ASPECTOS BIOMÉTRICOS DA DETECÇÃO DE QTL'S ("QUANTITATIVE TRAIT LOCI") EM ESPÉCIES CULTIVADAS

HEYDER DINIZ SILVA

Tese apresentada à Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, para a obtenção do título de Doutor em Agronomia, Área de Concentração: Estatística e Experimentação Agronômica.

PIRACICABA
São Paulo - Brasil
Outubro - 2001

ASPECTOS BIOMÉTRICOS DA DETECÇÃO DE QTL'S
("QUANTITATIVE TRAIT LOCI") EM ESPÉCIES CULTIVADAS

HEYDER DINIZ SILVA

Engenheiro Agrônomo

Orientador: Prof. Dr. Roland Vencovsky

Tese apresentada à Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, para a obtenção do título de Doutor em Agronomia, Área de Concentração: Estatística e Experimentação Agronômica.

PIRACICABA
São Paulo - Brasil
Outubro - 2001

Dados Internacionais de Catalogação na Publicação (CIP)
DIVISÃO DE BIBLIOTECA E DOCUMENTAÇÃO - ESALQ/USP

Silva, Heyder Diniz

Aspectos biométricos da detecção de QTL'S ("Quantitative Trait Loci")
em espécies cultivadas / Heyder Diniz Silva. - - Piracicaba, 2001.
166 p.

Tese (doutorado) - Escola Superior de Agricultura Luiz de Queiroz, 2001.
Bibliografia.

1. Biometria 2. Genética estatística 3. Genética quantitativa 4. Marcador
genético 5. Método estatístico 6. Milho I. Título

CDD 575.10212

“Permitida a cópia total ou parcial deste documento, desde que citada a fonte – O autor”

AGRADECIMENTOS

Gostaria de agradecer a todos aqueles que, direta ou indiretamente, contribuíram para realização desta tese, reservo-me, contudo, o direito de não citar nomes, porém, dois não posso omitir:

A Deus, fonte de toda sabedoria e conhecimento;

Ao Prof. Roland Vencovsky, um exemplo a ser seguido;

SUMÁRIO

	Página
RESUMO	vii
SUMMARY	ix
1 INTRODUÇÃO	1
2 REVISÃO DE LITERATURA	4
2.1 Uso de Marcadores no Melhoramento Vegetal	4
2.2 Base genética	9
2.2.1 Desequilíbrio de ligação	9
2.2.2 Ligação e recombinação	11
2.2.3 Funções de mapeamento	13
2.3 Métodos de Mapeamento de QTLs	17
2.3.1 Análises de marcas simples	18
2.3.2 Mapeamento por Intervalo	32
2.3.3 Mapeamento por Intervalo Composto	38
2.3.4 Mapeamento Por Intervalo Múltiplo	41
2.3.5 Mapeamento de Múltiplos Caracteres	45
2.4 Testes Múltiplos e seleção de marcas associadas a QTL's	47
2.5 Interação QTLs x Ambiente	58
3 TESTES MÚLTIPLOS E PODER DE DETECÇÃO DE QTL's	61
Resumo	61
Summary	62

3.1	Introdução	62
3.2	Metodologia	73
3.3	Resultados e Discussão	76
3.4	Conclusões	84
4	PARTIÇÃO DA VARIÂNCIA DA INTERAÇÃO GENÓTIPOS X AMBIENTES EM EFEITOS EXPLICADOS PELOS MAR- CADORES E DESVIOS	86
	Resumo	86
	Summary	87
4.1	Introdução	87
4.2	Metodologia	88
4.3	Resultados e discussão	91
4.3.1	Modelo matemático e partição das somas de quadrados	91
4.3.2	Esperanças matemáticas das somas de quadrados e dos quadrados médios.	95
4.3.3	Estimadores dos componentes de variância	100
4.3.4	Testes de hipóteses.	101
4.3.5	Estimadores das proporções da variância genética e da variância da interação genótipos x ambientes explicadas pelos marcadores moleculares	110
4.4	Conclusões	119
5	ANÁLISE DA INTERAÇÃO GENÓTIPOS X AMBIENTES EM MILHO, UTILIZANDO MARCADORES MOLECULARES CODOMINANTES	120
	Resumo	120
	Summary	121
5.1	Introdução	121
5.2	Material e Métodos	123
5.2.1	Material	123
5.2.2	Métodos	123

	vi
5.3 Resultados e Discussao	126
5.4 Conclusões	141
6 CONCLUSÕES	143
APÊNDICE	152

ASPECTOS BIOMÉTRICOS DA DETECÇÃO DE QTL'S ("QUANTITATIVE TRAIT LOCI") EM ESPÉCIES CULTIVADAS

Autor: HEYDER DINIZ SILVA

Orientador: Prof. Dr. ROLAND VENCovsky

RESUMO

O mapeamento de QTL's difere dos demais tipos de pesquisas conduzida em genética. Por se tratar basicamente de um procedimento de testes múltiplos, surge, neste contexto, um problema que se refere ao nível de significância conjunto da análise, e consequentemente, seu poder. Deste modo, avaliou-se, via simulação computacional de dados, o poder de detecção de QTL's da análise de marcas simples, realizada por meio de regressão linear múltipla, utilizando o procedimento "stepwise" para seleção das marcas e procedimentos baseados em testes individuais, utilizando os critérios FDR e de Bonferroni para determinação nível de significância conjunto. Os resultados mostraram que o procedimento baseado em regressão múltipla, utilizando o procedimento "stepwise" foi mais poderoso em identificar as marcas associadas a QTL's e, mesmo nos casos em que este procedimento apresentou poder ligeiramente inferior aos demais, verificou-se que o mesmo tem como grande vantagem selecionar apenas as marcas mais fortemente ligadas

aos QTL's. Dentre os critérios FDR e de Bonferroni, o primeiro mostrou-se, em geral, mais poderoso, devendo ser adotado nos procedimentos de mapeamento por intervalo. Outro problema encontrado na análise de QTL's refere-se à abordagem da interação QTL's x ambientes. Neste contexto, apresentou-se uma partição da variância da interação genótipos x ambientes em efeitos explicados pelos marcadores e desvios, a partir da qual obtiveram-se os estimadores da proporção da variância genética (p_m), e da variância da interação genótipos x ambientes (p_{ms}), explicadas pelos marcadores moleculares. Estes estimadores independem de desvios das frequências alélicas dos marcadores em relação às esperadas (1:2:1 em uma geração F_2 , 1:1 em um retrocruzamento, etc.), porém, apresentam uma alta probabilidade de obtenção de estimativas fora do intervalo paramétrico, principalmente para valores elevados destas proporções. Contudo, estas probabilidades podem ser reduzidas com o aumento do número de repetições e/ou de ambientes nos quais as progênies são avaliadas. A partir de um conjunto de dados de produtividade de grãos, referentes à avaliação de 68 progênies de milho, genotipadas para 77 marcadores moleculares codominantes e avaliadas em quatro ambientes, verificou-se que as metodologias apresentadas permitiram estimar as proporções p_m e p_{ms} , bem como classificar as marcas associadas a QTL's, conforme seu nível de interação. O procedimento permitiu ainda a identificação de regiões cromossômicas envolvidas no controle genético do caractere sob estudo conforme sua maior ou menor estabilidade ao longo dos ambientes.

