população, número de repetições com 3 grupos de ligação e número de inversões nas populações

Saturação	*Tamanho da*	*Repetições*		*Correlação de Spearman*	
do genoma	*população*	*Multiloco**	*S/multiloco***	*Multiloco**	*S/multiloco***
	Co-dominante	5	5	0	0
5	Dominante	5	5	2	0
	Retrocruzamento	5	5	1	0
	Co-dominante	4	5	0	0
10	Dominante	5	5	2	0
	Retrocruzamento	5	5	1	0
	Co-dominante	4	3	0	0
15	Dominante	4	4	0	0
	Retrocruzamento	5	5	0	0

*Repetições em que foi utilizada análise multiloco para construção dos mapas

**Repetições em que não foi utilizado análise multiloco para construção dos mapas

4.2.2 Tamanhos do grupos de ligação

Com relação ao comprimento médio de cada grupo de ligação, foi obtida a média aritmética dos comprimentos dos grupos de ligação dos mapas consenso obtidos a partir de mapas construídos com e sem análise multiloco. Os valores dos comprimentos médios são apresentados nas Tabelas 6a e 6b.

O comprimento esperado nos grupos de ligação após a integração dos mapas das populações F_2 co-dominante, F_2 dominante e retrocruzamento era de 50 cM quando se integrou mapas com saturação de 5cM e de 100 e 150 cM quando houve a integração de mapas com 10 e 15 cM de saturação, uma vez que estes são os comprimentos de cada grupo de ligação no nível de saturação do genoma utilizado para a simulação das populações.

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

Para os mapas consenso entre populações (Tabela 6a), observou-se que, em geral, que à medida que aumenta o tamanho da população segregante aumenta o tamanho dos genomas tendendo a aproximar-se cada vez mais do tamanho do genoma original. Uma explicação para a aproximação do tamanho médio dos grupos de ligação ao valor esperado do genoma original é que à medida que aumenta o tamanho da população segregante ocorre a redução da amplitude de variação dos tamanhos de grupos de ligação obtidos entre as repetições. Esta redução da amplitude de variação dos tamanhos de grupos de ligação em torno da média pode ser observada pela redução do desvio padrão (Tabela 9a) com o aumento do número de indivíduos nas populações segregantes. A exceção foi para os mapas de saturação de 10 cM com tamanho 100, 200 e 400 construídos com análise multiloco, os quais os tamanhos médios dos grupos de ligação diminuíram com o aumento do tamanho da população. Esse resultado poderia ser explicado, pela imprecisão das estimavas da análise multiloco. Mais mesmo assim, à

medida que aumentou o número de indivíduos, o desvio padrão entre as classes decaiu. Por exemplo, para a saturação de 10 cM tem-se que o desvio padrão foi de 3,864 e 0,746 para os tamanhos de população de 100 e 400 indivíduos.

Além da explicação em termos estatísticos, há uma explicação biológica dada pela melhoria das estimativas das freqüências de recombinação à medida que se aumenta o tamanho das populações segregantes utilizadas no mapeamento, consequentemente melhorando as estimativas de freqüência de recombinação, melhor a recuperação dos mapas integrados.

O teste de médias permitiu a avaliação estatísticas dos efeitos do tamanho da população na recuperação do tamanho dos grupos de ligação formado na integração dos mapas genéticos. Com o aumento no número de indivíduos das populações, para os mapas construídos com e sem análise multiloco foi observado uma diferença significativa entre pelo menos uma das classes de médias, pelo teste de tukey a 5%, seja entre as médias dos grupos de ligação ou das médias gerais do comprimento dos grupos de ligação. Isso foi observado principalmente entre as populações de 100 e 400 indivíduos. Dentro de cada grupo analisado observa-se que as populações com 400 indivíduos obtevou-se uma média mais ajustada aos valores observados nos mapas "ideais". O que pode ser corroborado com a queda do desvio padrão e do estresse quando se aumenta o número de indivíduos na população, os menores valores de desvio padrão são observados neste tamanho populacional. Por exemplo, para a saturação de 10 cM, sem análise multiloco, tem-se que o desvio padrão foi de 4,476 e 0,630 para os tamanhos de população de 100 e 400 indivíduos, respectivamente. Podendo levar a uma inferência que mesmo para integração de mapas o tamanho da população influência na maior confiabilidade da reconstituição dos mapas, maior número de indivíduos maior confiança nos processos de integração e tamanho dos grupos de ligação.

Para os mapas consenso dentro de população não foi observado um padrão nas médias dos grupos de ligação (Tabela 6b), uma vez que a integração ocorreu entre mapas construídos com populações de diferentes tamanhos. A realização do teste de Tukey ($P<0,05$) permitiu a avaliação estatística dos efeitos do tipo de população na integração de mapas genéticos. Para as populações F_2 co-dominates e F_2 dominante houve diferença significativa entre as médias gerais de quase todos os níveis de saturação estudados com e sem análise multiloco, com exceção para os mapas com saturação de 5 cM, utilizando análise multiloco.

Tabela 6a. Mapa consenso entre populações. Média aritmética do comprimento, em cM, de três grupos de ligação em três tamanhos de população com marcadores eqüidistantes a 5, 10 e 15 cM com/sem análise multiloco para construção dos mapas genéticos.

Saturação	*Tamanho populacional*	*Com análise multiloco*				*Sem análise multiloco*			
		Grupos de ligação			*Média Geral*	*Grupos de ligação*			*Média Geral*
		1	*2*	*3*		*1*	*2*	*3*	
5	100	41.447 (b)[1]	37.530 (b)	43.014 (b)	40.664 (b)	47.00808 (a)	49.226 (a)	49.61462 (a)	48.616 (ab)
	200	40.326 (b)	37.664 (b)	40.624 (b)	39.539 (b)	46.58598 (a)	49.137 (a)	47.20846 (a)	47.643 (b)
	400	52.390 (a)	50.715 (a)	51.777 (a)	51.627 (a)	52.66906 (a)	50.490 (a)	52.15622 (a)	51.771 (a)
10	100	92.096 (a)[1]	95.608 (a)	88.006 (a)	91.903 (a)	90.07396 (b)	88.004 (b)	87.01572 (b)	88.364 (b)
	200	82.604 (a)	78.732 (b)	80.139 (a)	80.491 (b)	104.7658 (a)	106.93 (a)	104.0996 (a)	105.26 (a)
	400	85.611 (a)	82.676 (ab)	85.239 (a)	84.509 (ab)	104.5164 (a)	107.33 (a)	97.9616 (ab)	103.27 (a)
15	100	106.84 (a)	108.33 (a)	101.40 (a)	105.52 (b)	114.0468 (a)	109.24 (a)	114.4249 (b)	112.57 (b)
	200	109.06 (a)	110.89 (a)	110.79 (a)	110.24 (ab)	117.2177 (a)	115.23 (a)	118.8223 (a)	117.09 (a)
	400	114.86 (a)	113.83 (a)	114.01 (a)	114.23 (a)	118.8172 (a)	118.70 (a)	118.214 (ab)	118.57 (a)

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

Tabela 6b. Mapa consenso dentro de populações. Média aritmética do comprimento, em cM, de três grupos de ligação em três tipos de população com marcadores eqüidistantes a 5, 10 e 15 cM com/sem análise multiloco para construção dos mapas genéticos.

Saturação	*Tipo de População*	*Com análise multiloco*				*Sem análise multiloco*			
		Grupos de ligação			*Média Geral*	*Grupos de ligação*			*Média Geral*
		1	*2*	*3*		*1*	*2*	*3*	
5	F_2 Co-dominante	43.035 (a)	42.265 (a)	40.516 (ab)	41.939 (ab)	46.782 (a)	46.591 (a)	46.589 (a)	46.654 (b)
	F_2 dominante	41.924 (a)	40.121 (a)	39.570 (b)	40.462 (b)	50.121 (a)	49.767 (a)	50.396 (a)	50.274 (a)
	Retrocruzamento	43.699 (a)	40.305 (a)	44.573 (a)	42.863 (a)	47.974 (a)	49.727 (a)	48.605 (a)	48.902 (ab)
10	Co-dominante	83.328 (a)	80.668 (a)	83.028 (a)	82.398 (a)	86.392 (b)	86.093 (b)	86.253 (b)	86.246 (c)
	F_2 dominante	77.711 (a)	75.114 (a)	75.974 (b)	76.266 (b)	92.429 (ab)	93.969 (a)	92.768 (ab)	93.055 (b)
	Retrocruzamento	83.167 (a)	79.491 (a)	83.167 (a)	82.123 (a)	96.798 (a)	96.262 (a)	95.351 (a)	96.137 (a)
15	F_2 Co-dominante	112.92 (a)	114.45 (a)	112.64 (a)	113.34 (a)	118.45 (a)	116.95 (a)	120.05 (a)	118.48 (a)
	F_2 dominante	109.41 (a)	105.37 (b)	109.58 (a)	108.12 (b)	117.96 (a)	114.95 (a)	117.95 (ab)	116.96 (b)
	Retrocruzamento	111.42 (a)	111.50 (ab)	108.77 (a)	110.56 (ab)	115.88 (a)	115.77 (a)	115.88 (b)	115.84 (b)

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

4.2.3 Média das distâncias entre marcas adjacentes

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

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

O método de mapeamento proposto por Lander e Botstein (1989) emprega pares de marcadores vizinhos para obter o máximo de informação da ligação de QTLs, dentro do segmento de cromossômico analisado. Esses autores investigaram acurácia do método por simulação, e os resultados indicaram a existência de uma probabilidade razoável de detecção de QTL que explique pelo menos 5% da variância. Em média, um QTL com capacidade para explicar 5 ou 10% da variância é mapeado com um intervalo de 40 ou 20 cM, respectivamente. QTLs com maiores efeitos genotípicos serão localizados com maior precisão.

A distância média de marcas adjacentes dentro de cada grupo de ligação esperada nos grupos de ligação após a integração dos mapas das populações F_2 co-dominante, F_2 dominante e retrocruzamento era de 5 cM quando se integrou mapas gerados a partir do genoma original com saturação de 5cM, de 10 e 15 cM quando houve a integração de mapas gerados com 10 e 15 cM de saturação do genoma original.

Analisando o efeito do tamanho da população dentro do nível de saturação de 5 cM, para mapas consenso integrados dentro de populações (Tabela 7a), verificou-se que ocorre aproximação do valor de distância média entre as marcas adjacentes ao valor esperado de 5 cM à medida que aumenta o tamanho de população de 100 para 400 indivíduos. Uma explicação para esse resultado é a redução da amplitude da variação

dos valores de distância média obtidos entre as repetições. Esta redução da amplitude de variação dos valores de distância em torno da média pode ser observada pela redução dos valores de desvio padrão (Tabela 9a) à medida que se aumentou o número de indivíduos das populações segregantes Por exemplo, para a população de 100 indivíduos com análise multiloco o desvio padrão foi de 0,464 ao enquanto que para a de 400 indivíduos foi de 0,163. Exceção a este comportamento foi observada para as populações de 200 indivíduos em que a distância média entre os grupos de ligação diminuíram em relação à população de 100 indivíduos, porém o desvio padrão, decaiu.

Pelo teste de média para os tamanhos de população de 5 cM, as médias gerais das populações de 200 e 400 indivíduos se diferiram significativamente, ao passo os tamanhos de população de 100 e 200 indivíduos não diferiram significativamente pelo teste de Tukey a 5% de probabilidade.

Para os mapas consenso gerados a partir da saturação de 10 e 15 cM, como observado para os de 5 cM, houve uma tendência a aproximação do valor de distância média entre as marcas adjacentes ao valor esperado de 10 e 15 cM com o aumento do número de indivíduos da população de mapeamento de 100 para 400 indivíduos. Por exemplo para o grupo de ligação 1 dos mapas gerados com análise multiloco e saturação 10 cM foi obtida a distância média de 9,007 e 10,451, para os tamanhos de população de 100 e 400 indivíduos, respectivamente. Também os valores de desvio padrão geral diminuíram com o aumento do tamanho das populações (Tabela 9a). Por exemplo, para o mapa oriundo de saturação 15 cM, sem análise multiloco, o desvio padrão foi de 0,572 e 0,197 para os tamanhos populacionais de 100 e 400 indivíduos, respectivamente. Exceção a este padrão foi encontrada no nível de saturação de 10 cM, com análise multiloco para a construção dos mapas genéticos em que houve uma queda das distâncias médias entre as marcas.

É interessante observar que, com o aumento do tamanho da população, existe a tendência clara à diminuição da amplitude de variação nos valores das médias das distâncias entre marcas adjacentes, para os mapas gerados sem análise multiloco. Esta diminuição pôde ser observada tanto nos valores pertencentes ao mesmo grupo de ligação, quanto na média geral de cada tamanho populacional.

Visscher et al. (1996), utilizando técnicas de simulação, avaliaram a eficiência da introgressão de genes em populações de retrocruzamento derivadas de linha endogâmicas. Após a obtenção dos dados, ficou constatado que marcadores foram

eficientes tanto para introgressão de alelos quanto para seleção de genótipos receptores, simultaneamente. Ao serem consideradas marcas espaçadas de 10 e 20 cM, houve vantagem de uma a duas gerações de seleção de retrocruzamentos, em relação ao tempo gasto no melhoramento clássico, quanto à seleção fenotípica.

Considerando que o teste de médias ($P<0,05$) indicou diferença entre as médias gerais das populações de 100 e 400 indivíduos para os mapas gerados com análise multiloco com saturação 15 cM e sem análise multiloco para saturações de 10 e 15 cM, o teste não foi significativo entre os tamanhos populacionais para o nível de saturação de 10 cM com análise multiloco.

Para os mapas consenso integrados dentro de população (Tabela 7b), a distância média das marcas tende a ser menor para os mapas construídos com análise multiloco quando comparados com os mapas construídos sem análise multiloco, e que os valores das estimativas das distâncias médias entre as marcas adjacentes tende a se aproximar mais do valor original nos mapas construídos sem análise multiloco.

A realização do teste de Tukey ($P<0,05$) permitiu a avaliação estatística dos efeitos do tipo de população na integração de mapas genéticos. Para as populações F_2 co-dominates e F_2 dominante houve diferença significativa entre as médias gerais de quase todos os níveis de saturação estudados com e sem análise multiloco, com exceção para os mapas com saturação de 5 e 15 cM, quando foi utilizada análise multiloco e não foi utilizada análise multiloco

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

Tabela 7a. Consenso entre populações. Média das distâncias, em cM, entre marcas adjacentes nas repetições que recuperaram três grupos de ligação, em três tamanhos de população quando se trabalhou com marcadores eqüidistantes a 5, 10 e 15 cM com/sem análise multiloco para construção dos mapas genéticos.

Saturação	*Tamanho populacional*	*Com análise multiloco* — *Grupos de ligação* 1	2	3	*Média Geral*	*Sem análise multiloco* — *Grupos de ligação* 1	2	3	*Média Geral*
5	100	4,1448 (b)[1]	3,7530 (b)	4,3014 (b)	4,0664 (b)	4,70082 (a)[1]	4,9226 (a)	4,96146 (a)	4,8616 (ab)
	200	4,0328 (b)	3,7664 (b)	4,0624 (b)	3,9539 (b)	4,65862 (a)	4,91374 (a)	4,72086 (a)	4,7644 (b)
	400	5,239 (a)	5,0715 (a)	5,1778 (a)	5,1627 (a)	5,2669 (a)	5,0490 (a)	5,21562 (a)	5,1771 (a)
10	100	8,8611 (a)[1]	9,2005 (a)	8,3426 (a)	8,8014 (a)	9,00738 (b)	8,8004 (b)	8,70158 (b)	8,8364 (b)
	200	8,2604 (a)	7,8732 (a)	8,0138 (a)	8,0491 (a)	10,47656 (a)	10,693 (a)	10,40996 (a)	10,526 (a)
	400	8,5611 (a)	8,2676 (a)	8,5239 (a)	8,4509 (a)	10,45164 (a)[1]	10,733 (a)	9,79614 (ab)	10,327 (a)
15	100	11,404 (a)[1]	10,924 (a)	11,442 (b)	11,257 (b)	10,68436 (a)[1]	10,833 (a)	10,14072 (a)	10,552 (b)
	200	11,721 (a)	11,523 (a)	11,882 (a)	11,709 (a)	10,90644 (a)	11,089 (a)	11,07916 (a)	11,024 (ab)
	400	11,881 (a)	11,870 (a)	11,821 (ab)	11,857 (a)	11,48622 (a)	11,383 (a)	11,40104 (a)	11,423 (a)

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

Tabela 7b. Consenso entre populações. Média das distâncias, em cM, entre marcas adjacentes nas repetições que recuperaram três grupos de ligação, em três tamanhos de população quando se trabalhou com marcadores eqüidistantes a 5, 10 e 15 cM com/sem análise multiloco para construção dos mapas genéticos.