BIOMETRICAL ASPECTS OF QTL DETECTION IN CULTIVATED SPECIES

Author: HEYDER DINIZ SILVA

Adviser: Prof. Dr. ROLAND VENCovsky

SUMMARY

In general terms, QTL mapping differs from other research activities in genetics. Being basically a multiple test procedure, problems arise which are related to the joint level of significance of the analysis, and consequently, to its power. Using computational simulation of data, the power of simple marker analysis, carried out through multiple linear regression, using stepwise procedures to select the markers was obtained. Procedures based on single tests, using both the FDR and the Bonferroni criteria to determinate the joint level of significance were also used. Results showed that the procedure based on multiple regression, using the stepwise technique, was the most powerful in identifying markers associated to QTL's. However, in cases where its power was smaller, its advantage was the ability to detect only markers strongly associates with QTL's. In comparision with the Bonferroni method, the FDR criterion was in general more powerful, and should be adopted in the interval mapping procedures. Additional problems found in the QTL

analysis refer to the QTL x environment interaction. We consider this aspect by partitioning the genotype x environment interaction variance in components explained by the molecular markers and deviations. This allowed estimating the proportion of the genetic variance (p_m), and genotype x environment variance (p_{ms}), explained by the markers. These estimators are not affected by deviations of allelic frequencies of the markers in relation to the expected values (1:2:1 in a F2 generation, 1:1 in a backcross, etc). However, there is a high probability of obtaining estimates out of the parametric range, specially for high values of this proportion. Nevertheless, these probabilities can be reduced by increasing the number of replications and/or environments where the progenies are evaluated. Based on a set of grain yield data, obtained from the evaluation of 68 maize progenies genotyped for 77 codominant molecular markers, and evaluated as top crosses in four environments, the presented methodologies allowed estimating proportions p_m and p_{ms} as well the classification of markers associated to QTL's, with respect to its level of genotype x environment interaction. The procedure also allowed the identification of chromosomic regions, involved in the genetical control of the considered trait, according to its stability, in relation to the observed environmental variation.

1 INTRODUÇÃO

Desde o início da história, o melhoramento vegetal e animal baseou-se na seleção de indivíduos com fenótipos desejáveis para serem os genitores da próxima geração. Os métodos clássicos de melhoramento fundamentam-se no fato de que o fenótipo é resultado da ação conjunta e independente do genótipo e do ambiente ($F = G + E$), expressa em termos de suas variâncias por $\sigma_F^2 = \sigma_G^2 + \sigma_E^2$. Quantificando-se esses componentes, pode-se inferir sobre a viabilidade de programas de melhoramento, prever ganhos a serem obtidos com a seleção e, então, optar por estratégias que maximizem esses ganhos. Comparando as populações domésticas com as selvagens, verifica-se que a seleção artificial obteve grandes sucessos na alteração dos fenótipos. Porém os materiais genéticos disponíveis para o melhoramento em algumas espécies, como o milho e a soja, por exemplo, já foram muito melhorados em relação a seus ancestrais selvagens e, atualmente, apenas pequenos ganhos estão sendo obtidos, pois a variabilidade genética explorada nesses materiais tornou-se pequena em muitos programas de melhoramento.

Com o advento das técnicas moleculares modernas de análise genômica, um novo horizonte abriu-se aos melhoristas, vislumbrando-se maiores possibilidades de ganhos, à medida que o controle genético dos caracteres quantitativos sob seleção é desvendado e incorporado aos métodos clássicos de melhoramento. Dentre as principais contribuições que os marcadores moleculares podem trazer ao melhoramento de caracteres quantitativos, podem-se listar: i) o estudo da diversidade e das distâncias genéticas entre materiais, visando à escolha de progenitores, no intuito de monitorar a variabilidade genética disponível para o melhoramento; ii) a identificação

e a localização de QTL's ("Quantitative Trait Loci") com grandes efeitos individuais, podendo-se, deste modo, selecionar os indivíduos com base nas informações diretas de seu genótipo; iii) incorporação das informações quanto ao genótipo de um indivíduo, obtidas via marcadores moleculares, aos métodos clássicos de melhoramento, chamada seleção assistida por marcadores ("Marker Assisted Selection"-MAS); iv) a seleção precoce em espécies perenes, para as quais são necessários vários anos até que entrem na fase produtiva e possam, então, ser avaliadas quanto a caracteres quantitativos, tais como a produção de frutos, o rendimento e a qualidade da madeira etc. De posse do conhecimento de quais genes afetam esses caracteres, a seleção pode ser realizada nos estádios juvenis da planta, mediante sua genotipagem para os genes de interesse; v) obtenção de um melhor conhecimento das correlações genéticas entre caracteres.

Na análise de QTLs, existem alguns problemas que ainda se encontram a descoberto, dentre os quais, podem-se destacar:

- A existência da interação QTLs x ambientes, expressa, principalmente, pela inconstância dos efeitos dos QTLs, quando avaliados em vários ambientes, chegando ao extremo de uma marca identificada como estando associada a QTLs em um ambiente não ser detectada em outro.

- O nível de significância conjunto da análise, uma vez que esta se caracteriza por um procedimento de testes múltiplos, dos quais, de acordo com Liu (1998), alguns podem ser independentes entre si, e outros não. Tal fato torna a abordagem clássica para determinação do nível de significância conjunto inapropriada para experimentos de mapeamento de QTLs.

Portanto, objetivou-se nesta tese:

- Avaliar, via simulação computacional de dados, duas metodologias baseadas em testes múltiplos, e uma baseada em regressão múltipla, quanto ao poder de detecção de marcas associadas a QTL's;
- Apresentar uma partição da variância da interação genótipos x ambientes em

efeitos explicados pelos marcadores e desvios

- Estender o estimador p da proporção da variância genética explicada pelos marcadores moleculares, apresentado por Bearzoti & Vencovsky (1998), para experimentos conduzidos em um único ambiente, a situação em que os genótipos são avaliados em vários ambientes;
- Propor um estimador da proporção da variância devida à interação genótipos x ambientes ($G \times E$), que é devida à interação QTLs x ambientes ($QTL \times E$);
- Aplicar as metodologias desenvolvidas a um conjunto de dados reais, referentes à avaliação de 67 progênies de milho, genotipadas para 77 marcadores moleculares codominantes e avaliadas em quatro ambientes.

2 REVISAO DE LITERATURA

2.1 Uso de Marcadores no Melhoramento Vegetal

Mendel assentou os fundamentos da genética por meio do estudo das diferenças que dividiam as suas ervilhas em categorias claramente distintas. Assim, nunca havia dúvida quanto às sementes produzidas por suas plantas serem lisas ou enrugadas, ou as flores vermelhas ou brancas, e assim por diante: as categorias não se sobrepunham. Ele foi capaz de mostrar que cada classe fenotípica correspondia a um, ou, em todo o caso, a somente poucos genótipos e que, quando havia mais de um genótipo na classe fenotípica, eles podiam ser separados por testes apropriados de cruzamento, isto é, pelas classes de plantas produzidas após acasalamentos e testes apropriados. Foi, assim, capaz de inferir os genes, ou fatores, como os chamava, de cujos comportamentos depende a transmissão hereditária.

A maioria dos caracteres agrônômicos de interesse, tais como a produtividade de grãos, altura de plantas, concentração de sólidos solúveis, produtividade de biomassa, rendimento de madeira, dentre outros, não apresentam classes fenotípicas claramente distintas, como os estudados por Mendel, isto é, apresentam uma variação contínua, sendo denominados, por isso, caracteres quantitativos. Tomando como exemplo o peso das espigas produzidas por uma planta de milho, vemos que o número de classes nas quais os indivíduos podem ser agrupados, de acordo com a manifestação desse caráter, é limitado unicamente pela exatidão da medida que se pode fazer. Portanto, esses caracteres não se prestam para uma análise das causas da variação da maneira como Mendel mostrou ser possível para a variação

descontínua ou discreta.