Saturação	*Tipo de População*	*Com análise multiloco* — *Grupos de ligação* 1	2	3	*Média Geral*	*Sem análise multiloco* — *Grupos de ligação* 1	2	3	*Média Geral*
5	F_2 Co-dominante	4.303 (a)	4.226 (a)	4.051 (ab)	4.193 (ab)	4.6782 (a)	4.6591 (a)	4.6589 (a)	4.6654 (b)
	F_2 dominante	4.192 (a)	4.0121 (a)	3.934 (b)	4.046 (b)	5.0121 (a)	5.0307 (a)	5.0396 (a)	5.0275 (a)
	Retrocruzamento	4.369 (a)	4.030 (a)	4.459 (a)	4.286 (a)	4.8374 (a)	4.9727 (a)	4.8604 (a)	4.8902 (ab)
10	F_2 Co-dominante	8.333 (a)	8.083 (a)	8.303 (a)	8.2398 (a)	8.5962 (b)	8.6733 (b)	8.6042 (b)	8.6246 (c)
	F_2 Dominante	7.7711 (a)	7.511 (a)	7.597 (b)	7.6266 (b)	9.2429 (ab)	9.3969 (a)	9.2768 (ab)	9.3055 (b)
	Retrocruzamento	8.316 (a)	7.971 (a)	8.349 (a)	8.2123 (a)	9.6798 (a)	9.6262 (a)	9.5351 (a)	9.6137 (a)
15	F_2 Co-dominante	11.292 (a)	11.445 (a)	11.264 (a)	11.334 (a)	11.745 (a)	11.915 (a)	11.883 (a)	11.848 (a)
	F_2 dominante	10.941 (a)	10.537 (b)	10.958 (a)	10.812 (b)	11.796 (a)	11.495 (a)	11.795 (ab)	11.696 (ab)
	Retrocruzamento	11.142 (a)	11.150 (ab)	10.877 (a)	11.056 (ab)	11.588 (a)	11.577 (a)	11.588 (b)	11.584 (b)

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

4.2.4 Estresse

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. Para obtenção dos valores médios de estresse, foi feita uma média aritmética dos valores de estresse obtidos nas repetições em que houve a integração dos 3 grupos de ligação. Os valores obtidos são apresentados nas tabelas 8a e 8b, juntamente com os resultados das comparações feitas entre as médias (Tukey a 5%).

Vale salientar que os valores de estresse apresentam significados diferentes em níveis de saturação diferentes, ou seja, devem ser comparados níveis de estresse entre genomas com o mesmo nível de saturação.

Analisando o efeito da integração de mapas dentro de populações (Tabela 8a) do nível de saturação de 5 cM nota-se uma redução dos valores de estresse médio a medida que aumenta o tamanho da população segregante. Isto foi observado para todos os grupos de ligação com exceção para os grupos de ligação um e dois construídos com análise multiloco para os tamanhos populacionais de 100 e 200 indivíduos, em que houve um aumento do estresse com o tamanho da população de 100 para 200. Porém, os valores de média geral de estresses decaíram com aumento da população, com estimativas de estresse médio de 30,616 e 17,332 observadas para os tamanhos populacionais de 100 e 400 indivíduos respectivamente, nos mapas construídos com análise multiloco.

Para os níveis de saturação de 10 e 15 cM verificou-se, assim como para o nível de saturação de 5 cM, uma redução nos valores de estresse médio a medida que se aumenta o tamanho das populações segregantes para todos os grupos de ligação. Por exemplo, o grupo de ligação 3, para mapas construídos sem análise multiloco com saturação 10 cM, tem-se que o estresse médio foi de 23,225 e 14,335 para os tamanho populacionais com 100 e 400 indivíduos, respectivamente. Além da redução dos valores de estresse médio, também ocorreu uma redução na amplitude de variação dos valores de estresse entre repetições à medida que se aumentou o tamanho da população segregante, o que é representado pela redução dos valores de desvio padrão (Tabela 9a). Esse resultado demonstra uma maior confiabilidade no mapeamento de populações com maior número de indivíduos. Por exemplo, o desvio padrão geral para os mapas

consenso com saturação de 15 cM construídos com análise multiloco foi de 8,120 e 0,915 para os tamanhos de população de 100 e 400 indivíduos.

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 obtidas com a utilização do teste de comparação das médias gerais dos grupos de ligação, onde os maiores valores associados às menores populações, diferem significativamente das maiores médias associadas as menores populações.

Para os mapas consenso integrados dentro de populações (Tabela 8b) observa-se que o estresse foi menor em populações F_2 co-dominante quando comparados com os demais tipos de população, apesar do teste de Tukey não ter se mostrado significativo para os diferentes tipos de populações.

Quanto ao nível de saturação de 5 cM os testes de média entre os grupos de ligação não se mostraram significativos, tanto para mapas construídos com e sem análise multiloco. Porém, na média geral para os mapas sem análise multiloco a população F_2 co-dominante se diferiu estatisticamente da população de retrocruzamento. Neste nível, a variação foi de 17,123 (b) a 24,866 (a), nas populações F_2 co-dominante e de retrocruzamento.

Analisando os tamanhos de população de 10 e 15 cM observa-se que não há diferenças significativas dos testes de médias entre os grupos de ligação das diferentes populações. Por exemplo, para mapas construídos com análise multiloco na saturação de 10 cM, o estresse médio foi de 12,977 (a), 17,428 (a) e 18,569 (a), nas populações F_2 co-dominante, F_2 dominante e de retrocruzamento, respectivamente. Diferença significativa no teste de Tukey a 5% somente foi identificada para a média geral dos mapas gerados sem análise multiloco e saturação de 10 cM.

Pelos testes estatísticos, podemos inferir que os estresses gerados pela integração de mapas consenso dentro de populações de diferentes tamanhos, não diferem significativamente entre os grupos de ligação das diferentes populações analisadas, dentro de cada nível de saturação estudado.

Tabela 8a. Mapa consenso entre populações.Valores do estresse médio (%) em função do tamanho da população em populações quando se trabalhou com marcadores eqüidistantes a 5, 10 e 15 cM com/sem análise multiloco para construção dos mapas genéticos.

Saturação	*Tamanho populacional*	*Com análise multiloco*				*Sem análise multiloco*			
		Grupos de ligação			*Média Geral*	*Grupos de ligação*			*Média Geral*
		1	*2*	*3*		*1*	*2*	*3*	
5	100	29,109 (a)	31,164 (ab)	31,576 (a)	30,616 (a)	26,84676 (a)	25,164 (a)	30,42114 (a)	27,477 (a)
	200	29,441 (a)	33,716 (a)	23,151 (ab)	28,769 (a)	22,56018 (ab)	22,503 (ab)	20,14644 (ab)	21,736 (b)
	400	16,169 (b)	17,893 (b)	17,933 (b)	17,332 (b)	17,00442 (b)	18,319 (b)	17,87746 (b)	17,733 (b)
10	100	29,853 (a)	26,202 (a)	40,330 (a)	32,128 (a)	24,7922 (a)	21,492 (a)	23,39118 (a)	23,225 (a)
	200	20,168 (a)	14,135 (a)	15,816 (b)	16,706 (b)	20,2913 (a)	16,736 (b)	16,98234 (a)	18,003 (a)
	400	20,629 (a)	13,605 (a)	18,957 (b)	17,730 (b)	14,04318 (b)	14,617 (a)	14,34616 (a)	14,335 (b)
15	100	24,744 (a)	21,678 (a)	24,850 (a)	23,757 (a)	26,98212 (a)	32,197 (a)	27,82516 (a)	29,001 (a)
	200	18,770 (a)	19,074 (a)	19,066 (a)	18,97 (b)	24,5673 (a)	26,695 (ab)	23,5768 (b)	24,946 (b)
	400	15,680 (a)	16,392 (a)	15,673 (a)	15,915 (b)	22,8167 (a)	24,502 (b)	23,5061 (b)	23,608 (b)

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

Tabela 8b. Mapa consenso dentro populações.Valores do estresse médio (%) em função do tamanho da população em populações quando se trabalhou com marcadores eqüidistantes a 5, 10 e 15 cM com/sem análise multiloco para construção dos mapas genéticos.

Saturação	*Tipo de População*	*Com análise multiloco*				*Sem análise multiloco*			
		Grupos de ligação			*Média Geral*	*Grupos de ligação*			*Média Geral*
		1	*2*	*3*		*1*	*2*	*3*	
5	F_2 Co-dominante	20.265 (a)	16.540 (a)	19.134 (a)	18.646 (a)	16.448 (a)	14.918 (a)	20.003 (a)	17.123 (b)
	F_2 dominante	21.081 (a)	27.771 (a)	25.087 (a)	24.646 (a)	19.590 (a)	22.230 (a)	17.242 (a)	19.687 (ab)
	Retrocruzamento	18.743 (a)	20.660 (a)	27.585 (a)	22.329 (a)	24.028 (a)	24.052 (a)	26.520 (a)	24.866 (a)
10	F_2 Co-dominante	12.977 (a)	16.727 (a)	16.202 (a)	15.302 (a)	15.623 (a)	14.255 (a)	12.931 (a	14.270 (ab)
	F_2 Dominante	17.428 (a)	20.303 (a)	20.800 (a)	19.510 (a)	16.929 (a)	14.612 (a)	16.343 (a)	15.961 (a)
	Retro cruzamento	18.569 (a)	21.057 (a)	18.755 (a)	19.460 (a)	14.291 (a)	12.089 (a)	13.333 (a)	13.238 (b)
15	F_2 Co-dominante	16.374 (a)	16.233 (a)	16.016 (a)	16.208 (b)	24.374 (a)	23.875 (a)	23.554 (a)	23.934 (a)
	F_2 dominante	18.308 (a)	22.072 (a)	18.816 (a)	19.732 (a)	23.545 (a)	28.088 (a)	24.206 (a)	25.280 (a)
	Retrocruzamento	20.098 (a)	18.118 (a)	19.873 (a)	19.363 (a)	25.010 (a)	26.649 (a)	25.296 (a)	25.652 (a)

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

4.2.5 Desvio padrão e variância entre marcas adjacentes

Uma forma adicional de observar o comportamento dos comprimentos médios dos grupos de ligação, média das distâncias entre as marcas adjacentes e os valores de estresse médio é através da análise do desvio-padrão. Na tabela 9a são apresentados os desvios padrões gerais do comprimento médio dos grupos de ligação (*), média das distâncias entre marcas adjacentes (**) e estresse médio (***). Com o aumento no número de indivíduos em um mesmo nível de saturação do genoma espera-se que o desvio-padrão diminua. Desta forma, torna-se evidente a tendência de redução na amplitude de variação observada nas médias com o aumento do tamanho populacional.

Para os mapas consenso entre populações (Tabela 9a), o desvio padrão do comprimento dos grupos de ligação e da média das distâncias entre marcas adjacentes observa-se que o desvio padrão decaiu com o aumento da população para os 2 tipos de mapas integrados, a única exceção foi para a saturação de 5 cM e tamanho populacional de 200 indivíduos para os mapas construídos sem análise multiloco o qual apresentou desvio padrão maior em relação à população de 100 indivíduos.

O desvio padrão geral do estresse também decaiu com o aumento da população, com exceção para a saturação de 5 cM e mapas construídos com análise multiloco e tamanho populacional 200, que se mostrou maior que desvio padrão do tamanho populacional 100 indivíduos.

A partir das distâncias entre marcas adjacentes obtidas nos grupos de ligação foi estimada a variância amostral. Como em quadros anteriores, para cada grupo de ligação são dadas médias aritméticas, nesse caso, são obtidas variânicas em cada repetição, em que houve a formação de três grupos de ligação no mapeamento das populações completamente informativas simuladas.

Os valores de variância encontrados na análise dos mapas obtidos da integração de mapas das populações simuladas são referentes aos erros do genoma com saturação e tamanho populacional dos quais eles foram comparados “genoma ideais”, pois os genomas utilizados para geração das populações segregantes tinham seus marcadores distribuídos de forma eqüidistante dentro dos três grupos de ligação.

Os valores de variância podem ser interpretados de forma que quanto menores os valores de variância mais eqüidistantes estarão distribuídas às marcas dentro dos grupos

de ligação e, conseqüentemente, menor o erro. Portanto, quanto menores os valores de variância, mais próximos estarão os valores do esperado, indicando uma boa recuperação do genoma com o mapeamento das populações segregantes ou uma menor tensão entre a ordem das marcas.

Analisando o efeito do tamanho da população, verificou-se redução da média da variância em relação às distâncias entre marcas adjacentes. Em relação à confiabilidade, quanto menores os valores de variância dos valores entre marcas adjacentes, maior a confiabilidade dos dados, sendo assim, com o aumento do tamanho populacional temos maior confiabilidade dos dados.

Além da redução na média das variâncias, observou-se redução na amplitude de variação dos valores de variância dos diferentes tamanhos populacionais, à medida que aumentou o tamanho da população.

Para mapas consenso dentro de populações (Tabela 9b), os valores de variância foram menores para populações F_2 co-dominantes nos 3 níveis de saturação analisados, quando comparados com os valores de populações F_2 dominante e de retrocruzamento.

Quanto ao desvio padrão, ele foi menor nos mapas simulados sem análise multiloco quando comparado nos mapas simulados com análise multiloco nos níveis de saturação de 10 e 15 cM. Caso contrário ocorreu para saturação de 5 cM, em que apenas a população F_2 dominante manteve este padrão.

Tabela 9a. Mapa consenso dentro populações.Valores de desvio padrão em função do tamanho da população, média das distâncias entre marcas adjacentes, estresse médio e variância em populações quando se trabalhou com marcadores eqüidistantes a 5, 10 e 15 cM com/sem análise multiloco, para construção dos mapas genéticos.

		Com análise multiloco				*Sem análise multiloco*			
Saturação	*Tipo de População*	Desvio Padrão *	Desvio Padrão **	Desvio Padrão ***	*Variância*	Desvio Padrão *	Desvio Padrão **	Desvio Padrão ***	Variância
5	100	4.642	0.464	6.687	1.312	3.415	0.341	6.633	1.798
	200	3.321	0.332	9.077	1.025	5.319	0.531	5.270	0.811
	400	1.633	0.163	2.845	0.677	1.640	0.164	2.474	0.694
10	100	15.707	1.570	14.707	3.864	9.257	0.925	6.453	4.479
	200	4.813	0.481	4.828	1.177	7.770	0.777	4.911	1.941
	400	3.916	0.391	4.070	0.746	4.353	0.435	4.019	0.630
15	100	10.273	1.0273	8.120	3.197	5.723	0.572	4.480	3.047
	200	4.624	0.462	2.809	1.144	3.384	0.338	2.422	0.797
	400	1.801	0.180	1.932	0.915	1.979	0.197	1.404	0.551

Desvio padrão do comprimento médio dos grupos de ligação (*)
Desvio padrão da média das distâncias entre marcas adjacentes (**)
Desvio padrão estresse médio (***)

Tabela 9b. Mapa consenso dentro populações.Valores de desvio padrão em função do tamanho da população, média das distâncias entre marcas adjacentes e estresse médio e variância em populações quando se trabalhou com marcadores eqüidistantes a 5, 10 e 15 cM com/sem análise multiloco, para construção dos mapas genéticos.

		Com análise multiloco				*Sem análise multiloco*			
Saturação	*Tipo de População*	Desvio Padrão *	Desvio Padrão **	Desvio Padrão ***	*Variância*	Desvio Padrão *	Desvio Padrão **	Desvio Padrão ***	Variância
5	F_2 Co-dominante	2.1148	0.2114	4.3085	0.2995	3.7788	0.3779	4.5305	0.3943
	F_2 Dominante	2.8877	0.2887	8.9998	1.0891	2.1651	0.2165	5.3032	0.7182
	Retrocruzamento	2.8877	0.2683	6.3779	0.7845	5.2989	0.5298	7.6075	1.2791
10	F_2 Co-dominante	4.8494	0.4849	5.0489	0.6052	3.8451	0.3845	3.7909	0.7224
	F_2 Dominante	4.4200	0.4420	11.094	2.8776	2.2737	0.2273	2.6315	1.7424
	Retrocruzamento	4.4618	0.4462	4.8170	1.6091	3.8123	0.3812	2.3247	1.3456
15	F_2 Co-dominante	3.1061	0.3106	2.4014	0.8361	2.7822	0.2782	2.2996	0.8121
	F_2 dominante	4.8059	0.4805	4.0111	1.6669	3.2567	0.3256	3.3002	1.1080
	Retrocruzamento	3.4973	0.3497	3.1058	1.7179	1.6124	0.1612	1.2986	0.8122

Desvio padrão do comprimento médio dos grupos de ligação (*)
Desvio padrão da média das distâncias entre marcas adjacentes (**)
Desvio padrão estresse médio (***)

4.3 ASPECTOS TEÓRICOS DA ANÁLISE MULTILOCO

Para obtenção dos mapas consenso pela metodologia proposta fez-se uso da análise multiloco, e para seu entendimento, alguns questionamentos devem ser formulados:

- A reutilização da análise multiloco para integração dos mapas poderia estar influenciando nas estimativas de distâncias entre as marcas.
- Existe diferença da integração entre mapas construídos com e sem análise multiloco.
- Como explicar a ocorrência de um resultado inesperado: distância negativa em um mapa integrado.

Mapas genéticos são ferramentas úteis em vários campos da pesquisa genética, tanto no aspecto da pesquisa básica quanto na aplicada. A construção de um mapa genético é essencialmente a procura de um arranjo linear de marcadores a partir de valores de recombinação. A distância de mapa entre dois marcadores é definida como número significativo de eventos de recombinação, envolvendo uma dada cromátide, em uma dada região por meiose, é expressa por centimorgans (cM). A relação entre distância de mapa e freqüência de recombinação (e vice-versa) é expressa por uma função de mapeamento genético. Diferentes funções de mapeamento correspondem a diferentes graus de interferência assumidos no *crossing-over* (Stam,1993). Para a construção de um mapa genético com 100 marcadores, por exemplo, não são necessárias apenas 99 estimativas de distâncias, apesar de que apenas estes valores serão descritos nas representações gráficas dos grupos de ligações. Todas as estimativas, num total de n(n-1)/2 (igual a 4950, quando n é igual a 100), devem contribuir para melhor representar a distância entre os pares de marcas e ratificar o ordenamento estabelecido. A incorporação destas estimativas é normalmente feita por procedimento denominado de análise multiloco.

Modelos de multilocos diferem de modelos dois-locos porque toda a informação dos intervalos entre dois locos de um grupo de ligação é considerada. Por exemplo, informações sobre intervalos entre os locos A e B e entre B e C serão incluídas para a estimação de todas as distâncias do mapa entre A e C se a ordem dos locos é ABC.

Em análise de ligação multilocos é comum assumir que a interferência é nula, possibilitando maior tratabilidade matemática à análise de mapeamento, porém existem fortes evidências biológicas de que, na maioria dos casos a interferência se manifesta, principalmente, quando os locos estão pouco espaçados. Uma aproximação para modelo que pressupõe interferência não nula é aquela em que a porcentagem de recombinação é expressa por uma função de mapeamento particular, como a de Kosambi. Entretanto, deve-se ter em mente que as várias funções de mapeamento não pressupõem a porcentagem de recombinação conjunta, ou seja, não são probabilidades de multilocos, para mais de três locos (Goldstein et al, 1995).

Funções de mapeamento são importantes ferramentas em um modelo multiloco útil. Entretanto, a adequação das funções de mapeamento depende se o que se assumiu fundamentalmente na função de mapeamento é verdadeiro ou falso. Muitas tentativas têm sido desenvolvidas para a construção de um modelo multilocos (Liu, 1998).

Deve ser entendido que as funções de mapeamento são fórmulas que expressam a porcentagem de recombinação entre duas marcas como função da distância entre elas. Elas têm sido comumente utilizadas para construção de mapas genéticos. Como as porcentagens de recombinação não são aditivas as funções de mapeamento são utilizadas para resolver esse problema.

A primeira função de mapeamento foi proposta por Haldane (1919), desde então muitas outras têm sido propostas e discutidas.

Sturtevant (1913) e Morgam (1928), em seus primeiros trabalhos com mapeamento genético em *Drosophila* utilizaram a porcentagem de recombinação como estimativa da distância no mapa. Para um pequeno segmento do genoma é considerado que a chance de que ocorra um duplo ou múltiplos *crossing-overs* é baixa. Neste caso a porcentagem de recombinação estimada tem a mesma expectativa que o número esperado de *crossing-overs*. Isto quer dizer que esta função de mapeamento só pode ser utilizada quando um pequeno segmento do genoma é analisado. Quando um segmento

maior é analisado o número esperado de múltiplos *crossing-overs* aumenta e a função de Morgan não é apropriadamente aplicada.

A função de mapeamento de Haldane é baseada no fato de que não há interferência e que os eventos de *crossing-over* possuem uma distribuição de Poisson. A derivação original da função de Haldane foi puramente matemática. Contudo essa função tem sido muito usada e estudada nas últimas décadas sendo que derivações mais biológicas tem sido desenvolvidas.

Quando a porcentagem de recombinação é pequena, à distância no mapa utilizando tanto a função de Haldane como a porcentagem de recombinação é aproximadamente igual. Quando o tamanho do segmento aumenta o número de múltiplos *crossing-over* aumenta e as distâncias no mapa são ajustadas para múltiplos *crossing-over* através da função de mapeamento de Haldane.

A função de mapeamento de Kosambi é uma extensão da função de Haldane. Ela é baseada em observações empíricas, mostrando que a interferência existe e crossing-over não ocorrem randomicamente.

A base da função de mapeamento de Kosambi é que a interferência depende do tamanho do genoma. A interferência é ausente quando o segmento é suficientemente grande (ex. c =1 quando r = 0,5). A interferência aumenta quando o comprimento do segmento decresce (ex. c = 0 quando r = 0,0).

A função de Kosambi fornece uma medida de distância genética que geralmente ajusta dados melhor que a função de Haldane, porque ela leva a interferência em conta na estimação. Por isso ela é mais popular na literatura.

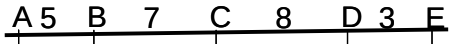
Para obtenção de mapas genético, mapas consenso e mapa integrado é indispensável à utilização da análise multiloco. Os efeitos que esta análise provoca devem ser melhor esclarecidos para que sua adoção seja feita de forma criteriosa. Informações sobre esta técnica é escassa na literatura e merece atenção especial no contexto deste trabalho.

Deve ser lembrado que no mapeamento genético são necessárias algumas etapas básicas tais como:

- Teste de segregação dos marcadores individuais.
- Cálculo da porcentagem de recombinação entre pares de marcadores, obtendo uma matriz D_{mxm}.

- Agrupamento dos marcadores baseado em critério de LOD mínimo e Recombinação máxima.
- Ordenamento, com base em algum critério, geralmente de menor sarf.

Assim, após todas as etapas de mapeamento concluídas, é comum apresentar, por exemplo, um mapa hipotético com o seguinte grupo de ligação



Construído o mapa com 5 marcadores, existirão dez estimativas de distância entre os pares de marcadores, sendo apresentadas no mapa apenas quatro delas. De maneira geral, com m marcadores tem-se m(m-1)/2 estimativas das quais m-1 são utilizadas na representação gráfica. Destaca-se o fato de que o valor de r_{BC} é exatamente aquele obtido entre pares de marcadores e se encontra na matriz D, ou seja, nenhuma outra informação contribui para a distância gráfica entre B e C. Entretanto, valores de (r_{AC}-r_{AB}) ou de (r_{BD}-r_{CD}), sob certas pressuposições, também fornecem valor da medida r_{BC} e, portanto, pode-se questionar qual seria a melhor estimativa para r_{BC}.

Para definir a melhor estimativa deve-se considerar alguns aspectos, tais como:

a) Aspecto biológico

Este é definido em função da existência de um fenômeno denominado de interferência. Na hipótese de que este fenômeno aconteça, tem-se que

$r_{AC} = r_{AB} + r_{BC} - 2c\, r_{AB}\, r_{BC}$, em que c é o coeficiente de coincidência

De maneira geral, tem-se que:

$$r_{BC} = \frac{r_{AC} - r_{AB}}{1 - 2c\, r_{AB}}$$

Desta forma, pode-se entender que $r_{BC} = r_{AC} - r_{AB}$ somente se aplica nos casos em que c=0, ou seja, há interferência total nas regiões consideradas.

Percebe-se que várias outras estimativas de r_{BC} poderão ser obtidas, mas dependerão fundamentalmente da pressuposição do valor da interferência. Assim, o pesquisador deve estar preocupado em como incorporar estas estimativas num processo de análise multiloco, uma vez que o valor da interferência é desconhecido.

b) Aspecto estatístico

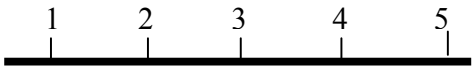
Outro aspecto a ser considerado é o fato de que as estimativas do valor de r são obtidas a partir de populações de tamanho relativamente pequeno, de forma que o valor obtido poderá estar associado a um grande desvio padrão, dependendo além do tamanho da população, do tipo de marcador, do tipo de população e até mesmo da fase de ligação. Assim, mesmo nos casos em que a interferência seja total, é possível não ter $r_{AC} = r_{AB} + r_{BC}$, pois os erros associados a cada estimativa poderá provocar a perda da aditividade.

Assim, pode-se questionar qual seria a melhor maneira de estimar r_{BC}. Levar apenas em consideração as informações de A e B ou incorporarmos as informações dos demais locos, sabendo dos problemas inerentes à precisão e o desconhecimento da interferência.

4.3.1. **Metodologia de análise multiloco**

Na literatura há duas metodologias citadas, a de mínimos quadrados e da máxima verossimilhança, para efetuar a análise multiloco. Nestas metodologias é considerada uma medida D de discrepância entre as distâncias observadas e seus valores esperados, dada por:

$$D= (m_1-r_{12})^2 + (m_1+m_2- r_{13})^2 + \ldots \ldots + (m_4-r_{45})^2$$



Distânicas estimadas	r_{12}	r_{23}	r_{34}	r_{45}
Distância paramétrica	m_1	m_2	m_3	m_4

Como os valores de m_j não são igualmente precisos, os termos da soma de D devem receber pesos apropriados. Na metodologia baseada em mínimos quadrados, o peso a ser utilizado é o valor de LOD. Na metodologia de máxima verossimilhança, o peso recomendado é a inversa da variância de r_{ij}.

Neste momento é necessária uma abordagem sobre o uso de pesos, pois como já foi tratado anteriormente, em uma análise multiloco, deve-se pressupor a existência, ou não, da interferência e utilizar uma função de mapeamento no processo de estimação de valores de m_i que melhor se ajustam na função D, por um critério de minimização (mínimos ponderados) ou maximização (máxima verossimilhança).

Para uma população F_2, tomada como ilustração, e marcadores co-dominantes, têm-se que o conteúdo médio de informação é dado por:

$$C(r) = \frac{2(1 - 3r + 3r^2)}{r(1-r)(1-2r+2r^2)}$$

O índice de informatividade de Fisher é, portanto:

$$I(r) = Nc(r)$$

e a variância dada por:

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

Se os valores de r_{ij} são transformados por uma função de mapeamento, tem-se novas estimativas de variância, dada por :

a) Para a transformação pela função de Haldane, em que :

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

Tem-se que:

$$V(h) = \frac{1}{(1-2r)^2}v(r)$$

b) Para a transformação pela função de Kosambi, em que:

$$k = -\frac{1}{4}\ln\left[\frac{1-2r}{1+2r}\right]$$

Tem-se

$$V(k) = \frac{1}{(1-4r)^2}V(r)$$

Assim, um quadro comparativo de valores de peso, considerando uma população F_2, de tamanho 100, e marcadores co-dominantes, é apresentado a seguir:

Tabela 10 : Valores de V(r) em função de Morgan, Haldane e Kosambi

função	r = 0,02	r = 0,05	r = 0,10	r =0,15	r =0,20	r =0,40
Morgan	5,88	14,74	29,73	45,24	61,53	146,88
Haldane	6,38	18,20	46,46	92,33	170,94	367206
Kosambi	5,90	15,04	32,26	54,63	87,21	93665

* valores de V(r), devem ser multiplicados por 10^{-5}

Verifica-se, portanto, que pares de marcadores com pequenos valores de distância (abaixo de 5 cM), os pesos estabelecidos pelas diferentes funções são

equivalentes. Mas, para valores acima de 10, função como a de Haldane pode ponderar de forma bem diferenciada daquela obtida por Kosambi e Morgan. Deve ser lembrado que, apesar dos pesos serem diferentes, os valores de distância são bem próximos da faixa de 0 a 20, como é ilustrado a seguir:

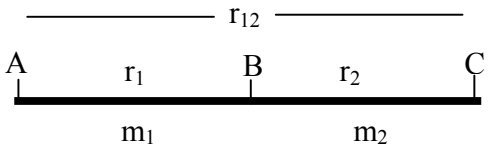
Tabela 11: transformação dos valores de r (Morgan)

Função	*0,02*	*0,05*	*0,10*	*0,15*	*0,20*	*0,49*
Haldane	0,02	0,053	0,112	0,178	0,255	1,956
Kosambi	0,02	0,05	0,101	0,155	0,212	1,149

A seguir serão tratados alguns casos para ilustrar o efeito da ponderação nas análises multiloco

Caso1- Três marcadores e Interferência total

Como ilustração é considerado o esquema:



Será inicialmente, admitido:

Sendo: $r_1 = r_2 = r$ $\quad r_{12} = 2r$
$p_1 = p_2 = p$ $\quad p_{12} = \theta$

De forma que:

$D = p_1(m_1 - r_1)^2 + \theta(m_1 - m_2 - r_{12})^2 + p_2(m_2 - r_2)^2$

de onde resulta o sistema de equações:

$$Pm=Q$$

$$\begin{pmatrix} p+\theta & \theta \\ \theta & p+\theta \end{pmatrix} \begin{pmatrix} m_1 \\ m_2 \end{pmatrix} = \begin{pmatrix} r\,(p+2\,\theta) \\ r\,(p+2\,\theta) \end{pmatrix}$$

e a solução

$m_1 = m_2 = r$

Assim, por exemplo, se $r_1 = 5$, $r_2 = 5$ e $r_{12} = 10$ tem-se os resultados

a) usando Morgan

$m_1 = m_2 = 5$, ou seja, a análise multiloco confirma os valores de distância entre pares de marcadores

b) Usando Haldane

$m_1 = 5{,}48$
$m_2 = 5{,}48$

Transformando os valores para porcentagem de recombinação, tem-se:

$r = 0{,}5\,[\,1 - e^{-2n}\,]$

tem-se

$m_1 = 5{,}19$
$m_2 = 5{,}19$

Note que os valores obtidos na análise multiloco são mais elevados. A distorção foi provocada pelo fato de se admitir a interferência nula, numa situação em que a interferência era total.

c) Usando Kosambi

$m_{k1} = 5{,}05$
$m_{k2} = 5{,}05$

Transformando os valores para porcentagem de recombinação, tem-se:

$$r = 0{,}5 \left[\frac{e^{4k} - 1}{e^{4k} + 1} \right]$$

$m_1 = 5{,}03$
$m_2 = 5{,}03$

Os valores na análise multiloco são mais elevados e a distorção é provocada pelo fato de se admitir que a coincidência seja igual a $2r_{12}$, numa situação em a coincidência era nula ou interferência total.

Veja que ao pressupor a existência de interferência o valor $h_{12} \neq h_1 + h_2$ e $k_{12} \neq k_1 + k_2$, alterando as estimativas de m_1 e m_2.

De forma generalizada, tem-se para $r_1 = r$

$$r_2 = kr \quad \text{e} \quad r_{12} = (k+1)r$$

e independente de p_1, p_2 e p_{12}, se verifica

$m_1 = r$

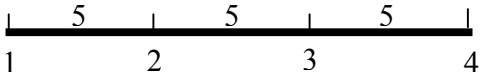
$m_2 = kr$

Caso 2- Vários marcadores com interferência total

Como ilustração, seja a matriz de distância dada a seguir obtida de uma população F_2 de tamanho 200, com marcadores co-dominantes.

$$D = \begin{pmatrix} 0 & 5 & 10 & 15 \\ & 0 & 5 & 10 \\ & & 0 & 5 \\ & & & 0 \end{pmatrix}$$

O agrupamento e ordenamento serão dados por



Após análise multiloco, por meio da metodologia baseada na máxima verossimilhança são, obtidos os resultados:

a) Usando a porcentagem de recombinação

Mapa Original: [1]—5,0—[2]— 5,0—[3]— 5,0—[4]

Após análise multiloco:

[1]— 5,0— [2]— 5,0 — [3]— 5,0— [4]

b) Usando a função de Haldane

Após análise multiloco: [1] —5,53 —[2]— 5,56 —[3]— 5,53— [4]

Convertendo em porcentagem de recombinação, tem-se:

[1] —5,24 —[2]— 5,26 —[3]— 5,24— [4]

Novamente são encontrados valores alterados em relação aos originais

c) Usando a função de Kosambi

Após análise multiloco

[1] —5,08 —[2]— 5,07 —[3]— 5,08— [4]

Convertendo em porcentagem de recombinação tem-se:

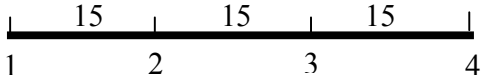
[1] —5,06 —[2]— 5,05 —[3]— 5,06— [4]

Nota-se que variações são geradas pela análise multiloco em razão do estabelecimento de uma função de mapeamento não apropriada.

Em um outro exemplo, em que o mapa será menos saturado, pode-se considerar:

$$D = \begin{pmatrix} 0 & 15 & 30 & 45 \\ & 0 & 15 & 30 \\ & & 0 & 15 \\ & & & 0 \end{pmatrix}$$

O agrupamento e ordenamento, numa análise multiloco proporciona o seguinte resultado:



Após análise multiloco, tem-se:

a) Usando a porcentagem de recombinação

[1]─15─[2]─15─[3]─15─[4]

b) Usando a função de Haldane

Em Haldane:

[1]─19,19─[2]─20,17─[3]─19,19─[4]

e, em porcentagem de recombinação:

[1]─15,93─[2]─16,6─[3]─15,46─[4]

Neste exemplo verifica-se uma alteração mais pronunciada dos valores encontrados pela análise multiloco que pressupôs interferência nula.

c) Usando a função Kosambi

Em Kosambi:

[1]─16,23─[2]─16,64─[3]─16,23─[4]

e, em porcentagem de recombinação:

[1]─15,68─[2]─16,05─[3]─15,68─[4]

Observa-se que considerando interferência total ou seja c=0, dependendo da função que se utiliza no processo de mapeamento, a distância entre as marcas pode ser viesada e que este viés tende a ser maior em mapas menos saturados. A função de Haldane, nos exemplos considerados, proporcionou maiores discrepâncias.

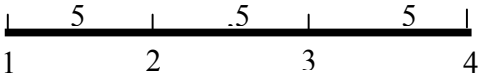
Caso 3 – Vários marcadores – Interferência Nula

Nas situações a seguir serão considerados casos que a interferência é nula e, portanto, a melhor opção de análise multiloco seria, com o uso da função de Haldane. As discrepâncias geradas por não se adotar esta função de mapeamento poderá ser observada pelo leitor.