Dada a natureza contínua da variação dos caracteres quantitativos, o seu estudo deve começar pela distribuição de frequências originada pelo caráter, quando observado um certo número de indivíduos. Tal distribuição é caracterizada por certos parâmetros estatísticos, dos quais a média e a variância são os mais importantes, aos quais se podem acrescentar covariâncias e correlações. Fisher (1918) mostrou que os achados biométricos não apenas podiam ser facilmente interpretados, mas, de fato, em alguns casos, exigiam virtualmente uma interpretação em termos dos fatores de Mendel, então já chamados de genes, e cujos portadores eram reconhecidamente os cromossomos. Desse modo, Fisher forneceu as bases para o surgimento da genética quantitativa.

O primeiro grande princípio da genética é que o fenótipo é resultante do genótipo do indivíduo e do meio-ambiente no qual ele se desenvolve e passa a sua vida. De fato, Johanssen (1909)¹, citado por Mather & Jinks (1984), demonstrou, em seus experimentos com feijão, que fatores ambientais eram a maior fonte de variação em caracteres quantitativos, levando à conclusão de que, para esse tipo de caráter, o fenótipo não é um bom indicador do genótipo.

Os genes ou locos cromossômicos que controlam uma determinada característica quantitativa são denominados QTLs ("Quantitative Trait Loci"). Segundo Falconer & Mackay (1996), um QTL é um segmento de cromossomo que afeta uma determinada característica, não sendo necessariamente um único gene.

Usando *Drosophila*, Morgan (1919) indicou que os genes Mendelianos são ligados e podem ser mapeados em grupos de ligação. Sax (1923) demonstrou que o efeito de um loco ou gene individual sobre um caráter quantitativo podia ser identificado, com a ajuda de um gene marcador, a partir de uma série de cruzamentos. Os genes marcadores utilizados nesse trabalho eram genes associados à coloração de sementes de feijão, sendo, portanto, denominados marcadores morfológicos.

Os delineamentos experimentais usados para mapear e estimar os

¹JOHANNSEN, W. Element der exacten erblichkeitslehre. FISHER, JENA

efeitos de QTLs são extensões dos métodos clássicos de mapeamento de genes simples, sendo fundamentados no desequilíbrio de ligação entre os "alelos" de um loco marcador e os do QTL. Para se mapear um QTL, é necessário um mapa de ligação de marcadores polimórficos, que cubram todo o genoma, e que exista variabilidade para o caráter quantitativo que se deseja estudar. As características ideais que locos marcadores devem possuir são:

- (i) serem altamente polimórficos;
- (ii) cobrirem todo o genoma;
- (iii) serem seletivamente neutros, no sentido de não afetarem o caráter em questão nem as características reprodutivas do indivíduo e
- (iv) serem co-dominantes para que todos os possíveis genótipos dos locos marcadores possam ser identificados.

Apesar de toda a teoria necessária para o estudo de QTLs estar disponível desde a década de 20, até meados da década de 60, obteve-se pouco sucesso na identificação de marcadores ligados a QTLs, pois os marcadores utilizados nesse período eram controlados por genes associados a caracteres morfológicos, em geral, fenótipos de fácil identificação visual, como nanismo, deficiência clorótica, cor da pétala, grupo sanguíneo etc, mas, que são marcadores pouco polimórficos, não cobrem todo o genoma, além de não serem seletivamente neutros.

Segundo Ferreira & Grattapaglia (1998), a revolução nesse quadro iniciou-se com o desenvolvimento de marcadores isoenzimáticos, o que ampliou o número de marcadores genéticos disponíveis, e com o advento das técnicas modernas de biologia molecular, que permitiram o surgimento de diversos métodos de detecção de polimorfismo ao nível de DNA. Os marcadores genéticos obtidos com base em fenótipos moleculares são denominados marcadores moleculares.

De acordo com Milach (1998), os distintos tipos de marcadores moleculares hoje disponíveis diferenciam-se pela tecnologia utilizada para revelar variabilidade ao nível de DNA, podendo ser classificados em dois grupos, conforme a metodologia utilizada para identificá-los: hibridização ou amplificação de DNA.

Entre os identificados por hibridizao, esto os marcadores RFLP ("Restriction Fragment Length Polymorphism") e Ministites ou locos VNTR ("Variable Number of Tandem Repeats"). J aqueles revelados por amplificao incluem os marcadores do tipo: RAPD ("Random Amplified Polymorphic DNA"); SCAR ("Sequence Characterized Amplified Regions"); STS ("Sequence Tagged Sites"); microsstites ou SSR (Single Sequence Repeats") e AFLP ("Amplified Fragment Length Polymorphism").

A Tabela 1, adaptada de Ferreira & Grattapaglia (1998), fornece algumas caractersticas dos principais tipos de marcadores moleculares utilizados no melhoramento vegetal:

Tabela 1. Anlise comparativa dos principais tipos de marcadores moleculares usados na anlise gentica de plantas

Vari�vel	Tipo de marcador				
	Isoenzimas	RFLP	RAPD	micross�t�ites	AFLP
Express�o g�nica	codominante	codominante	dominante	codominante	dominante
N�mero de alelos/loco	2 a 4 alelos	multial�lico	2 alelos	multial�lico	2 alelos
Disponibilidade de marcadores no genoma	20 a 50	ilimitada	ilimitada	ilimitada	ilimitada
Distribui��o no genoma	regi�es de c�pia �nica	regi�es de baixo n�mero de c�pias	mais ou menos ao acaso	mais ou menos ao acaso	mais ou menos ao acaso

A deteco de ligao gentica entre marcadores e QTLs depende da existncia de desequilbrio de ligao entre os alelos do loco marcador e do QTL. Esse desequilbrio gera efeitos quantitativos associados ao marcador, que podem ser detectados e estimados por meio de anlises estatsticas adequadas.

O procedimento de procura, localização e estimação dos efeitos de QTLs é denominado de mapeamento de QTLs. Os delineamentos usuais utilizados para tal, nas principais espécies cultivadas, baseiam-se no uso de populações segregantes, F_2 , ou retrocruzamentos, as quais são obtidas pelo cruzamento de duas linhagens homozigóticas fenotipicamente divergentes quanto ao caráter quantitativo em questão, existindo outros procedimentos, utilizados em espécies perenes e animais, para os quais a utilização desses cruzamentos controlados é inviável, seja pelo longo intervalo entre as gerações ou por questões éticas.

Um determinado número de indivíduos dessas populações (geralmente, algumas centenas) são avaliados para os caracteres quantitativos de interesse, e genotipados para uma série de marcadores moleculares, distribuídos a intervalos regulares pelo genoma. Se houver uma diferença na média do fenótipo entre as classes genotípicas do marcador, pode-se inferir que existe um QTL ligado a aquele marcador. Falconer & Mackay (1996) afirmam que o número de QTLs detectados é sempre uma subestimação do número real de locos, devido ao fato de, dois QTLs, muito próximos um do outro, poderem aparecer como apenas um, se estiverem em associação (ligação entre alelos favoráveis ao caráter nos dois locos), ou não serem detectados se em repulsão (ligação entre um alelo favorável em um loco e um desfavorável no outro). Porém, como são utilizadas técnicas estatísticas para verificar a ligação entre o QTL e o marcador, nesse procedimento, está-se sujeito ao erro tipo I que, no caso, equivale à identificação de um QTL inexistente, também chamado QTL "fantasma", e, desse modo, a afirmação de Falconer & Mackay (1996) pode nem sempre ser verdadeira.