Como ilustração, seja a matriz:

$$D = \begin{pmatrix} 0 & 5 & 9,5 & 13,55 \\ & 0 & 5 & 9,5 \\ & & 0 & 5 \\ & & & 0 \end{pmatrix}$$

O agrupamento e ordenamento, proporcionariam o seguinte resultado:



Se for considerada a análise multiloco, baseada no método da máxima verossimilhança, seriam obtidos os seguintes resultados:

a) Baseado em porcentagem de recombinação

Mapa original:

[1]—5,0—[2]— 5,0—[3]— 5,0—[4]

Após análise multiloco:

[1]—4,51—[2]— 4,82—[3]—4,51— [4]

Ao assumir interferência total numa situação que ela é nula proporcionaria uma subestimação diferenciada nos valores de distância

b) Baseado em Haldane:

Mapa Original: [1]—5,0—[2]— 5,0—[3]—5,0—[4]

Mapa em Haldane: [1]—5,27—[2]— 5,27—[3]—5,27— [4]

Mapa em Porcentagem de Recombinação: [1]—5,0—[2]— 5,0—[3]—5,0—[4]

O mapa gerado com função de mapeamento adequada (Haldane), reproduz exatamente o mapa de ligação verdadeiro.

c) Baseado em Kosambi:

Mapa Original: [1]—5,0—[2]— 5,0—[3]—5,0—[4]

Mapa em Kosambi: [1]—4,83—[2]—4,81—[3]—4, 83— [4]

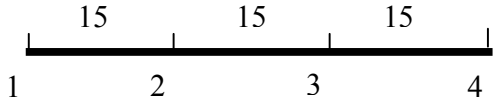
Mapa em Porcentagem de Recombinação: [1]—4,81—[2]— 4,80—[3]—4,81— [4]

Apesar de menores magnitudes, as alterações provocadas pela função de Kosambi são manifestadas no mapa gerado.

Para uma situação menos saturada, pode-se ilustrar:

$$D = \begin{pmatrix} 0 & 15 & 25,5 & 32,85 \\ & 0 & 15 & 25,5 \\ & & 0 & 15 \\ & & & 0 \end{pmatrix}$$

O agrupamento e ordenamento, proporcionaram o seguinte resultado:



Se for considerada a análise multiloco, baseada no método da máxima verossimilhança, seriam obtidas os seguintes resultados:

c) Baseado em porcentagem de recombinação

Mapa original: (1)-15.0 -(2)-15.0 -(3)-15.0 -(4)

Após análise multiloco: (1)-11.96 -(2)-12.75 -(3)-11.96 -(4)

Verificam-se discrepâncias consideráveis de cerca de 3 cM em razão do uso inadequado da função de mapeamento

d) Baseado em Haldane:

Mapa Original: (1)-15.0 -(2)-15.0 -(3)-15.0 -(4)

Mapa em Haldane: (1)-17.83 -(2)-17.83 -(3)-17.83 -(4)

Mapa em Porcentagem de Recombinação: (1)-15.0 -(2)-15.0 -(3)-15.0 -(4)

O mapa foi perfeitamente reconstituído em função do mapeamento adequado

e) Baseado em Kosambi:

Mapa Original: (1)-15.0 -(2)-15.0 -(3)-15.0 -(4)

Mapa em Kosambi: (1)-14.55 -(2)-14.29 -(3)-14.55 -(4)

Mapa em Porcentagem de Recombinação: (1)-14.15 -(2)-13.91 -(3)-14.15 -(4)

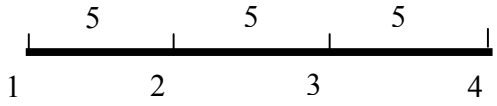
Discrepâncias em torno de 2 cM ocorreram em razão do uso da função de Kosambi, numa situação em que a interferência é nula.

Observa-se que quando se considera interferência nula c=1, ocorre um viés nas estimativas da freqüência de recombinação que tende a diminuir a distância entre as marcas principalmente após análise multiloco, o que confirma o fato observado nos mapas simulados para o processo de integração.

Caso 4- Vários marcadores – Coincidência igual a 2r

$$\begin{pmatrix} 0 & 5 & 9,9 & 14,60 \\ & 0 & 5 & 7,5 \\ & & 0 & 5 \\ & & & 0 \end{pmatrix}$$

O agrupamento e ordenamento, proporcionou o seguinte resultado:



Se for considerada a análise multiloco, baseada no método da máxima verossimilhança, seriam obtidas os seguintes resultados:

f) Baseado em porcentagem de recombinação

Mapa original: (1)-5.0 -(2)-5.0 -(3)-5.0 -(4)

Mapa em Morgan : (1)-4.86 -(2)-4.98 -(3)-4.86 -(4)

Após análise multiloco: (1)-4.86 -(2)-4.98 -(3)-4.86 -(4)

g) Baseado em Haldane:

Mapa Original: (1)-5.0 -(2)-5.0 -(3)-5.0 -(4)

Mapa em Haldane: (1)-5.47 -(2)-5.50 -(3)-5.47 -(4)

Mapa em Porcentagem de Recombinação: (1)-5.18 -(2)-5.20 -(3)-5.18 -(4)

h) Baseado em Kosambi:

Mapa Original (1)-5.0 -(2)-5.0 -(3)-5.0 -(4)

Mapa em Kosambi : (1)-5.02 -(2)-5.02 -(3)-5.02 -(4)

Mapa em Porcentagem de Recombinação: (1)-5.0 -(2)-5.0 -(3)-5.0 -(4)

Observa-se que quando se considera coincidência c=2r, a distância entre as marcas também diminui em uma proporção maior quando se considera interferência nula.

Outra situação a ser considerada é o efeito da imprecisão das estimativas da porcentagem de recombinação que pode conduzir a erros grosseiros na estimativa da porcentagem de recombinação e, principalmente, tornar o próprio ordenamento e agrupamento, originalmente estabelecido, de baixa confiabilidade.

Como ilustração será considerada a matriz:

$$D = \begin{pmatrix} 0 & 5 & 10 & 15 \\ & 0 & r & 10 \\ & & 0 & 5 \\ & & & 0 \end{pmatrix}$$

Já foi relatado que a situação em que r = 5 cM, retrata um mapa de agrupamento e ordenamento perfeito cuja análise multiloco, baseada em porcentagem de recombinação conduz a valores exatamente iguais ao mapa original.

Entretanto, os valores de r da matriz são estimativas e, portanto, apresentam desvios padrão de modo a dificultar tanto a obtenção das melhores estimativas multiloco do mapa quanto o ordenamento. Assim, serão consideradas várias situações enfatizando o efeito da imprecisão de apenas um valor da matriz, denotado por r, sobre o ordenamento e mapeamento.

Nos exemplos apresentados anteriormente, foi enfatizado os problemas inerentes à incorporação de informações conjunta, tendo em vista o desconhecimento prévio da taxa de interferência e do verdadeiro ordenamento entre locos. Fatores adicionais referem-se à precisão das estimativas e a proporcionalidade dos pesos de ponderação dados a eles, tendo em vista as diferentes funções de mapeamento.

Tabela 12 – ordem dos marcadores variando apenas uma estimativa de r.

r	Ordem	Mapa (Análise Multiloco em % Recombinação)
1	1-2-3-4	(1)-5.0 -(2)-1.0 -(3)-5.0 -(4)* (1)-6.43 -(2)-1.45 -(3)-6.43 -(4)** (1)-5.27 -(2)-4.51 -(3)-5.27 -(4)*** (1)-6.37 -(2)-1.48 -(3)-6.37 -(4)****
2	1-2-3-4	(1)-5.0 -(2)-2.0 -(3)-5.0 -(4) (1)-6.06 -(2)-2.66 -(3)-6.06 -(4) (1)-5.36 -(2)-4.35 -(3)-5.36 -(4) (1)-5.94 -(2)-2.65 -(3)-5.94 -(4)
3	1-2-3-4	Estimativas imprecisas

4	1-2-3-4	(1)-5.0 -(2)-4.0 -(3)-5.0 -(4) (1)-5.47 -(2)-4.54 -(3)-5.47 -(4) (1)-5.2 -(2)-4.64 -(3)-5.47 -(4) (1)-5.3 -(2)-4.39 -(3)-5.3 -(4)
5	1-2-3-4	(1)-5.0 -(2)-5.0 -(3)-5.0 -(4) (1)-5.24 -(2)-5.26 -(3)-5.24 -(4) (1)-5.0 -(2)-5.0 -(3)-5.0 -(4) (1)-5.06 -(2)-5.05 -(3)-5.06 -(4)
6	1-2-3-4	(1)-5.0 -(2)-6.0 -(3)-5.0 -(4) (1)-5.03 -(2)-5.87 -(3)-5.03 -(4) (1)-4.75 -(2)-5.45 -(3)-4.75 -(4) (1)-4.85 -(2)-5.62 -(3)-4.85 -(4)
7	1-2-3-4	(1)-5.0 -(2)-7.0 -(3)-5.0 -(4) (1)-4.86 -(2)-6.38 -(3)-4.86 -(4) (1)-4.45 -(2)-5.98 -(3)-4.45 -(4) (1)-4.68 -(2)-6.1 -(3)-4.68 -(4)
8	1-2-3-4	(1)-5.0 -(2)-8.0 -(3)-5.0 -(4) (1)-4.71 -(2)-6.81 -(3)-4.71 -(4) (1)-4.12 -(2)-6.58 -(3)-4.12 -(4) (1)-4.52 -(2)-6.51 -(3)-4.52 -(4)
9	1-2-3-4	(1)-5.0 -(2)-9.0 -(3)-5.0 -(4) (1)-4.59 -(2)-7.17 -(3)-4.59 -(4) (1)-3.37 -(2)-7.17 -(3)-3.77 -(4) (1)-4.39 -(2)-6.86 -(3)-4.39 -(4)
10	4-3-1-2	(4)-5.0 -(3)-10.0 -(1)-5.0 -(2) (4)-4.19 -(3)-7.74 -(1)-3.1 -(2) (4)-3.6 -(3)-9.78 -(1)- -0.27 -(2) (4)-4.01 -(3)-7.6 -(1)-2.82 -(2)

*Mapa original sem análise multiloco **Mapa após análise multiloco – Haldane

*** Mapa com análise multiloco - Morgan **** Mapa após análise multiloco-Kosambi

Como observado os resultados obtidos pela técnica de análise multiloco estarão sujeitos a perda de acurácia na medida em que a pressuposição sobre a interferência for equivocada e a precisão das estimativas for baixa proporcionando, inclusive, o estabelecimento de um ordenamento dúbio. A tabela 12 apresenta variações de um único valor de r em um mesmo agrupamento. As estimativas de distância entre os pares de marcas variam em função da alteração deste único valor de r, que pode provocar alterações drásticas nos agrupamento como imprecisão das estimativas, inversão de marcas e mesmo distâncias negativas entre os pares de marcas, como no caso em que r e igual a 10. A incorporação de toda a informação genética, na análise genômica de uma ou várias populações, é a maneira mais segura de inferir sobre as marcas moleculares associadas.

4.3.2 Comparação dos mapas consenso obtidos entre populações construídos com e sem análise multiloco

.

Após análise de cada um dos tipos mapas gerados nos itens, anteriormente discutidos, analisou-se os resultados obtidos nos dois tipos de mapas consenso obtidos (com e sem análise multiloco). O objetivo foi comparar os efeitos da análise multiloco na construção de mapas consenso, uma vez que os mapas simulados e consenso, obtidos com análise multiloco se mostraram menores, mais distantes dos valores do genoma original. Quanto aos valores de estresse, desvio padrão e variâncias para mapas construídos com análise multiloco não se mostraram discrepantes, devido às comparações realizadas com o genoma ideal.

4.3.2.1 Comprimento dos grupos de ligação

Como já mencionado, o comprimento médio esperado para cada grupo de ligação após o mapeamento das populações era de 50, 100 e 150cM, uma vez que esse é o comprimento de cada grupo de ligação no nível de saturação do genoma utilizado para a simulação das populações.

A média dos comprimentos de grupos de ligação está plotada nos Gráficos 1a, 1b, 1c. Com variação encontrada entre as medidas do comprimento, ao realizar o Teste de Tukey (5%), foi observada diferença estatística significativa entre as situações analisadas.

A análise referente ao comprimento médio do grupo de ligação, ao se observar os gráficos 1a, 1b e 1c, nota-se que o tamanho dos grupos de ligação gerados pela análise multiloco é inferior a aqueles obtidos pelos mapas construídos sem análise multiloco. Observa-se ainda que para os níveis de saturação de 5, 10 e 15 cM, pelo teste de Tukey ($P<0,05$) o comprimento dos mapas gerados, com e sem análise multiloco, pelas populações de tamanho 100 e 200 indivíduos se mostraram significativos em todos os níveis de saturação estudados. Para os mapas gerados com tamanho populacional de 400 indivíduos o teste de média não foi significativo.

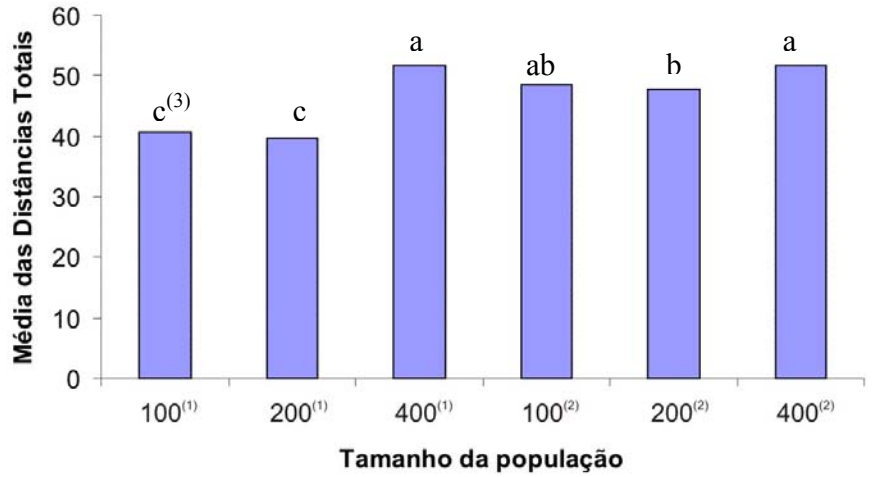


Gráfico 1a. (1) mapas em que foi utilizada a análise multiloco (2) mapas construídos sem análise multiloco, referentes aos tamanhos de população e comprimento médio dos grupos de ligação, e Teste Tukey a 5% de significância, saturação do mapa 5 cM.
(3) indica as letras correspondentes ao resultado da análise de médias feita pelo teste de Tukey a 5%, em que, médias seguidas pela mesma letra não diferem estatisticamente entre si.

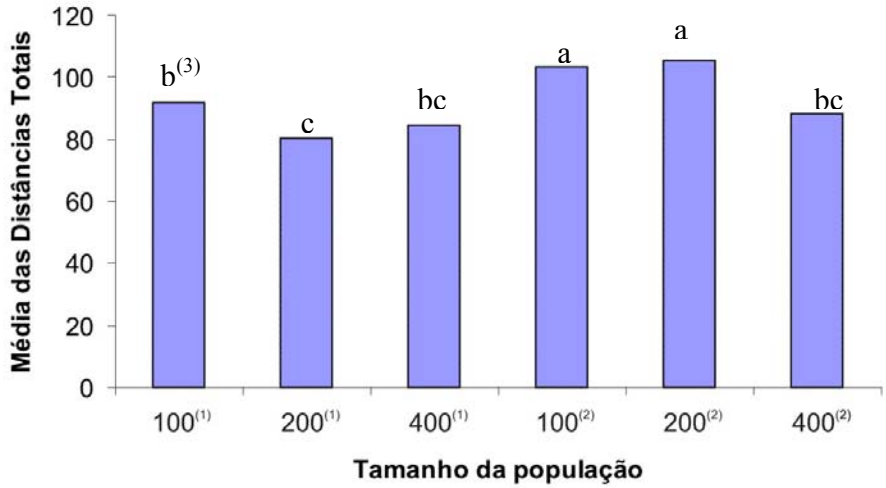


Gráfico 1b. (1) mapas em que foi utilizada a análise multiloco (2) mapas construídos sem análise multiloco, referentes aos tamanhos de população e comprimento médio dos grupos de ligação, e Teste Tukey a 5% de significância, saturação do mapa 10 cM.
(3) indica as letras correspondentes ao resultado da análise de médias feita pelo teste de Tukey a 5%, em que, médias seguidas pela mesma letra não diferem estatisticamente entre si.

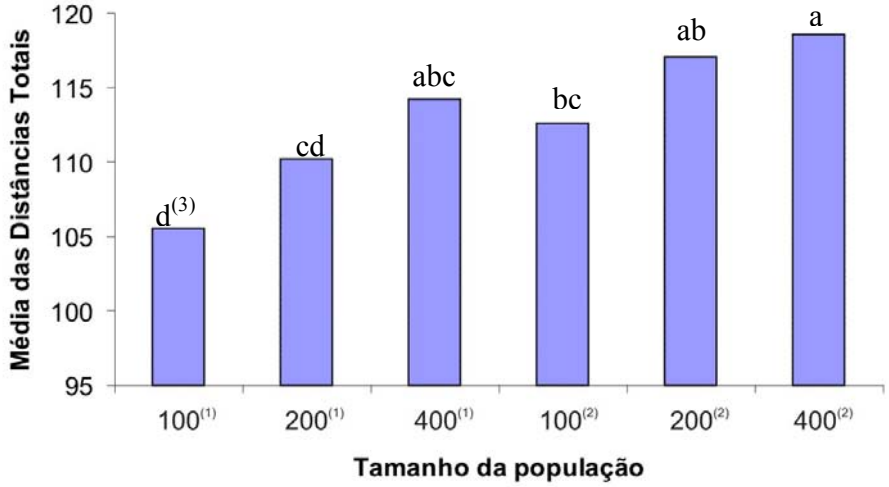


Gráfico 1c. (1) mapas em que foi utilizado a análise multiloco (2) mapas construídos sem análise multiloco, referentes aos tamanhos de população e comprimento médio dos grupos de ligação, e Teste Tukey a 5% de significância, saturação do mapa 15 cM.
(3) indica as letras correspondentes ao resultado da análise de médias feita pelo teste de Tukey a 5%, em que, médias seguidas pela mesma letra não diferem estatisticamente entre si.

4.3.2.2 Média das distâncias entre marcas adjacentes

A comparação feita utilizando os valores de média das distâncias entre marcas adjacentes para os vários tamanhos de população utilizados para construção de mapas com e sem análise multiloco estão representados nos gráficos 2a, 2b, e 2c, respectivamente.

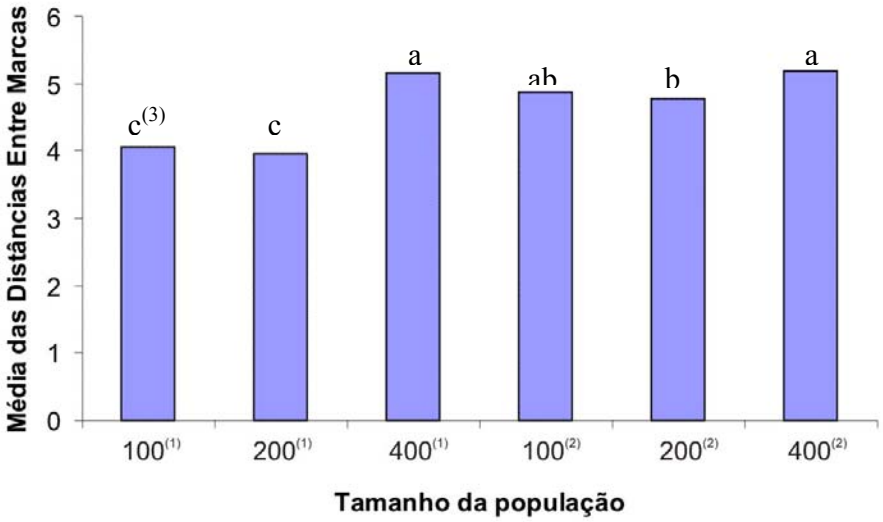


Gráfico 2a.(1) mapas em que foi utilizada a análise multiloco (2) mapas construídos sem análise multiloco, referentes aos tamanhos de população e média das distâncias de marcas adjacentes, e Teste Tukey a 5% de significância, saturação do mapa 5 cM.
(3) indica as letras correspondentes ao resultado da análise de médias feita pelo teste de Tukey a 5%, em que, médias seguidas pela mesma letra não diferem estatisticamente entre si.

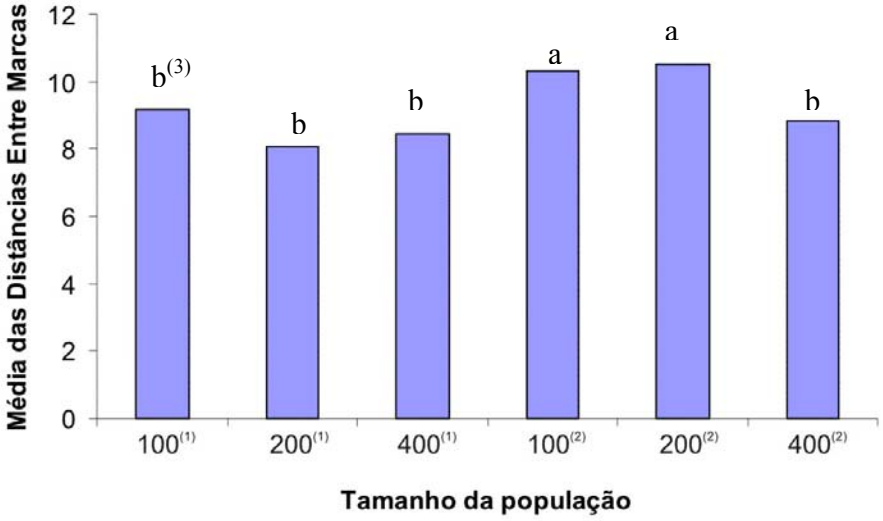


Gráfico 2a.(1) mapas em que foi utilizada a análise multiloco (2) mapas construídos sem análise multiloco, referentes aos tamanhos de população e média das distâncias de marcas adjacentes, e Teste Tukey a 5% de significância, saturação do mapa 10 cM.
(3) indica as letras correspondentes ao resultado da análise de médias feita pelo teste de Tukey a 5%, em que, médias seguidas pela mesma letra não diferem estatisticamente entre si.

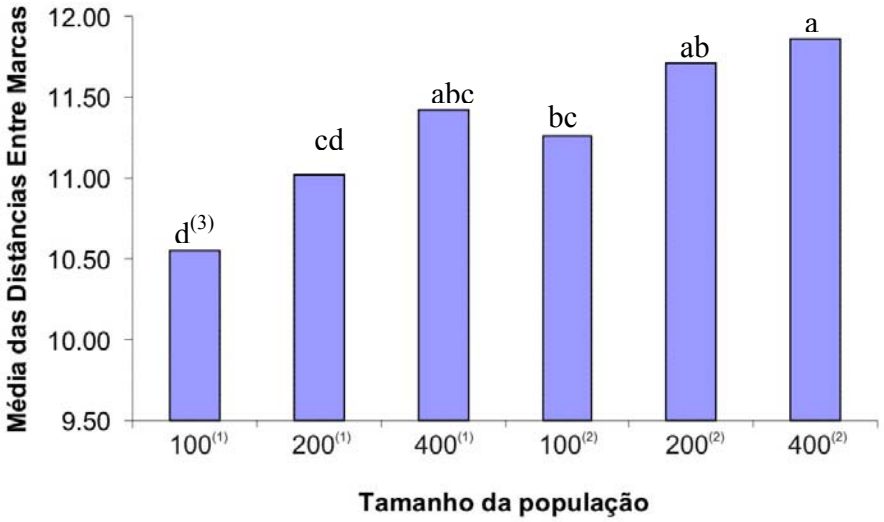


Gráfico 2a.(1) mapas em que foi utilizada a análise multiloco (2) mapas construídos sem análise multiloco, referentes aos tamanhos de população e média das distâncias de marcas adjacentes, e Teste Tukey a 5% de significância, saturação do mapa 15 cM.
(3) indica as letras correspondentes ao resultado da análise de médias feita pelo teste de Tukey a 5%, em que, médias seguidas pela mesma letra não diferem estatisticamente entre si.

Na análise referente a média das marcas adjacentes, nota-se que a média das distâncias das marcas geradas pela análise multiloco é inferior a aqueles obtidos pelos mapas construídos sem análise multiloco. Para o nível de saturação de 5 e 10 cM, o teste de Tukey se mostrou significativo entre as duas classes estudas para populações de tamanho 100 e 200 indivíduos. O teste de Tukey somente não foi significativo para populações de tamanho 400 indivíduos nos 3 níveis de saturação estudados.

4.3.2.3- Estresse

Os valores de estresse obtidos são apresentados a seguir, nos Gráficos 3a, 3b e 3c, juntamente com o resultado de comparações feitas com teste de médias. Analisando os efeitos do tamanho da população observou-se claramente a tendência de redução dos valores de estresse médio à medida que aumenta o tamanho da população nas duas situações avaliadas e que o estresse tende a ser maior nos mapas construídos com análise multiloco e saturação 5 cM e o inverso foi observado para saturação de 15 cM com o teste de Tukey significativo para todos os três pares de classes avaliados (100, 200 e

400). Para a saturação de 10 cM o teste somente se mostrou significativo para mapas com construídos com 100 indivíduos.

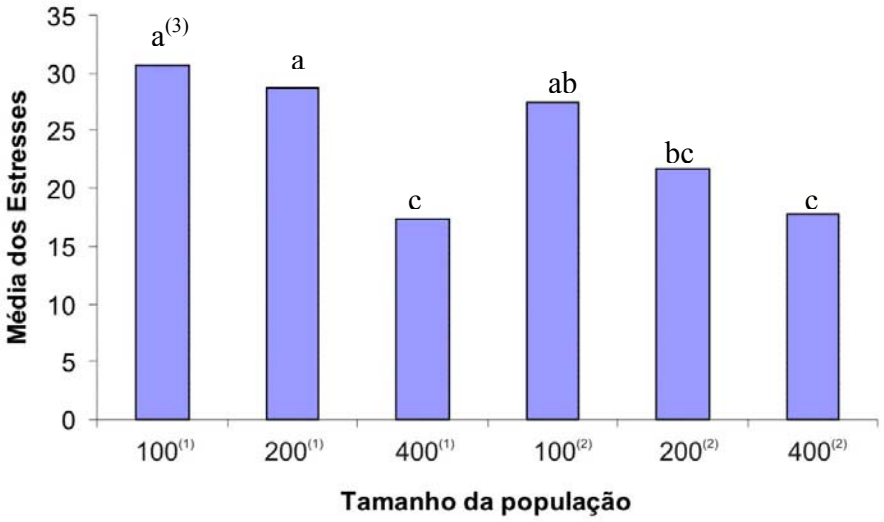


Gráfico 4a.(1) mapas em que foi utilizada a análise multiloco (2) mapas construídos sem análise multiloco, referentes médias dos estresses, e Teste Tukey a 5% de significância, saturação do mapa 5 cM.
(3) indica as letras correspondentes ao resultado da análise de médias feita pelo teste de Tukey a 5%, em que, médias seguidas pela mesma letra não diferem estatisticamente entre si.

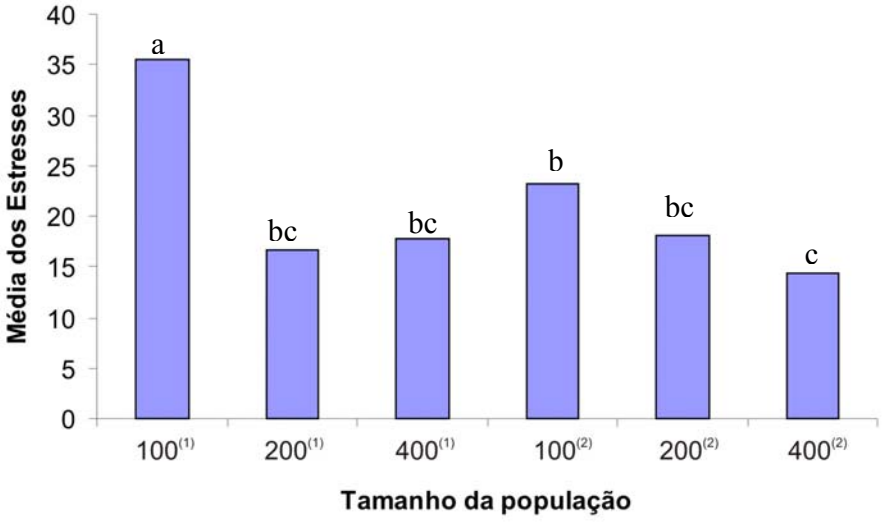


Gráfico 3b.(1) mapas em que foi utilizada a análise multiloco (2) mapas construídos sem análise multiloco, referentes médias dos estresses, e Teste Tukey a 5% de significância, saturação do mapa 10 cM.
(3) indica as letras correspondentes ao resultado da análise de médias feita pelo teste de Tukey a 5%, em que, médias seguidas pela mesma letra não diferem estatisticamente entre si.

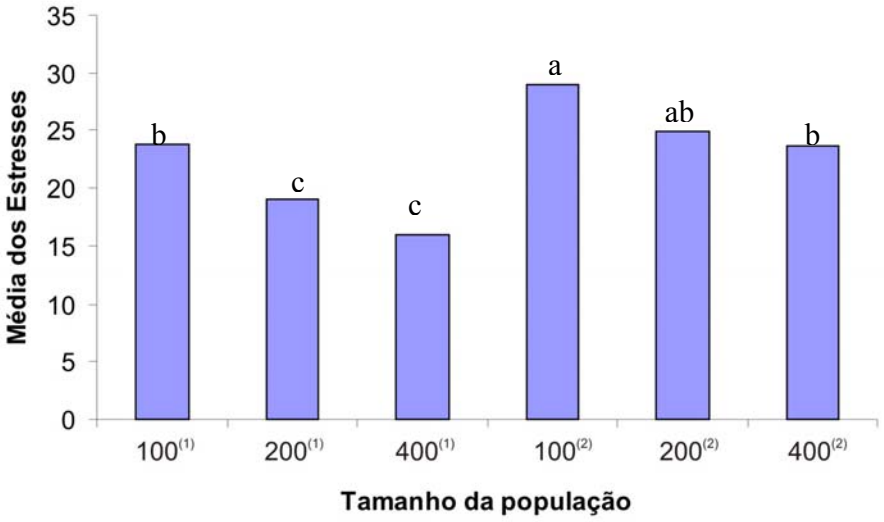


Gráfico 3c.(1) mapas em que foi utilizada a análise multiloco (2) mapas construídos sem análise multiloco, referentes médias dos estresses, e Teste Tukey a 5% de significância, saturação do mapa 15 cM.
(3) indica as letras correspondentes ao resultado da análise de médias feita pelo teste de Tukey a 5%, em que, médias seguidas pela mesma letra não diferem estatisticamente entre si.

Pelas análises observou-se que as diferenças entre os mapas consenso gerados com e sem análise multiloco são maiores para tamanhos menores tamanhos de população e à medida que o tamanho da população aumenta esta diferença tende a se anular.

4.4 – Integração de Mapas Genéticos

Mapas de ligação tem uma ampla variedade de aplicações tanto na genética quantitativa como na genômica. Numerosos mapas genéticos baseados em vários tipos de marcas genéticas tem se acumulado nos últimos anos. Classicamente, um único tipo de cruzamento é utilizado para construir mapas de ligação, resultando em uma população segregante (Maliepaard et al. 1997; Wu et al. 2002), ou alguns tipos específicos de população tais como RIL, retrocruzamentos, duplo-haplóides. Programas

bem estabelecidos de computadores, tais como MAP-MAKER, LINKAGE-1 e GQMOL (Lander and Green 1987; Suitter et al. 1983 e Cruz 2008), tem sido muito utilizados para análise destes dados. Mapas genéticos saturados têm sido construídos para diversas espécies de plantas cultivadas (Paterson et al.2000).

Existem pelo menos duas fontes de múltiplos mapas. Primeiro, o mapeamento de populações surge a partir de diferentes cruzamentos experimentais usando independentes materiais (ex. Sewell et al. 1999; Lombard and Delourme 2001). Segundo, quando alelos segregam em ambos os pais de espécies de fecundação cruzada, o mapa de ligação específico para cada pai é frequentemente construído, tal como na estratégia de mapeamento do pseudocruzamento teste (ex. Grattapaglia and Sederoff 1994; Marques et al. 1998; Testolin et al. 2001). O problema seria então como combinar múltiplos mapas genéticos em um único mapa com grande conteúdo de informação.

Stam (1993) considera uma metodologia no programa computacional "JoinMap" para integrar mapas individuais de ligação resultante de diferentes experimentos. Este método como dito anteriormente tem sido amplamente utilizado por diversos pesquisadores. O procedimento básico empregado pelo JoinMap é começar com estimativas de freqüência de recombinação individuais derivadas de diferentes experimentos. Estas estimativas são linearmente combinadas dentro de uma única estimativa usadas como peso.

Hu et al (2004) consideram um método de integração de mapas baseado em estimativas de verossimilhança. Este método propõe uma estimativa da função de verossimilhança comum que combina informações entre todos os cruzamentos, para obter uma estimativa comum de recombinação.

Programas como MapMaker e Carthagene não informam qual metodologia utilizam para integração de mapas genéticos bem como os artigos que fizeram uso destes programas computacionais, apenas informam qual programa utilizou e os comandos computacionais utilizados para se obter o mapa integrado.

Em vista das dificuldades de entendimento das metodologias propostas dos processos utilizados para integração dos mapas genéticos encontrados na literatura e da carência de pesquisa sobre a confiabilidade destes mapas gerados pelos diferentes programas e métodos, foi proposto um método de integração de mapas genéticos derivado de diferentes cruzamentos, implementado no programa de análises genômicas GQMOL (Cruz 2008).

Um mapa genético é essencialmente um arranjo linear de marcadores a partir de valores de recombinação. O método proposto baseia neste princípio e tem como principal ferramenta para integração de mapas genético de ligação a análise multiloco, assim como no método de Stam (1993). Em contraste com JoinMap, que estima a informação sobre a recombinação em um dado cruzamento a partir dos valores de LOD e então combina estimativas entre os cruzamentos assumindo uma distribuição binomial, o método proposto considera a informação de recombinação com base nos valores de variância, as quais dependem ou são influenciadas pelo tipo de população utilizado, tamanho da população, tipo de marcador, da freqüência de recombinação e da fase de ligação.

O grande problema da utilização do método proposto por Hu et al. (2004) é que eles não apresentaram um programa computacional específico para construção de mapas genéticos integrados, o que restringe substancialmente a utilização desta metodologia. Segundo Hu et al (2004), a metodologia proposta se mostrou, por simulações, mais eficiente que o método proposto por Stam (1993) quando marcas dominantes ou uma mistura de marcas co-dominantes e dominantes estão presentes em diferentes mapas que são utilizados na integração. Porém, para o emprego de sua metodologia somente a fase de ligação mais provável pode ser usada. Entretanto, isto seria mais eficiente com tamanho maior de progênies, onde a fase incorreta pode ter probabilidade muito menor que a fase correta, e ter pequeno efeito na probabilidade total. Entretanto, quando o tamanho da progênie é pequeno, escolhendo a fase mais provável, pode-se introduzir um desvio (viés) devido aos efeitos do pequeno tamanho da amostra. Pequenos tamanhos de amostra podem conduzir a viés levando a incongruência da posição das marcas como observado em Trigo (Daryl et al 2004).