2.2 Base genética

2.2.1 Desequilíbrio de ligação

Desequilíbrio de ligação é definido como sendo qualquer desvio das frequências alélicas em relação às frequências esperadas sob independência, indicando a existência de uma associação entre os dois locos. Deste modo, para que os marcadores moleculares possam contribuir nos estudos de herança e no melhoramento, é necessário que a população esteja em desequilíbrio de ligação, pois, caso contrário, a probabilidade de ocorrência de determinada classe de marcador seria independente da segregação dos alelos do QTL em questão. Em vista disso, é que se utilizam, principalmente no melhoramento vegetal, os cruzamentos controlados, chamados delineamentos genéticos, que visam maximizar o desequilíbrio de ligação.

Uma medida do desequilíbrio de ligação é apresentada por Crow & Kimura (1970) como sendo:

$$z = \frac{P_{MQ}P_{mq}}{P_{Mq}P_{mQ}}, \quad (1)$$

em que P_{MQ} , P_{mq} , P_{Mq} e P_{mQ} são as frequências dos gametas MQ, mq, Mq e mQ, respectivamente.

Sob independência, o valor de z é igual a 1 (um) e qualquer desvio desse valor indica a existência de desequilíbrio de ligação.

Weir (1996) define outra medida do desequilíbrio genético entre dois locos como:

$$D_{MQ} = P_{MQ} - P_M \cdot P_Q \quad (2)$$

em que P_{MQ} é a frequência do gameta MQ, e P_M e P_Q as frequências dos alelos M e Q respectivamente. O estimador de máxima verossimilhança é:

$$\hat{D}_{MQ} = P_{MQ} - P_M \cdot P_Q; \quad (3)$$

que é obtido diretamente das frequências, tendo

$$E[\hat{D}_{MQ}] = \frac{2n-1}{2n} D_{MQ}$$

e

$$Var[\hat{D}_{MQ}] = \frac{1}{2n} [P_M(1-P_M)P_Q(1-P_Q) + (1-2P_M)(1-2P_Q)D_{MQ} - D_{MQ}^2]$$

com n sendo o número de indivíduos;

No caso, uma estatística Qui-quadrado para o teste da hipótese $H_0 : D_{MQ} = 0$ (ausência de desequilíbrio) pode ser obtida elevando-se ao quadrado a variável z:

$$z = \frac{\hat{D}_{MQ} - E[\hat{D}_{MQ}]}{\sqrt{Var[\hat{D}_{MQ}]}}; \quad (4)$$

a qual, sob H_0 é

$$\hat{A}_{MQ}^2 = z^2 = \frac{2n\hat{D}_{MQ}}{P_M(1-P_M)P_Q(1-P_Q)}; \quad (5)$$

Segundo Weir (1996), o desequilíbrio de ligação entre dois locos também pode ser avaliado por:

$$D_{MQ} = P_{MQ}P_{mq} - P_{Mq}P_{mQ}; \quad (6)$$

Esse coeficiente de desequilíbrio genético pode ser positivo ou negativo, sendo sua amplitude dependente das frequências alélicas nos dois locos:

$$\min[P_M(1-P_Q); (1-P_M)P_Q] > D_{MQ} > -1: \max[P_MP_Q; (1-P_M)(1-P_Q)]:$$

Segundo Ferreira e Grattapaglia (1998), a ligação física entre dois locos no mesmo cromossomo é apenas uma das causas de desequilíbrio de ligação, entretanto, em experimentos de mapeamento, uma associação significativa entre marcador e gene é invariavelmente interpretada como sendo evidência de ligação física.

2.2.2 Ligação e recombinação

Ligação gênica é definida como a associação entre genes localizados num mesmo cromossomo, sendo que, para esses genes, as proporções genotípicas esperadas após um determinado cruzamento afastam-se das proporções mendelianas, isto é, os genes tendem a ser herdados juntos. Assim sendo, suponha-se que existam em uma determinada população um gene marcador M, com os alelos M e m, e um QTL Q, com os alelos Q e q, estando, ambos, em um mesmo cromossomo e completamente ligados, isto é, não existe recombinação entre estes locos. Considerando que o QTL em questão apresenta os valores genotípicos esquematizados na figura 21, na qual a é o efeito aditivo do QTL, e d é o efeito de dominância, então, os valores genotípicos dos indivíduos $\frac{M \ Q}{M \ Q}$, $\frac{M \ Q}{m \ q}$ e $\frac{m \ q}{m \ q}$ serão +a, d e -a respectivamente.

Figura 1 - Representação de um QTL, com seus valores genotípicos

Desse modo, embora os genótipos do QTL não possam ser observados,

os valores de a e d podem ser estimados pelos seguintes contrastes entre as médias dos indivíduos MM , Mm e mm :

$$a = \frac{\bar{y}_{MM} - \bar{y}_{mm}}{2} \quad (7)$$

$$d = \frac{(2\bar{y}_{Mm} - \bar{y}_{MM} - \bar{y}_{mm})}{2} = \bar{y}_{Mm} - \left(\frac{\bar{y}_{MM} + \bar{y}_{mm}}{2} \right) \quad (8)$$

Porém, durante a meiose, as cromátides homólogas sofrem um processo de quebra e reunião, e, se no processo de reunião ocorrer troca de partes de cromossomos entre cromátides não irmãs, ocorre o fenômeno chamado "crossing over" ou permuta gênica. O resultado dessa recombinação é o aparecimento de gametas não parentais, ou recombinantes, a uma determinada frequência c , chamada frequência ou taxa de recombinação, como esquematizado na Figura 2, na qual os gametas MQ e mq são os tipos parentais enquanto Mq e mQ são recombinantes.

Figura 2 - Ilustração do fenômeno da recombinação entre dois alelos

Deste modo, a geração F_2 desse cruzamento será composta pelos genótipos apresentados na Tabela 2, e neste caso, o contraste $\bar{y}_{MM} - \bar{y}_{mm} = \frac{1}{2}a(1-2c)$: Assim, verifica-se que a estimativa do valor do efeito aditivo (a), associado ao marcador (M), decresce linearmente com o aumento da frequência de recombinação entre o loco marcador e o QTL (c).