Silfverberg-Dilworth et al. (2006) também relataram que alguns casos de inversão na ordem das marcas pode ocorrer, quando marcas então muito intimamente ligadas. O problema na identificação de inversões em um mapa integrado seria como identificá-las, uma vez que não há um mapa referência. Quando se trabalha com espécies de fecundação cruzada, tal como na estratégia de mapeamento do pseudocruzamento, as comparações podem ser feitas com os mapas individuais construídos para cada pai como realizada por Doligez et al. (2006).

Pelo método proposto foi testada a integração de mapas que somente possuíam marcadores âncoras. Foram gerados os mapas consenso entre populações, na menor

distância entre as marcas 5 cM no genoma original. O número de inversões foi relativamente baixo para integração entre mapas provenientes de populações de diferentes tamanhos, apenas uma em trinta repetições.

A proposta de se obter um mapa genético integrado basicamente envolve três passos: alinhamento, ordenamento e integração.

Na literatura o processo de alinhamento entre mapas, tem sido considerado como a identificação das marcas consideradas âncoras entre diferentes mapas. Este alinhamento entre mapas é feito "a mão" sem ajuda de um programa computacional como descrito em Doligez et al. (2006) e N'Diaye et al. (2008) (Figura 14) Nestes trabalhos foram integrados mapas genético de maçã e uva, respectivamente e utilizaram o programa Carthagene. O alinhamento também recebe nomes diferentes com a mesma finalidade na literatura como: comparação da ordem das marcas, ordem de leitura e consenso entre mapas.

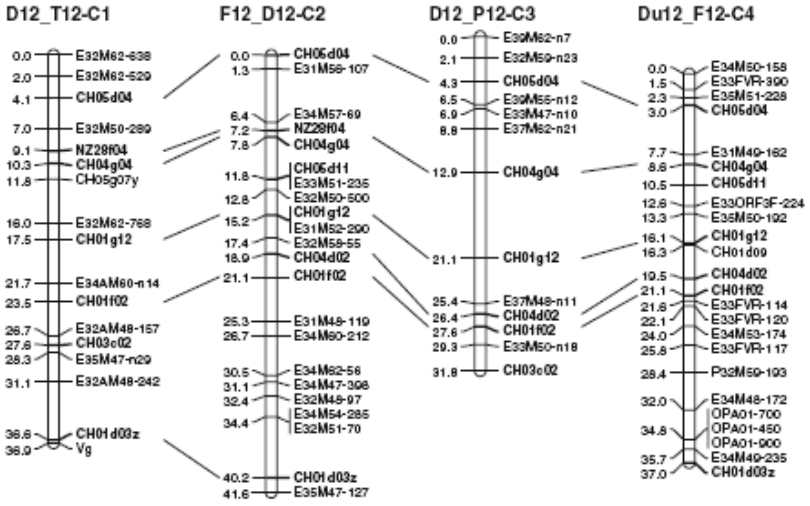


Figura 14- Alinhamento entre mapas genéticos de maçã apresentado por N'Diaye et al.(2008)

O processo de alinhamento de mapas proposto neste trabalho é um processo simples do qual se institui uma distância arbitrária entre as marcas em que uma delas é considerada como a primeira do mapa e as demais marcas posicionadas em relação à primeira. Neste mapa alinhado estão presentes todas as marcas inclusive os marcadores âncoras repetidas vezes. (Figura 15)

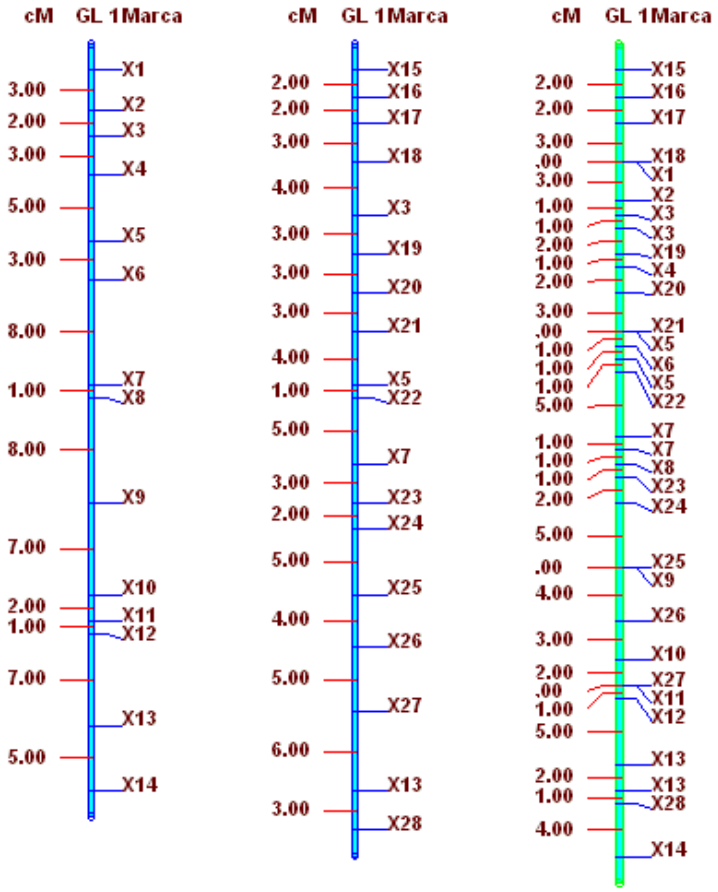


Figura 15 – Em azul grupo de ligação 1 de diferentes experimentos e em verde mapa alinhado em que todos os marcadores são representados inclusive os âncoras repetidas vezes.

O próximo passo para integração de mapas seria o ordenamento das marcas, no qual consiste simplesmente em organizar o mapa alinhado instituindo uma distância média entre cada par de marcas. Assim, no mapa ordenado passa a estar presente todas as marcas inclusive os marcadores âncoras apenas uma vez. (Figura 16)

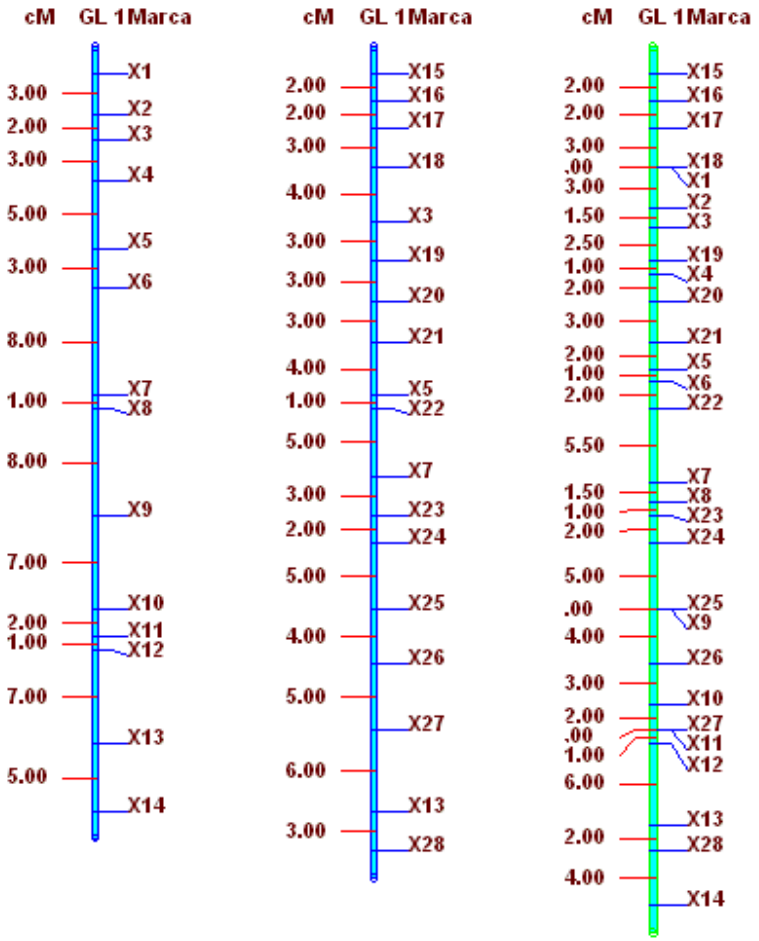


Figura 16 - Mapa ordenado. Em azul grupo de ligação 1 de diferentes experimentos e em verde mapa ordenado pela metodologia proposta.

Ordenado e alinhado os diferentes mapas genéticos, se processa a integração dos mapas genéticos pela análise multiloco. Uma vez adotada a análise multiloco, deve lembrar que as funções de mapeamento passam a exercer papel fundamental no resultado a ser obtido. A adoção de uma, entre as diferentes funções de mapeamento, depende das pressuposições a respeito da distribuição da permuta, do grau de interferência e do comprimento do segmento cromossômico considerado.

Para o estudo de mapas integrados e comprovação da funcionalidade da metodologia proposta foram utilizadas populações F_2 co-dominante e retrocruzamento com tamanhos de 100, 150, 200 e 400 indivíduos, com 21 marcas por grupo de ligação e marcadores eqüidistantes 5 cM, em um total de quatro simulações para F_2 co-dominante e quatro para retrocruzamentos,(Figura 17a e 17b).

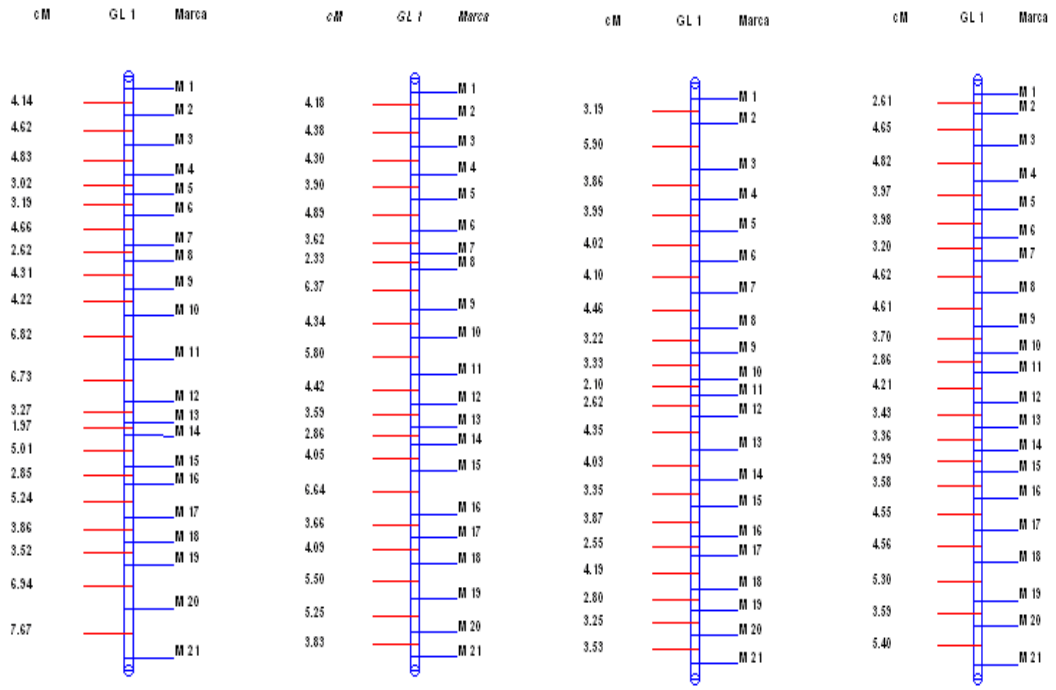


Figura 17a - Genomas estudados no processo de integração de mapas de populações F_2 co-dominante com tamanhos de 100(a), 150(b), 200(c) e 400(d) indivíduos.

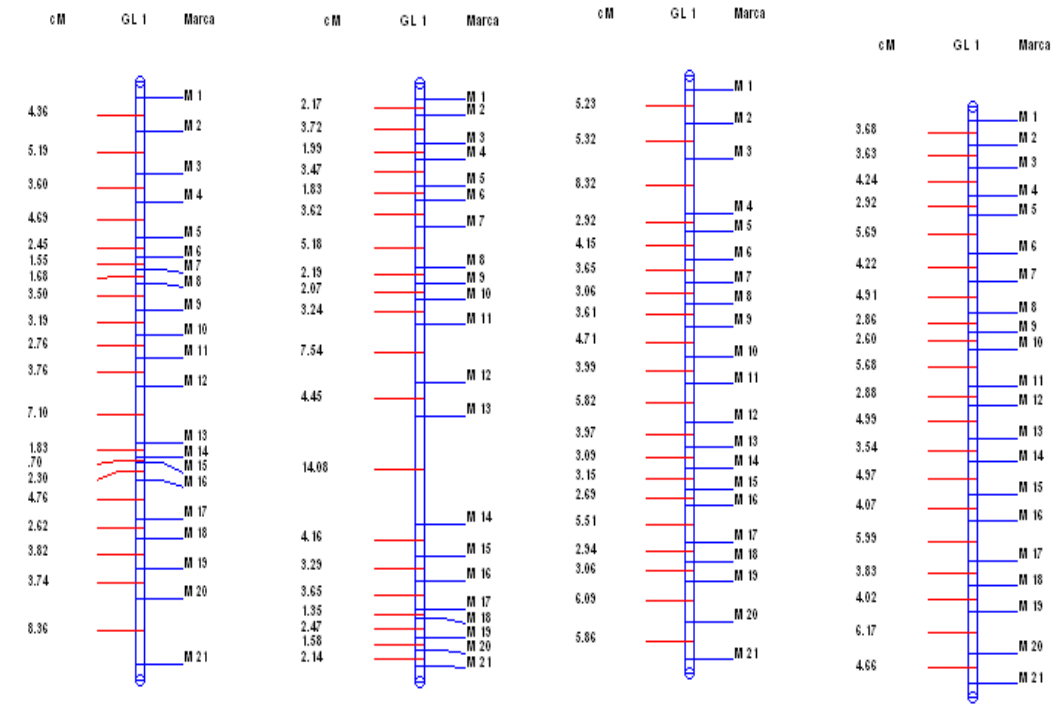


Figura17b - Genomas estudados no processo de integração de mapas de populações de retrocruzamentos com tamanhos de 100(a), 150(b), 200(c) e 400(d) indivíduos.

Quando se trabalha com vários mapas genético de uma determinada espécie, para integração entre os grupos de ligação de diversos mapas construídos com diferentes técnicas é necessário que entre os mesmos grupos de ligação tenha pelo menos um marcador que seja âncora. Assim, o passo inicial para o processo de integração seria a identificação destes marcadores que são âncoras. Identificados os marcadores âncora o passo seguinte seria obter o mapa alinhado, em seqüência o mapa ordenado e,por fim, o mapa integrado efetivo.

Cada genoma simulado foi fragmentado em quatro novos mapas de modo que foram obtidos três mapas com oito marcadores e um com nove marcadores, cada um destes mapas contendo quatro marcadores âncoras entre os quatro mapas (Figura 18), foram alinhados, ordenados, integrados e em seguida comparados com o mapa de origem.

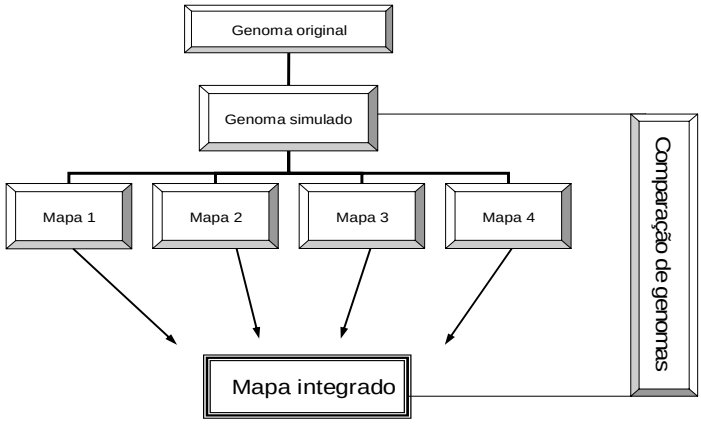


Figura 18 – Resumo do processo de estudo de mapas integrados

A integração de mapas gera um mapa com todos os marcadores com a melhor ordem, cujas distâncias entre os marcadores âncora e não âncoras são estimadas por análise multiloco. A veracidade das informações do mapa dependerá do tipo de população, do tipo de marca molecular e, principalmente, do número de indivíduos considerados na obtenção da porcentagem de recombinação entre pares de marcas.

Na tabela 13, estão os tipos e tamanhos de mapas utilizados no processo de integração, bem como a distância média de cada mapa, a variância, correlação de Spearman e o estresse avaliados em relação ao mapa de origem.

Tabela 13: Tipos e tamanhos de mapas utilizados no processo de integração, distância média de cada mapa, a variância, correlação de Spearman e o Estresse avaliados em relação ao mapa de origem .

Tamanho	*Tipo de Mapa*	*População*	*Tamanho*	*dist média*	*Variância*	*r Spearman*	*Estresse*
100	integrado	co-dominante	99.581	4.979	4.4891	1	16.4315
100	ordenado	co-dominante	89.4225	4.4711	2.4412	1	0.4237
150	integrado	co-dominante	97.1799	4.859	1.6399	1	13.3624
150	ordenado	co-dominante	88	4.4	1.1766	1	0
200	integrado	co-dominante	78.9327	3.9466	1.0607	1	11.1122
200	ordenado	co-dominante	72.71	3.6355	0.7025	1	0
400	integrado	co-dominante	87.9047	4.3952	1.057	1	13.3325
400	ordenado	co-dominante	79.99	3.9995	0.6434	1	0
100	integrado	retrocruzamento	78.8587	3.9429	4.2338	1	10.6942
100	ordenado	retrocruzamento	71.96	3.598	3.4173	1	0
150	integrado	retrocruzamento	82.6401	4.132	11.3211	1	15.2943
150	ordenado	retrocruzamento	74.19	3.7095	8.0796	1	0.4861
200	integrado	retrocruzamento	97.7246	4.8862	3.0727	1	17.6581
200	ordenado	retrocruzamento	87.14	4.357	2.1534	1	0
400	integrado	retrocruzamento	99.3229	4.9624	2.2587	1	51.9245
400	ordenado	retrocruzamento	80.89	4.4939	1.9859	1	47.483

As correlações de Spearman se mostraram constantes em todos os mapas integrados, tanto nas populações F_2 co-dominantes como nas de retrocruzamento, não apresentaram valores diferente da unidade em nenhum dos mapas integrados quando foram comparados com os mapas dos quais originaram. O que pode ser considerado um bom indicativo de confiabilidade do processo de integração dos mapas.

Para populações F_2 co-dominates, pelos mapas integrados, observa-se que quanto maior a população maior é tendência de aproximação do tamanho do grupo de ligação do mapa integrado ao tamanho do mapa dos quais foram fragmentados. O mesmo ocorre com a distância média das marcas. Quanto à variância pode-se observar que há uma queda da variância com o aumento da população; o mesmo não ocorre com o estresse que decresce somente até o tamanho populacional de 200 indivíduos.

Segundo Soller e Beckmann (1983), quando marcadores co-dominantes estão disponíveis, análises baseadas em gerações F_2 serão mais úteis que aquelas com base em gerações de retrocruzamento, por fornecerem informações tanto em relação à dominância quanto ao efeito maior do QTL identificado. Isso também pode ser uma explicação para a necessidade de maior tamanho de população para a obtenção de mapas mais confiáveis em populações de retrocruzamentos.

Para as populações de retrocruzamento observa-se que ocorreu o inesperado, com o aumento do número de indivíduos houve aumento no tamanho dos grupos de ligação e das distâncias médias entre as marcas. Estas medidas se afastaram daquelas observadas para os mapas dos quais eles se originaram, o que pode ser verificado com o aumento do estresse com o aumento da população, mas se aproximaram das medidas do mapa original que era de 100 cM e marcadores eqüidistantes a 5 cM.

Outro fato a ser destacado refere-se à construção dos mapas ordenados que em quase todas as situações eles apresentaram estresse zero ou bem próximo de zero. A única exceção foi para a população de retrocruzamento de tamanho 400 que apresentou um estresse elevado, ou seja, os mapas reconstituídos são exatamente iguais aos mapas de origem quando o estresse foi zero ou o estresse é muito pequeno em relação ao mapa dos quais eles foram originados. Este fato pode levar a conclusões equivocadas, pois poderia-se pensar que os mapas ordenados seriam mais indicados para integração dos mapas genéticos. Isto não seria uma realidade, pois quando se obtém o mapa integrado efetivo, deve-se lembrar que ele foi submetido a uma análise multiloco, ou seja, a ordem

bem como as distâncias entre as marcas foram analisadas em conformidade com todas as marcas presentes no grupo de ligação.

Como discutido anteriormente o programa mais utilizado para integração de mapas genéticos é o JoinMap que utiliza a metodologia proposta por Stam (1993) . Yan et al. (2005), utilizando mapas genéticos construídos para os sete grupos de ligação dos pais de uma progênie diplóide de rosas, obteveram o mapa integrado entre as dois progênies. Para comparar a metodologia proposta com a de Stam (1993), foi realizada a integração dos mapas referentes ao grupo de ligação três da espécie estudada, assim como realizado por Yan et al (2005), porém se utilizou o módulo integração de mapas genéticos do programa computacional GQMOL (Cruz 2008), baseado na proposta deste trabalho. (Figura 19)

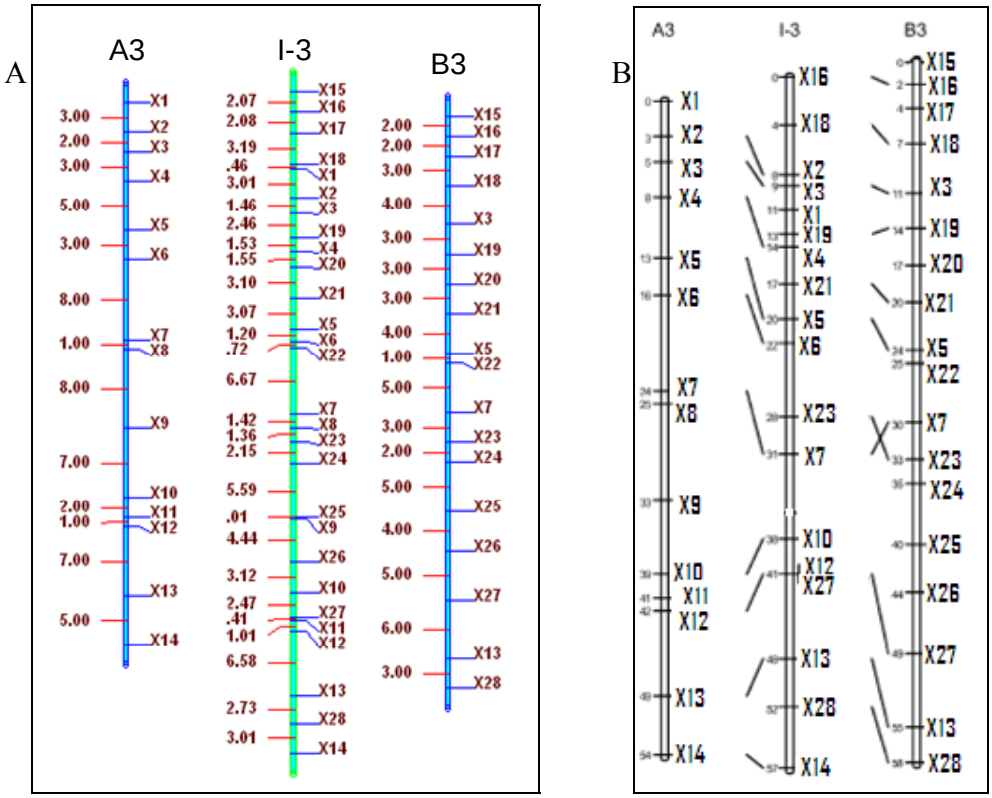


Figura19. A- Mapa integrado GQMOL e B - Mapa integrado JoinMap (Yan et al. 2005)

Para integrar os mapas a primeira marca do grupo A3 foi chamada de X1 e subseqüentemente até X14, o mesmo foi realizado para o B3 que a, primeira marca foi

chamada de X15 até a última X28, as marcas consideradas como âncoras receberam a mesma nomeclatura nos dois mapas. Em seguida ocorreu a integração dos mapas.

Quando se compara os mapas integrados pelos dois programas computacionais, observa-se que há pequenas diferenças entre as distâncias das marcas, comparações mais detalhadas são difíceis uma vez que os autores aproximam as distâncias no mapa para números inteiros e também não apresentaram os marcadores com distância inferior a dois cM, para melhor visualização. A maior diferença observada entre os dois mapas é a inversão em dois pares de marcas X7 com X23 e X27 com X12 que no mapa apresentado pelo JoinMap se apresentaram na mesma posição.

Diferenças na construção de mapas genéticos pelos diferentes programas são relatados na literatura. Sewell et al. (1999) trabalhando com mapas genéticos de pinus, relataram que os mapas genéticos obtidos pelo JoinMap são ligeiramente diferentes dos mapas obtidos pelo MAPMAKER. Qi et al. (1996) relataram que as diferenças também foram observadas em mapas de ligação construídos para cevada, e foi atribuída a maneira que cada programa calcula as distâncias, quando a interferência real difere da assumida.

Doligez et al (2006) integraram diferentes mapas de ligação com o programa Carthagene usando dados provenientes de família de irmãos completos em uva com tamanhos de 96, 45, 112, 139 e 153 indivíduos, o total do comprimento do mapa integrado de leitura foi 1485 cM com média de distância inter-locus de 6,2 cM. Os mesmos mapas foram integrados com o programa JoinMap, e os mapas integrados nos dois programas foram comparados e foram observadas algumas diferenças entre eles, principalmente inversões entre algumas marcas.

5- CONCLUSÕES

Para obtenção de informações suficientes, de modo que sejam gerados mapas integrados confiáveis, as análises devem ser associadas à utilização de tamanho de amostra e número de marcas adequadas. Mapas com distorções serão obtidos com uma utilização inadequada de indivíduos, mesmo que com uma grande quantidade de marcas e vice-versa. Recomenda-se não utilizar informações de populações com 100 indivíduos

ou menos para a geração de mapas e integração de mapas, tanto em populações F_2 como de retrocruzamento, pois neste nível é observada maior ocorrência de inversões bem como maiores problemas quanto à recuperação dos grupos de ligação.

A confiabilidade dos mapas consenso aumenta com o aumento do número de indivíduos. Entretanto a partir do tamanho 200 indivíduos não se observa grandes melhorias na qualidade dos mapas.

Para obtenção de mapas simulados, consenso e integrados a análise multiloco pode influenciar nas distâncias entre as marcas, gerando um maior viés nas estimativas à medida que a saturação do mapa diminui. O viés também é influenciado pelo tipo de função de mapeamento utilizado e, para mapas pouco saturados, a função de mapeamento de Haldane gera maiores variâncias, sendo menos indicada para mapas consenso e integrados.

O processo proposto de integração de mapa se mostrou eficiente para as populações estudadas, foram gerados mapas que apresentavam pequena tensão interna quando comparados com os mapas dos quais eles se originaram. A Integração de mapas se mostrou dependente do tipo de população utilizada, tamanho da população, tipo de marcador utilizado, da freqüência de recombinação e da fase de ligação.

Referências Bibliográficas

ACHERE, V.; FAIVRE-RAMPANT, P.; JEANDROZ, S.; BESNARD, G.; MARKUSSEN, T.; ARAGONES, A.; FLADUNG, M.; RITTER, E.; FAVRE,J.M. A full saturated linkage map of Picea abies including AFLP, SSR, ESTP, 5S rDNA and morphological markers. **Theoretical and Applied Genetics** 108: 1602-1613, 2004.

AHN, S.; TANKSLEY, S.D. **Comparative linkage maps of the rice and mize genomes. Proceedings of the National Academic of Sciences of the United States of America**, Washigton, v.90, n.17, p.7980-7984, 1993.

BAILEY, N.T.J. (1961) ***Introduction to the Mathematical Theory of Genetic Linka*ge. Oxford**: Clarendon Press.

BEAVIS, W.D.; GRANT, D. A Linkage map based on information from four F2 populations of maize (Zea Mays L.). **Theor Appl Genetic** 82(5): 636-644, 1991.

BLANC, G.; CHARCOSSET, A.; MARGIN, B.; GALLAIS, A.; MOREAU, L.;(2006) Connected populations for detecting quantitative trait loci and testing for epistasis: an application in maize. **Theor Appl Genetic** 113(2):206-224.

BRONDANI, R.P.V.; CAMPINHOS, E.; GRATTAPAGLIA,D. Mapped RAPD markers are transferable among Eucalyptus trees from the same populations. In: **Procedings of the International IUFRO conference on Eucalyptus genetics and silviculture**, Salvador, Brasil, Vol.2, p. 111-115. 1997.

BROWN, D.M., MATISE, T.C., KOIKE, G., SIMON, J.S., WINER, E.S., ZANGEN, S., MCLAUGHLIN, M.G., SHOIZAWA, M., ATKINSON, O.S., JR., J.R.H.,

CHAKRAVARTI, A., LANDER, E.S, JACOB, H.J. (1998) An integrated genetic linkage map of the laboratory rat. **Mammalian Genome** 9: 521-530.

BROWN, G.R.; KADEL III, E.E.; BASSONI, D.L.; KIEHNE, K. L.; TEMESGEN, B.; VAN BUIJTENEN, J.P.; SEWELL,M.M.; MARSHALL, K.A.; NEALE, D.B. 2001. Anchored reference loci in loblolly pine (pinus taeda L.) for integrating pine genomics. **Genetics** 159: 799-809, 2001.

BUROW, M. D.; COORS, J. G. Diallel: a microcomputer program for the simulation and analysis of diallel crosses. **Agronomy Journal,** Madison, v. 86, n. 1, p. 154-158, Jan./Feb. 1994.

CARNEIRO, M.S.; M.L.VIEIRA.Mapas genéticos em plantas. **Bragantia**, v.61, n.2. p. 89-100.2002.

CHANG, Y., TAO, Q., SCHEURING, C., DING, K., MEKSEM, K., ZHANG, H. (2001) An integrates Map of Arabidopsis thaliana for functional analysis of its genome sequence. **Genetics**, 159: 1231-1242.

COE, H.E.; NEUFFER, M.G.; HOISINGTON, D.A. The genetics of corn. In: SPRAGUE, G.F.; DUDLEY, J.W. (Eds.). **Corn and corn improvement**, 1988. p.81-237.

COELHO, A.S.G. Considerações gerais sobre a análise de QTL´s. In: PINHEIRO, J.B.; CARNEIRO,I.F.(Eds.). **Análise de QTL no melhoramento de plantas**. Goiânia: FUNAPE, 2000. p.1-36.

CRUZ, C. D. **Programa Genes**: aplicativo computacional em estatística aplicada à genética. Genetics and Molecular Biology, Ribeirão Preto, v. 21, n. 1, p. 135-138, Mar. 1998.

CRUZ, C. D. **A informática no melhoramento genético**. In: NASS, L. L. et al. (Eds).

Recursos Genéticos e Melhoramento de Plantas**.** Rondonópolis: Fundação-MT, 2001. p.1086-1118.

CRUZ, C.D. (2001) **Programa Genes***, versão Windows*: *aplicativo computacional em genética e estatística*. Imprensa Universitária, Viçosa, 648p.

CRUZ, C.D. SCHUSTER, I.(2001); **Programa GQMOL**: *Programa para análise de genética quantitativa e molecular*. versão: 2008.

CRUZ, C.D., CARNEIRO, P.C.S. (2003) **Modelos Biométricos aplicados ao melhoramento genético vol 2**, Editora UFV.

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

DIRLEWANGER, E.; GRAZIANO E.; JOOBEUR T.; GARRIGA-CALDERE, F.; CASSON, P.; HOWARD, W.; ARUS, P. Comparative mapping and marker-assisted selection in Rosaceae fruit crops. **Proc Natl Acad Sci** USA 101(26):9891-9896. 2004.

DOLIGEZ, A., ADAM-BLONDON, A.F., CIPRIANI, G., DI GASPERO, G., LAUCOU, V., MERDINOGLU, D., MEREDITH, C.P., RIAZ, S., ROUX, C., THIS, P. (2006). An integrated SSR map of grapevine base don five mapping populations. **Theoretical and Applied Genetics**. 113: 369-382.

EUCLYDES, R. F. **Uso de sistemas Genesys na avaliação de métodos de seleção clássicos e associados a marcadores moleculares.** 1996. 135 p. Tese (Doutorado) – Universidade Federal de Viçosa, Viçosa, MG.

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

FERREIRA, M.E.; GRATTAPAGLIA, D. **Introdução ao uso de marcadores moleculares em análise genética. Brasília**: EMBRAPA/CENARGEM, 1996.220p.

FERREIRA, D. F. Uso de simulação no melhoramento. In: NASS, L. L.; VALOIS, A. C. C.; MELO, I. S. de; VALADARES-INGLIS, M. C. (Ed). Recursos genéticos melhoramento de plantas**.** Rondonópolis: Fundação-MT, 2001. p. 1119-1141.

FREYRE, R.; P.W.SKROCH; V.GEFFROY; A.F.ADAM-BLONDON; A. SHIRMOHAMADALI; W.C.JOHNSON; V.LLACA; R.O.NODARI; P.A.PEREIRA; S. M. TSAI; J. TOHME; M.DROM; J.NEINHUIS; C.E.VALLEJOS; P.GEPTS. Towards an integrated linkage map of commun bean. 4. Development of a core linkage map and alignment of RFLP maps. **Theor Appl Genet**, v.97, p.847-856. 1998.

GARCIA, A.A.F., KIDO, E.A., MEZA, A.N., SOUZA, H.M.B., PINTO, L.R., PASTINA, M.M., LEITE, C.S., SILVA, J.A.G., ULIAN, E.C., FIGUEIRA, A., SOUZA, A.P. (2006). Development of an integrated genetic map of a sugarcane (Saccharum spp.) commercial cross, based on a maximum-likelihood approach for estimation of linkage and linkage phases. **Theoretical and Applied Genetics** 112: 298-314.

GOLDSTEIN, D. R.; ZHAO, H.; SPEED, T. P.(1995) Relative efficiencies of $\chi 2$ models of recombination for exclusion mapping and gene ordering. **Genomics**, v.27, p. 265-273.

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

GRIFFITHS, A.J.F.; MILLER, J.H.; SUZUKI, D.T.; LEWONTIN,R.C.; GELBART, W.M. **Introdução à genética**. Rio de Janeiro: Guanabara, 1998.856p.

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

FRASER, A. S. Simulation of genetics systems by automatic digital computers. I: Introduction. **Australian Journal of Biological Science**, Melboourne, v. 10, p. 484-491, 1957a.