Tabela 2. Frequências genotípicas esperadas numa geração F_2 , considerando dois locos ligados em fase de associação, com uma frequência de recombinação c

Genótipo	Frequência	Valor Genotípico	Média da classe
MQ/MQ	$\frac{1}{4}(1 - c)^2$	+a	
MQ/Mq	$\frac{c}{2}(1 - c)$	d	$\frac{1}{4}a(1 - 2c) + \frac{d}{2}c(1 - c)$
Mq/Mq	$\frac{c^2}{4}$	-a	
MQ/mQ	$\frac{c}{2}(1 - c)$	+a	
$\left\{ \begin{array}{l} \text{MQ/mq} \\ \text{Mq/mQ} \end{array} \right\} \text{MmQq}$	$\frac{1}{2}(1 - 2c + 2c^2)$	d	$\frac{d}{2}(1 - 2c + 2c^2)$
Mq/mq	$\frac{c}{2}(1 - c)$	-a	
mQ/mQ	$\frac{c^2}{4}$	+a	
mQ/mq	$\frac{c}{2}(1 - c)$	d	$-\frac{1}{4}a(1 - 2c) + \frac{d}{2}c(1 - c)$
mq/mq	$\frac{1}{4}(1 - c)^2$	-a	

2.2.3 Funções de mapeamento

Mapear um QTL implica estimar seus efeitos (a e d) e posição no genoma (geralmente, em relação a um marcador conhecido). Um problema de utilizar a frequência de recombinação como medida de distância entre genes, para a construção de mapas de ligação, é que esta não é uma escala aditiva, como pode ser verificado a seguir:

Considere-se o seguinte segmento cromossômico, com três genes (A, B e C), cujas frequências de recombinação entre estes são $c_{A;B} = 0,12$ e $c_{B;C} = 0,35$

$$\underline{\quad A \quad 0,12 \quad B \quad 0,35 \quad C \quad}$$

Desse modo, a probabilidade de recombinação entre A e C, admitindo a ausência de interferência, isto é, que os "crossing overs" sejam independentes, será:

$$c_{A;C} = c_{A;B} + c_{B;C} - 2c_{A;B}c_{B;C} \quad (9)$$

Substituindo os valores de $c_{A;B}$ e $c_{B;C}$ em 9, tem-se que: $c_{A;C} = 0,12 + 0,35 - 2(0,12)(0,35) = 0,386$, valor que é diferente da soma $0,12 + 0,35 = 0,47$, mostrando que a escala não é aditiva, ou seja, $c_{A;C} \leq c_{A;B} + c_{B;C}$.

De acordo com Bearzoti (2000), para solucionar esse problema, propõe-se outra escala (μ), definida como o número médio de pontos de permuta no segmento em questão. A unidade dessa escala foi denominada Morgan (M). Dessa forma, um Morgan (1M) corresponde a um segmento cromossômico no qual ocorre, em média, uma permuta gênica por gameta por geração, sendo muito comum a utilização da unidade cM (centimorgan). Na prática, o que realmente interessa saber, ao se consultar um mapa de ligação, é a frequência de recombinação (c) entre dois locos. Portanto, é preciso que exista uma relação que permita obter c a partir de μ . Essa regra de associação é chamada de função de mapeamento, sendo que as funções de mapeamento de Haldane (Haldane, 1919) e de Kosambi (Kosambi, 1944) são as mais utilizadas.

Para obter a função de mapeamento de Haldane, são admitidas duas pressuposições básicas:

- As ocorrências de permutas gênicas são independentes (não há interferência).

- As permutas gênicas ocorrem ao acaso, ao longo do cromossomo.

Deste modo, o número de permutas gênicas pode ser modelado segundo uma distribuição de Poisson.

Para que haja a recombinação entre dois locos separados por uma distância μ , é necessário que ocorra um número ímpar de permutas gênicas, no segmento cromossômico entre eles. Assim, a frequência de recombinação (c) corresponde à probabilidade de ocorrência de um número ímpar de permutas gênicas no intervalo μ . Como o número de "crossing overs", em um segmento de tamanho μ , pode ser modelado segundo a distribuição de Poisson, com média μ , tem-se:

$$\begin{aligned}
c &= P(\text{número ímpar de permutas gênicas}) \\
&= \sum_{y \text{ ímpar}} \frac{e^{-c} c^y}{y!} \\
&= e^{-c} \left[\frac{c}{1!} + \frac{c^3}{3!} + \frac{c^5}{5!} + \dots \right] \\
&= e^{-c} \frac{(e^c - e^{-c})}{2} \\
c &= \frac{1}{2}(1 - e^{-2c}) \tag{10}
\end{aligned}$$

cujas função inversa é:

$$c = -\frac{1}{2} \ln(1 - 2c) \tag{11}$$

Aplicando-se a função de mapeamento de Haldane ao exemplo anterior, tem-se:

$$c_1 = 0;12 \Rightarrow c_1 = 0;1372M(13;72cM);$$

$$c_2 = 0;35 \Rightarrow c_2 = 0;6020M(60;20cM);$$

Nota-se aqui que: $c_1 + c_2 = 0;7392M \Rightarrow c = 0;386$. Mostrando a aditividade da escala.

Segundo Liu (1998), desvios em relação aos valores obtidos pela equação 9 têm sido observados. Esses desvios são atribuídos ao fenômeno da interferência, que se caracteriza pelo fato de a ocorrência de um "crossing over" em uma região diminuir a probabilidade de ocorrência de "crossing over" em uma região adjacente. Assim sendo, uma fórmula mais geral, capaz de contemplar o fenômeno da interferência, é:

$$c_{A:C} = c_{A:B} + c_{B:C} - 2(1 - \pm)c_{A:B}c_{B:C}; \tag{12}$$

em que $\pm \theta$ um parâmetro que varia de zero, se os "crossing over" forem independentes (ausência de interferência) a um, quando a ocorrência de um "crossing over" em uma dada região impedir a ocorrência de outro em uma região adjacente (interferência completa).

Segundo Lynch & Walsh (1998), existem várias funções de mapeamento capazes de contemplar o fenômeno da interferência, sendo que a mais utilizada pelos geneticistas é a função de mapeamento de Kosambi (Kosambi, 1944):

$$c = \frac{1}{2} \left(\frac{e^{4\theta} - 1}{e^{4\theta} + 1} \right) = \frac{1}{2} \tanh(2\theta); \quad (13)$$

em que $\tanh$ é a tangente hiperbólica, e cuja função inversa é:

$$\theta = \frac{1}{4} \ln \left(\frac{1 + 2c}{1 - 2c} \right); \quad (14)$$

Bearzoti (2000) e Lynch & Walsh (1998) comparam as funções de mapeamento de Haldane e de Kosambi e afirmam que, para pequenos valores de c ($c < 0,15$), ambas as funções fornecem estimativas de $\theta \simeq c$, diferindo para frequências de recombinação maiores e tornando-se a aproximar para valores altos de c ($c \simeq 0,5$), como pode ser observado na Figura 3, na qual estão apresentadas as distâncias em Morgans, obtidas pelas funções de mapeamento de Haldane (11) e Kosambi (14) para frequências de recombinação variando de 0 a 0,5.

Lynch & Walsh (1998) comentam ainda que não existe uma relação universal entre distância de mapeamento (M) e o tamanho físico do segmento de DNA (em pares de bases), pois um centi-Morgan pode variar de 10000 a um milhão de nucleotídeos (10Kb a 1000Kb), dependendo da espécie, podendo, inclusive, existir grandes diferenças entre segmentos de um mesmo cromossomo.

Figura 3 - Distâncias em Morgans, obtidas pelas funções de mapeamento de Haldane e Kosambi para frequências de recombinação variando de 0 a 0,5.

2.3 Métodos de Mapeamento de QTLs

A associação entre marcador e QTL pode ser avaliada utilizando-se uma, duas ou mais marcas simultaneamente. Na análise de marcas simples, a distribuição dos valores do caráter é examinada separadamente para cada loco marcador. Assim, cada teste é realizado independentemente da informação dos outros locos e, deste modo, um cromossomo com t marcas fornece t testes de marcas simples. Esse tipo de análise é, geralmente, uma boa opção quando o objetivo é simplesmente a detecção de QTLs ligados a uma marca em detrimento da estimação da posição e dos efeitos do mesmo.

No mapeamento por intervalo, proposto por Lander & Botstein (1989), as análises são realizadas considerando cada par de marcas adjacentes, resultando então em $t-1$ testes de associação QTL-marcador por cromossomo com t marcas (um para cada intervalo). Essa metodologia de mapeamento oferece um incremento no poder de detecção de QTLs e estimativas mais precisas da posição e dos efeitos do mesmo.

Tanto o mapeamento por intervalo quanto as análises de marcas simples são viesadas quando múltiplos QTLs estão ligados ao intervalo/marca considerado. Em vista disso, Jansen (1992, 1993) e Zeng (1993, 1994) propuseram a técnica denominada Mapeamento por Intervalo Composto, que combina, em um único modelo, as técnicas de mapeamento por intervalo e regressão múltipla, no intuito de eliminar as influências de outros QTLs sobre o intervalo considerado.

2.3.1 Análises de marcas simples

A identificação de QTLs ligados a marcadores moleculares, utilizando marcas simples ("Single Marker Analysis") é fundamentada na comparação entre as médias das classes genotípicas dos marcadores, podendo ser realizada por meio dos seguintes testes:

- Teste t
- Regressão linear simples
- Análise de variância
- Razão de verossimilhança (LOD score)

Uma diferença significativa entre as médias das classes genotípicas dos marcadores indica a possibilidade de existência de pelo menos um QTL associado a aquele marcador. As análises de marcas simples são muito fáceis do ponto de vista de análise de dados e implementação, pois não apresentam aspectos computacionais complicados e não é necessário o conhecimento da posição dos marcadores no genoma (mapa de marcadores). Porém, Liu (1998) cita as seguintes desvantagens desse método:

- O efeito de um determinado QTL fica confundido com sua localização. Ou seja, não é possível distinguir um QTL de pequeno efeito, situado muito próximo ao marcador de um QTL de grande efeito, situado mais longe.

- A posição do QTL não pode ser precisamente determinada, devido à não independência entre os testes de hipóteses para ligação, que confundem o efeito do QTL e sua posição.

Esses fatos limitam a técnica de análise de marcas simples mais a detecção do que ao mapeamento de QTLs, pois, por mapeamento de QTLs, entende-se a estimação de seus efeitos e posição no genoma. Contudo, a Análise de Marcas Simples é muito utilizada, principalmente, como um passo preliminar para a utilização de técnicas mais complexas, mediante de uma varredura de todas as marcas existentes no genoma e eliminação daquelas chamadas não informativas, o que certamente contribui em muito para o aumento da eficiência e, muitas vezes, para a exequibilidade das técnicas mais rebuscadas, como, por exemplo, o mapeamento por intervalo composto e múltiplo.

Teste t

Como discutido anteriormente, em uma geração F_2 , resultante do cruzamento entre duas linhagens homozigóticas contrastantes, o efeito aditivo de um QTL pode ser estimado pelo contraste entre as médias dos indivíduos portadores dos genótipos MM e mm ($\bar{y}_{MM}$ e $\bar{y}_{mm}$). Desse modo, caso a hipótese de igualdade entre essas médias seja rejeitada, pode-se inferir a existência de um QTL associado à marca em questão.

Denominando a média populacional dos indivíduos MM por μ_{MM} e a dos indivíduos mm por μ_{mm} , uma estatística t, para o teste das hipóteses:

$$\begin{cases} H_0 : \mu_{MM} - \mu_{mm} = 0 \\ H_a : \mu_{MM} - \mu_{mm} \neq 0 \end{cases}$$

é dada por:

$$t_{\text{calc}} = \frac{\bar{x}_{\text{MM}} - \bar{x}_{\text{mm}}}{\sqrt{s^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \quad (15)$$

em que:

n_1 e n_2 são o número de indivíduos com genótipos MM e mm, respectivamente e

s^2 é o estimador da variância comum a MM e mm.

Sob H_0 , $t_{\text{calc}} \sim t_{(n_1+n_2-2)}$.

Uma vez que a esperança matemática do contraste $\bar{x}_{\text{MM}} - \bar{x}_{\text{mm}}$ é $E[\bar{x}_{\text{MM}} - \bar{x}_{\text{mm}}] = \frac{1}{2}a(1 - 2c)$, depreende-se, que a verdadeira hipótese testada é que o efeito aditivo associado a marca é nulo, e/ou a frequência de recombinação entre o QTL e o marcador é igual a 0,5 (QTL e marcador segregam independentemente).

Regressão linear simples

Um modelo de regressão linear simples pode ser utilizado para identificar QTLs ligados a marcadores moleculares do seguinte modo:

Considere-se o modelo de regressão linear simples:

$$y_i = \beta_0 + \beta_1 X_i + e_i;$$

em que:

y_i é o valor fenotípico observado do i -ésimo indivíduo;

X_i é uma variável "dummy" codificadora dos genótipos do marcador.

Para o caso de uma geração F_2 , X_i pode assumir os valores 1, 0 ou -1, se o genótipo do i -ésimo indivíduo for MM, Mm ou mm, respectivamente.;

β_0 é o intercepto do modelo de regressão, que corresponde ao valor genotípico para $X=0$, ou seja, a classe Mm;

β_1 é o coeficiente angular do modelo;

e_i é o erro aleatório associado a observação y_{ij} ; $e_{ij} \sim N(0; \frac{1}{2})$;

Ajustando o modelo de regressão com as médias dos indivíduos para cada classe genotípica do marcador, o modelo será:

$$Y = X\beta + \epsilon$$

$$\begin{bmatrix} \bar{y}_{MM} \\ \bar{y}_{Mm} \\ \bar{y}_{mm} \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \epsilon$$

Para o qual, o estimador de mínimos quadrados é: $\hat{\beta} = (X'X)^{-1}X'Y$

Então, nesse caso específico,

$$\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = \begin{bmatrix} \frac{1}{3} & 0 \\ 0 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} (\bar{y}_{MM} + \bar{y}_{Mm} + \bar{y}_{mm}) \\ (\bar{y}_{MM} - \bar{y}_{mm}) \end{bmatrix} \quad \text{e a esperança matemática de } \hat{\beta} \text{ será:}$$

$$E \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = E \begin{bmatrix} \frac{1}{3}(\bar{y}_{MM} + \bar{y}_{Mm} + \bar{y}_{mm}) \\ \frac{1}{2}(\bar{y}_{MM} - \bar{y}_{mm}) \end{bmatrix}$$

$$E \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = \begin{bmatrix} \frac{1}{3}[4dc(1-c) + d(1-2c+2c^2)] \\ \frac{1}{4}a(1-2c) \end{bmatrix} :$$

Verifica-se, dessa forma, que, $\hat{\beta}_0$ e $\hat{\beta}_1$; respectivamente, são estimadores viesados para os efeitos de dominância (d) e aditivo (a). Assim, testar a hipótese:

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_a : \beta_1 \neq 0 \end{cases}$$

equivale a testar a hipótese de que o efeito aditivo a seja nulo, e/ou $c = 0,5$.

Análise de variância

Seja a geração F_2 do cruzamento MMQQ x mmqq, anteriormente descrita. Sabe-se que nessa geração existirão as seguintes classes genotípicas: MM $_ _$, Mm $_ _$ e mm $_ _$. Considerando as classes fenotípicas como um efeito no modelo, com três níveis, e cada indivíduo com mesmo genótipo como uma repetição, tem-se que o valor observado de um determinado indivíduo pode ser modelado por:

$$y_{ij} = \mu + g_i + e_{ij};$$

em que:

y_{ij} é o valor observado no j -ésimo indivíduo, que apresenta o genótipo i ;

μ é uma constante inerente a todas as observações;

g_i é o efeito do i -ésimo genótipo (MM , Mm , ou mm);

e_{ij} é o erro aleatório associado à observação y_{ij} ; que sob H_0 ; $e_{ij} \sim N(0, \sigma^2)$;

O esquema da análise de variância encontra-se apresentado na Tabela 3.