FRASER, A. S. Simulation of genetics systems by automatic digital computers. II: Effect of linkage on rates of advance under selection. **Australian Journal of Biological Science**, Melbourne, v. 10, p. 492-499, 1957b.

FRISH, M., BOHN, M., MELCHINGER, A. E. Comparison of selection strategies for marker-assisted backcrossing of a gene. **Crop Sci**. 39: 1295-1301. 1999.

FUCHS, J.; KLOOS, D.; SCHUBERT I (1996) In situ localization of yeast artificial chomosome sequences on tomato and potato metaphase chromosome. **Chrom Res** 4:277-281.

HALDANE, J.B.S. (1919) The combination of linkage values, and the calculation of distance between linked factors. **Journal of Genetics**. 8, 299-309.

HARPER, L.; CANDE, W; (2000) Mapping a new frontier, development of integrated cytogenetic maps in plants. Func Integr Genom 1:89-98.

HOSPITAL, F.; CHEVALET, C.; MULSANT, P. Using markers in gene introgression breeding programs. **Genetics**, v.132, p. 1199-1210, December, 1992.

HU, X.S.; C. GOODWILLIE; K. M. RITLAND.(2004) Joining genetic linkage maps using a joint likelihood function. **Theoretical and applied Genetics**, v.109, n.5, Sep, p. 996-1004.

Hu, X.S.; C.GOODWILLIE; K.M.RITLAND. Joining genetic linkage maps using a joint likelihood function. **Theor Appl Genet**, v.109, n.5, Sep, p.996-1004. 2004.

JENSEN, J. and JORGENSEN, J.H. The barley chromosome 5 linkage map. **Hereditas**, 80, 5-16, 1975

KIANIAN, S.F.; QUIROS, C.F.; Geration of a Brassica oleracea composite RFLP map: linkage arrangements among various populations and evolutionary implications. **Theoretical an Applied Genetics**, New York, v.84, n.5/6, p.544-554, 1992.

KNAPP, S.J. Mapping quantitative trait loci using molecular markers: multilocus estimators of backcross, recombinant imbread, and doublet haploid parameters. **Theoretical and Applied Genetic**, New York, v.81, p.333-338, 1991.

KLEIN, P.E, KLEIN, R.R., CARTINHOUR, S.W., ULANCH, P.E., DONG, J., OBERT, J.A., MORISHIGE, D.T., SCHLUETER, S.D., CHILDS, K.L., ALE,M., MULLET, J.E.(2000). A high-throughput AFLP-based method for construction integrated genetic and physical maps: Progress toward a sorghum genome map. **Genome Research** 10: 789-807.

KOSAMBI, D.D. (1944). **The estimation of map distance from recombination values**. *Ann. Eugen*. 12, 172-175.

LALOUEL, J.M. Linkage mapping from pair-wise recombination data. **Heredity**, 38, 61-77, 1977.

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

LANDER, E.S.; WEINBERG, R.A. Genomics: journey to the century of biology. **Science**, Washington, v.287, n.5459, p.1777-1782,2000.

LANDER, E.S.; GREEN, P.; J. ABRAHANSON; A. BARLOW; M. J. DALY; S. E. LINCOLN; L. NEWBURG. MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. **Genomics**, v.1, n.2, Oct, p.174-81. 1987.

LANDER, E.S.; GREEN, P.; (1987) Construction of multilocos genetic linkage maps in humans. **Proc Natl Acad Sci** USA 84: 2363-2367.

LEE, M. **DNA markers and plant breeding programs. Advances in Agronomy**, San Diego, v.287, n.5459, p.1777-1782, 1995.

LESLEY, J.W. (1932) Trisomic types of the tomato and their relation to the genes, **Genetics** 17:545-559.

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

LEWONTIN, R.C.; HUBBY, J.1966. A molecular approach to the study of genetic heterozygosity in natural populations. II. Amounts of variation and degree of heterozygosity in natural populations of Drosophyla pseudoobscura. **Genetics** 54:595-609.

LIU, B.H. **Statistical genomics: linkage, mapping, and QTL analysis** (1998). Boca Raton, Flórida, USA: CRC Press. 605p.

LIU, B. H.; S. J. KNAPP. GMENDEL: a program for Mendelian segregation and linkage analysis of individual or multiple progeny populations using log-likelihood ratio's. **J. Hered**, v.81, p.407. 1990.

LOMBARD, V.; DELOURNE, R.; (2001) A consensus linkage map for repessed (Brassica napus L.): construction and integration of there individual maps from DH populations. **Theor Appl Genet** 103:491-507.

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

MACKAY, I.J. AND CALIGARI, P.D.S. (1999) Major errors in data and their effect on response to selection in plants. **Crop Science**. 39: 697-702.

MALIEPAARD, C.; JANSEN, J.; VAN OOJEN J.W.(1997)Linkage analysis in a full-sib family of an outbreeding plant species: overview and consequences for applications. **Genetc Res** 70: 727-250.

MARCEL, T. C.; R. K. VARSHNEY; M. BARBIERI; H. JAFARY; M. J. De KOCK; A. GRANER; R. E. NIKS. A high-density consensus map of barley to compare the distribution of QTLs for partial resistance to Puccinia hordei and of defence gene homologues. **Theor Appl Genet**, Nov 18. 2006.

MARQUES, C.M.; ARAUJO, J.A.; FERREIRA, J.G.; WHETTEN, R.; O'MALLEY D.M.; LIU, B.H.; SEDEROFF, R.; (1998) AFLP genetic maps of Eucalyptus globules and E. tereticornis. **Theor Appl Genet** 96:727-737.

MARTINEZ, O.; CURNOW, R.N. Estimating the locations and the size of the effects of quantitative trait using flanking markers. **Theor. Appl. Genet**., v.85, p. 480-488, 1992.

MARTINEZ, M.L.; VUKASINOVIC, N.; FREEMAN, A.E. Random model approach for QTL mapping in half-sib families. **Genet. Sel. Evol**., Elsevier, Paris, v.31, p.319-340, 1999.

MOORE, G.; DEVOS, K.M.; WANG, Z.; GALE, M.D. Cereal Genome Evolution: Grasses, line up and from a circle. **Current Biology**, London, v.5, n.7, p.737-749, 1995

MORGAN, T.H. **The theory of genes**. New Haven: Yale University Press, 1928.

MURPHY, R.W.; J.W.; BUTH, D.G.; HAUFLER, C.H. **Proteins I: isoenzyme electrophoresis**. In: HILLS, D.M.; MORITZ, C.(Eds). Sunderland: Sinauer Associates, 1990. p.45-126.

NICOLAS, S.D.; MIGNON, G.L.; EBER, F.; CORITON, O.; MONOD, H.; CLOUET, V.; HUTEAU, V.; LOSTANLEN, A.; DELOURME, R.; CHALHOUB, B.; RYDER, C.D.; CHEVRE, A.M.; JENCZEWSKI, E.(2007). Homologous recombination plays a major role in chromosome rearranjaments that occurs during meiosis of Brassica napus Haploids. **Genetics** 175(2): 487-503.

OOIJEN, J. W. V. Accuracy of mapping quantitative trait loci in autogamous species. **Theor, Appl. Genet**., v.84, p.803-811, 1992.

OWEN, A.R.G. (1950) **The theory of genetical recombination**. *Adv. Genet*. 3, 117-157.

PARAN, I.; J. R. VAN DER VOORT; V. LEFEBVRE; M. JAHN; L. LANDRY; M.V. SCHRIEK; B.TANYOLAC; C. CARANTA; A.B. CHAIN; K. LIVINGSTONE; A. PALLOIX; J. PELEMAN. (2004). An integrated genetic linkage map of pipper (*Capsicum spp.*) **Molecular Breeding**, v.13, p.251-261.

PELEMAN, J.; R.V. WIJK; J.V. OEVEREN; R. V. SCHAIK.(2000) **Linkage map integration: an integrate genetic map of Zea mays L**. Plant and Animal Genome Conference VIII. San Diego, California, USA., 472p.

PELGAS, B.; BEAUSEIGLE, S.; ACHERÉ, V.; JEANDROZ, S.; BOUSQUET, J.; ISABEL, N.(2006) Comparative genome mapping among Picea glauca, P. mariana

x P.rubens and P.abies, and correspondence with other Pinaceae. **Theor Appl Genet** 113(8):1371-1393.

QI, X.; T. S. PITTAAWAY; S. LINDUP; H. LIU; E. WATERMAN; F. K. PADI; C. T. HASH; J. ZHU; M. D. GALE; K. M. DEVOS. An integrated genetic map and a new set of simple sequence repeat markers for pearl millet, Pennisetum glaucum. **Theor Appl Genet**, v.109, n.7, Nov, p.1485-93. 2004.

RICK,C.M.; YODER,J.I. **Classical and molecular genetics of tomato highlights and perspectives**. Annual Review of Genetics, Palo Alto, v.22, p.281-300,1988.

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

SAHA, M.C.; MIAN, R.; ZWONITZER, J.C. CHEKHOVSKIY, K.; HOPKINS, A.A.; (2005) An SSR- and AFLP-based genetic linkage map of tall fescue (Festuca arundinacea Schreb.) **Theor Appl Genet** 110(2): 323-336.

SCHUSTER, I., CRUZ, C.D.(2004). **Estatística genômica aplicada a populações derivadas de cruzamentos controlados**. 1ª Ed. Viçosa: UFV. 568p.

SEWELL, M.M.; SHERMAN, B.K.; NEALE, D.B. A consensus map for Loblolly Pine (Pinus taeda L.): I. Construction and integration of individual linkage maps from two outbread three-geration pedigrees. **Genetics**, Baltimore, v.151, n.1, p.321-330, 1999.

SILVA, L. da C. E. **Simulação do tamanho da população e da saturação do genoma para mapeamento genético de RILs**. Viçosa, MG: UFV, 2005.120f. Dissertação (Mestrado em Genética e Melhoramento) – Universidade Federal de Viçosa, Viçosa.

SIMMONDS, N. W. Selection for local adaptation in a plant-breeding program. **Theoretical and Applied Genetics**, Berlin, v. 82, n. 3, p. 363-367, 1991.

SILVERVERG-DILWORTH, E.; MATASCI, C.; VAN DE WEG W.E.; VAN KAAUWEN, M.; WALSER, M.; KODDE, L.; SOGLIO, V.; GIANFRANCESCHI, L.; DUREL, C.E.; COSTA, F.; YAMAMOTO, T.; KOLLER, B.; GESSLER, C.; PATOCCHI, A. (2006) Microssatellite markers spanning the apple (Malus domestica Borkh.) **Genome. Tree Genetics & Genomes** 2(4):202-224.

SOLLER, M.; BECKMANN, J.S. Genetic polymorphism in varietal identification and genetic improvement. **Theor. Appl. Genet**., v. 67, p.25-33, 1983.

SONG, G.J., MAREK, L.F., SHOEMAKER, R.C., LARK, K.G., CONCIBIDO, V.C., DELANNAY, X., SPECHT, J.E., CREGAN, P.B. (2004) A new integrated genetic linkage map of the soybean. **Theoretical and applied genetics** 109: 122-128.

SOMERS, D. J.; P. ISAAC; K. EDWARDS. A high-density microsatellite consensus map for bread wheat (Triticum aestivum L.). **Theor Appl Genet**, v.109, n.6, Oct, p.1105-14. 2004.

SONG, Q. J.; L.F.MAREK; R.C.SHOEMAKER; K. G. LARK; V.C.CONCIBIDO; X. DELANNAY; J.E.SPECHT; P.B.CREGAN. A new integrated genetic linkage map of the soybean. **Theor Appl Genet**, v.109, n.1, Jun, p.122-8. 2004.

SUITER, K. A.; J. F. WENDEL; J. S. CASE. LINKAGE-1: a PASCAL computer program for the detection and analysis of genetic linkage**. J. Hered**, v.74, n.3, May-Jun, p.203-4. 1983.

STAM, P. (1993). Construction of integrated genetic linkage maps by means of a new computer package: JoinMap. **The Plant Journal,** v.3, n.5, p. 793-744.

STUTERVANT, A. H. The linear arrangement of six sex-linked factors in Drosophila, as shown by their mode of association. **Journal of Experiemental Zoology**, v.14, p. 43-59, 1913.

SYMONDS, V.V.; GODOY, A.V.; ALCONADA, T.; BOTTO, J.F.; JUENGER, T.E.; CASAL, J.J.; LLOYD, A.M. (2005). Mapping quantitative trait loci in multiple populations of Arabidopsis thaliana identifies natural allelic variation for trichome density. **Genetics** 169(3):1649-1658.

TANI, N.; TAKAHASHI, T.; IWATA, H.; MUKAI, Y.; UJINO-IHARA, T.; MATSUMOTO, A.; YOSHIMURA, K.; YOSHIMURU, H.; MURAI, M.; NAGASAKA, K.; TSUMURA, Y. A consensus linkage map for sugi (Cryptomeria japonica) from two pedigrees, based on microsatellites and expressed sequence tags. **Genetics**. 165: 1551-1568, 2003.

TANKSLEY, S.D.; BERNATZKY, R.; LAPITAN, N.L.; PRINCE, J.P. **Conservation of gene repertoire but not gene order in pepper and tomato**. Proceedings of the National Academic of Sciences of the United States of America, Washington, v.85, n.17, p. 6519-6423, 1988.

TESTOLIN, R.; HUANG, W.G.; LAIN, O.; MESSINA, R.; VECCHIONE, A.; CIPRIANI, G. A kiwifruit (actinidia spp.) linkage map based on microsatellites and integrated with AFLP markers. **Theoretical and Applied Genetics**, New York, v. 103, p.30-36, 2001.

ULLOA, M.; MERIDITH, W.R.; SHAPPLEY, Z.W. KAPHLER, A.L. (2002)RFLP genetic linkage maps from F2.3 populations and a joinmap of Gossypium hirsutum. **Theor Appl Genet** 104: 200-208.

UNIVERSIDADE FEDERAL DE VIÇOSA. GQMOL. Genética quantitativa e molecular. desenvolvido por Cosme Damião Cruz, 2004. Disponível em: <www.ufv.bg/dbg/gqmol/gqmol.htm>. Acesso em: 01/06/2008.

VAN DEYNZE, A.; VAN DER KNAAP, E.; FRANCIS, D. (2006) **Development and application o fan informative seto f anchored markers for tomato breeding**. In: Plant and Animal Genome XVI conf, San Diego, CA, USA. P.188.

VAN OOIJEN, J. W.;R. E. VOORRIPS. JoinMap 3.0: software for the calculation of genetic linkage maps. **Plant Research International, Wageningen, the Netherlands.** 2001.

VISSCHER, P. M.; HALEY, C. S.; THOMPSON, R. Marker-assisted introgression in backcross breeding programs. **Genetics,** v.144, p. 1923-1932, December, 1996.

VON Os, H.; S.ANDRZEJEWSKI; E.BAKKER; I.BARRENA; G.J.BRYAN; B. CAROMEL; B.GHAREEB; E.ISIDORE; W.De JONG; P.VAN KOERT; V.LEFEBVRE; D.MILBOURNE; E.RITTER; J.N.VAN DER VOORT; F.ROUSSELE-BOURGEOIS; J. VAN VLIET; R.WAUGH;R.G. VISSER; J.BAKKER; H.J.VAN ECK. Construction of a 10,000-marker ultradense genetic recombination map of potato: providing a framework for accelerated gene isolation and a genomewide physical map. **Genetics**, v.173, n.2, Jun, p.1075-87. 2006.

VERHAEGEN, D., PLOMION, C., GION, J.M.; POITEL, M., COSTA, P., KREME,A. Quantitative trait dissection analysis in Eucalyptus using RAPD markers: I. Detection of QTL in Interespecific hybrid progeny, stability of QTL, expression across different ages. **Theor appl Genet**, v.95:p. 597-608, 1997.

WAGNER, F. R. Tópicos Especiais III - Simulação**.** Disponível em: **<**http://www.inf.ufrgs.br/~flavio>. Acesso em: 23 fev. 2008.

WEEKS, D.E. and LANGE, K. Preliminary randing procedures for multilocus ordering. **Genomics**, 1, 236-242, 1987.

WU, J.; J.JEMKINS; J. ZHU; J. MCCARTY, Jr.; C.WATSON. Monte Carlo simulations on marker grouping and ordering. **Theor Appl Genet**, v.107, n.3, Aug, p.568-73. 2003.

WU, R.L.; Ma, C.X.; WU, S.S.; ZENG, Z.B.;(2002) Linkage mapping of sex-specific differences. **Genetic Res** 79:85-96.

YAN, Z., DENNEBOOM, C., HATTENDORF, A., DOLSTRA, O., DEBENER, T., STAM P., VISSER, P.B. (2005) Construction of an integrated map of rose with AFLP, SSR, PK, RGA, RFLP, SCAR and morphological markers. **Theoretical and applied Genetics**, 110: 766-777.

YIN, T.M.; DIFAZIO, S.P.; GUNTER, L.E. REIMENSCHNEIDER, D.; TUSKAN, G.A. 2004, Large-scale heterospecific segregation distortion in *Populus* revealed by a dense genetic map. **Theoretical and Applied Genetics** 109: 451-463.

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**, 1994. p. 39-57.

ZHONG, X.; FRANSZ, P.; VAN WENNEKES, E.; ZABEL, P.; VAN KAMMEN, A.; DE JONG, J. (1996) High-resolution mapping on pachytene chromosomes and extended DNA fibres. **Genomics**, p.1075-87.

ZHOU, Y; GWAZE, D.P.; REYES-VALDES, M.H.; BUI, T.; WILLIANS, C. G. No clustering for linkage map based on low-copy and undermethylated microssatellite. **Genome** 46: 809-816, 2003.