Tabela 3. Esquema da análise de variância para análise de marcas simples, considerando um loco marcador

Fonte de variação	GL	SQ	QM	F
Genótipos	(2)	SQM	QMM	$\frac{QMM}{QMR}$
Aditivo	1	SQA	QMA	$\frac{QMA}{QMR}$
Dominância	1	SQD	QMD	$\frac{QMD}{QMR}$
Resíduo	$\sum_{i=1}^3 (r_i - 1)$	SQR	QMR	
Total	$\sum_{i=1}^3 r_i - 1$	SQT		

r_i é o número de indivíduos com o genótipo i .

O efeito aditivo é estimado pelo contraste :

$$Y_A = \mu_{MM} - \mu_{mm};$$

e o efeito de dominância por

$$Y_D = 2\mu_{Mm} - (\mu_{MM} + \mu_{mm});$$

sendo que, evidentemente, esses contrastes são ortogonais e podem ser testados via uma estatística F, tendo como numerador o quadrado médio do contraste a ser testado e como denominador o quadrado médio do resíduo.

Deste modo, tem-se que:

$$E[Y_A] = E[1_{MM} - 1_{mm}] = \frac{1}{2}a(1 - 2c)$$

$$E[Y_D] = E[21_{Mm} - (1_{MM} + 1_{mm})] = \frac{1}{2}d(1 - 3c + 3c^2)$$

Trabalhando-se com um par de marcas, têm-se 9 genótipos, e podem-se testar interações entre genes (epistasia), como esquematizado na Tabela 4.

Tabela 4. Esquema da análise de variância, para análise de marcas simples considerando dois locos marcadores simultaneamente

Fonte de variação	GL	SQ	QM	F
Genótipos	(8)	SQG	QMG	$\frac{QMMG}{QMR}$
Marca 1	(2)	SQM1	QMM1	$\frac{QMM1}{QMR}$
Aditivo	1	SQA(1)	QMA(1)	$\frac{QMA(1)}{QMR}$
Dominância	1	SQD(1)	QMD(1)	$\frac{QMD(1)}{QMR}$
Marca 2	(2)	SQM2	QMM2	$\frac{QMM2}{QMR}$
Aditivo	1	SQA(2)	QMA(2)	$\frac{QMA(2)}{QMR}$
Dominância	1	SQD(2)	QMD(2)	$\frac{QMD(2)}{QMR}$
Epistasia (Marca1 x Marca 2)	(4)	SQE	QME	$\frac{QMDE}{QMR}$
Aditivo x Aditivo	1	SQAxA	QMAxA	$\frac{QMAxA}{QMR}$
Aditivo x Dominância	1	SQAxD	QMAxD	$\frac{QMAxD}{QMR}$
Dominância x Aditivo	1	SQDxA	QMDxA	$\frac{QMDxA}{QMR}$
Dominância x Dominância	1	SQDxD	QMDxD	$\frac{QMDxD}{QMR}$
Resíduo	$\sum_{i=1}^9 (r_{i-1})$	SQR	QMR	
Total	$\sum_{i=1}^9 r_{i-1}$	SQT		

Bearzoti & Vencovsky (1998), utilizando uma combinação de modelos

de regressão linear simples, propuseram um estimador da proporção da variância genética, explicada pelos marcadores moleculares ($\hat{\rho}$), a ser utilizado na estimação do escore molecular, do Índice de seleção de Lande & Thompson (1990), que é aplicável a experimentos conduzidos em um único ambiente, dado por:

$$\hat{\rho} = \frac{\sigma_m^2}{\sigma_m^2 + \sigma_h^2} \quad (16)$$

Para tanto, consideraram um experimento, conduzido no delineamento inteiramente casualizado, no qual foram avaliadas g progênies em r repetições, tendo adotado o seguinte modelo:

$$y_{ij} = \mu + g_i + e_{ij} \quad (17)$$

em que:

y_{ij} é o valor fenotípico observado para a progênie i , $i = 1, 2, \dots, g$, na repetição j , $j = 1, 2, \dots, r$;

μ é uma constante inerente a todas as observações;

g_i é o efeito da progênie i ;

e_{ij} é o erro aleatório associado à observação y_{ij} .

com, g_i sendo modelado por:

$$g_i = \sum_t \beta_t Z_{it} + h_i \quad (18)$$

em que:

β_t é o coeficiente de regressão da marca t ;

Z_{it} é uma variável dummy codificadora do genótipo da i -ésima progênie, quanto a marca t

h_i é a porção de g_i não explicada pelas marcas (desvios)

Assumindo que:

$$e_{ij} \sim N(0; \sigma_e^2)$$

$$g_i \sim N(0; \sigma_g^2);$$

e_{ij} e g_i são independentes;

$$\text{cov}(g_i; g_{i^0}) = 0.$$

$$\sigma_g^2 = \sigma_m^2 + \sigma_h^2$$

$$V(y_{ij} | m_i) = \sigma_e^2 + \sigma_h^2;$$

$$\text{Cov}(y_{ij}; y_{ij^0} | m_i) = \sigma_h^2.$$

Em que σ_m^2 é o componente de variância devido aos efeitos associados aos marcadores moleculares, e σ_h^2 o componente de variância devido aos desvios desses efeitos, obtiveram o quadro de análise de variância apresentado na tabela 5.

Tabela 5. Esquema da análise de variância individual, com as respectivas expressões das esperanças matemáticas dos quadrados médios, apresentado por Bearzoti & Vencovsky (1998).

FV	GL	SQ	E(QM)
Progenies	g-1	$y^0(P_P - P_1)y$	$\sigma_e^2 + r\sigma_h^2 + r\sigma_m^2$
Reg	k	$y^0(P_R - P_1)y$	$\sigma_e^2 + r\sigma_h^2 + \frac{(g-1)}{k}r\sigma_m^2;$
Desv	g-k-1	$y^0(P_P - P_R)y$	$\sigma_e^2 + r\sigma_h^2$
Resíduo	g(r-1)	$y^0(I - P)y$	σ_e^2
Total	rg-1	$y'y - y^011^0y$	

Método da máxima verossimilhança (LOD score)

O método da máxima verossimilhança para análise de marcas simples foi apresentado por Weller (1986), sendo que uma das suas grandes vantagens, em relação aos métodos anteriormente apresentados, é estimar a posição do QTL, separando a frequência de recombinação c dos efeitos aditivos (a) e de dominância (d).

Assumindo que os fenótipos y dos indivíduos genotipados têm distribuição normal e considerando um marcador M ligado a um QTL Q , a densidade do i -ésimo indivíduo terá a seguinte forma geral:

$$f(y_i=M_j) = \sum_{k=1}^g N(1_Q; \frac{3}{4}^2) P(Q_k=M_j); \quad (19)$$

em que:

M_j é o genótipo j do marcador M (MM, Mm ou mm);

Q_k é o genótipo k do QTL Q ;

$N(\cdot)$ é a densidade normal;

g é número de classes genotípicas;

$P(Q_k=M_j)$ é a probabilidade condicional do indivíduo i possuir o genótipo Q_k do QTL, dado que possui o genótipo M_j do marcador.

Pode-se notar que a densidade 19 consiste em uma mistura de distribuições normais, na qual as proporções de mistura, $P(Q_k=M_j)$, são funções da posição (distância) do QTL em relação ao marcador e do delineamento genético utilizado, enquanto que os efeitos do QTL afetam apenas os parâmetros da distribuição (1_Q e $\frac{3}{4}^2$).

Considerando uma geração F_2 , com um QTL ligado ao marcador, apresentando uma frequência de recombinação c com este, e assumindo-se que os fenótipos seguem uma distribuição normal, as densidades para os indivíduos portadores dos três possíveis genótipos, quanto ao marcador, serão:

$$f(y_i=MM) = (1 - c)^2 N(1_{QQ}; \frac{3}{4}^2) + 2c(1 - c) N(1_{Qq}; \frac{3}{4}^2) + c^2 N(1_{qq}; \frac{3}{4}^2);$$

$$f(y_i=Mm) = c(1 - c)^2 N(1_{QQ}; \frac{3}{4}^2) + (1 - 2c + 2c^2) N(1_{Qq}; \frac{3}{4}^2) + c(1 - c) N(1_{qq}; \frac{3}{4}^2);$$

$$f(y_i=mm) = c^2 N(1_{QQ}; \frac{3}{4}^2) + 2c(1 - c) N(1_{Qq}; \frac{3}{4}^2) + c(1 - c)^2 N(1_{qq}; \frac{3}{4}^2);$$

A função de verossimilhança ($y; \mu$) considerando-se n indivíduos (linhagens) F_2 , é dada por:

$$L(y; \mu) = \prod_{i=1}^n f(y_i; \mu) \quad (20)$$

Com $\mu = c; 1_{QQ}; 1_{Qq}; 1_{qq}; \frac{3}{4}^2$.

O método da máxima verossimilhança consiste, portanto, em determinar os valores de $c; 1_{QQ}; 1_{Qq}; 1_{qq}; \frac{3}{4}^2$ que maximizam $L(y; \mu)$:

O teste estatístico, para a ligação entre o marcador e o QTL, consiste num critério baseado na razão de verossimilhança, da seguinte forma:

$$\chi^2 = -2 \ln \frac{\max L(H_a)}{\max L(H_0)} \quad (21)$$

Em que H_a refere-se à presença de um QTL ligado ao marcador, cuja função de verossimilhança para o caso de uma geração F_2 , anteriormente apresentada, possui cinco parâmetros ($\mu = c; 1_{QQ}; 1_{Qq}; 1_{qq}; \frac{3}{4}^2$) e H_0 refere-se à ausência de QTL segregando, isto é, sob a pressuposição de que a distribuição dos fenótipos é uma única distribuição normal ($1_{QQ} = 1_{Qq} = 1_{qq} = 1$). Nesse caso, a função de verossimilhança $L(y; \mu)$ possui apenas dois parâmetros (1 e $\frac{3}{4}^2$). Assim sendo, a estatística 21 pode ser escrita como:

$$\chi^2 = -2 \ln \frac{\max L(y; c; 1_{QQ}; 1_{Qq}; 1_{qq}; \frac{3}{4}^2)}{\max L(y; 1; \frac{3}{4}^2)} \quad (22)$$

Essa estatística segue, aproximadamente, uma distribuição Qui-quadrado com 3 graus de liberdade (diferença entre o número de parâmetros estimados sob H_a e sob H_0). Deste modo, pode-se utilizar a distribuição $\chi^2_{(3)}$ para testar a hipótese de ausência de QTL segregando, ligado ao marcador em questão.

No mapeamento de QTLs, é muito comum o uso do LOD score (log of the ODDS). A diferença entre o LOD score e a estatística χ^2 é a base do logaritmo

usada no teste: enquanto que, no teste χ^2 , trabalha-se com logaritmos naturais, para o cálculo do LOD score, utiliza-se logaritmo na base 10. A principal razão de trabalhar com o LOD score, em detrimento a χ^2 , está na interpretação dos resultados. A estatística χ^2 é interpretada como a probabilidade de ocorrência dos dados, sob a hipótese nula, a qual é calculada usando a distribuição teórica de Qui-quadrado. Já a interpretação do LOD score é baseada no conceito de odds ratio (razão de chances), não sendo necessária nenhuma distribuição teórica para sua interpretação, uma vez que um LOD score igual a 3 indica que a hipótese alternativa é $10^3 = 1000$ vezes mais provável que a hipótese nula, sendo que o ponto de corte para rejeição da hipótese de ausência de QTL associado ao marcador, a um dado nível de significância, α , é dado por:

$$\text{LOD} \simeq \frac{\hat{A}_{\alpha}^2}{2 \ln 10}; \quad (23)$$

A identificação da posição do QTL, utilizando o método da máxima verossimilhança, é realizada calculando o valor do LOD score para os possíveis valores de c , tomando o valor de c correspondente ao máximo LOD obtido como sendo a posição mais provável do QTL, caso a hipótese H_0 tenha sido rejeitada. Para facilitar a visualização e interpretação dos resultados, é comum a apresentação gráfica da estatística χ^2 ou do LOD-score, em função da posição do QTL (c) no cromossomo, como esquematizado na Figura 4.

Nesta figura, encontram-se apresentados os valores dos LOD score para dois caracteres hipotéticos, (var 1 e var 2) ao longo de um cromossomo. Assume-se a possibilidade de existência de um QTL em uma determinada posição, quando o valor do LOD score supera o ponto de corte ("threshold"), representado pela linha horizontal. Pode-se observar, nesta figura, uma forte evidência de existência de um QTL, associado ao caráter var 1, localizado, aproximadamente no meio do cromossomo (posição 5), não existindo evidência da existência de QTLs associados ao caráter var 2, pois, em nenhuma posição do cromossomo, o LOD score desse caráter ultrapassou o "threshold".

Figura 4 - Apresentação gráfica dos LOD-score para dois caracteres (var 1 e var 2) em função da posição do QTL em um cromossomo. Assume-se a possibilidade de existência de um QTL, em uma determinada posição, quando o valor do LOD score supera o ponto de corte do teste ("threshold"), representado pela linha horizontal.

Para a obtenção dos estimadores de máxima verossimilhança, é computacionalmente ineficiente tentar obter o máximo de $L(y;\mu)$ simultaneamente para c , as médias dos genótipos e $\frac{3}{4}^2$. Uma alternativa útil consiste em obter as estimativas das médias e $\frac{3}{4}^2$, mantendo o valor de c fixo, o que pode ser implementado por meio do algoritmo ECM ("Expectation / Conditional Maximization"), que é uma variação do algoritmo EM ("Expectation Maximization").

Algoritmo EM e ECM

O algoritmo EM (Dempster et al., 1977) é utilizado em casos nos quais as observações podem ser consideradas como dados incompletos, ou seja, quando há mais categorias nos dados do que as que podem ser distinguidas. Isso ocorre no mapeamento de QTLs, pois apenas os genótipos dos marcadores podem ser distinguidos e, no caso de uma geração F_2 ; para cada classe genotípica de marcadores podem existir três classes de QTL. Cada iteração do algoritmo consiste em um pas-

so E ("expectation"), em que o modelo para o qual a verossimilhança foi definida é utilizado para estimar as frequências das classes que não podem ser distinguidas ("hidden categories"), e um passo M ("maximization"), que utiliza essas frequências para obter as estimativas de máxima verossimilhança.

A título de ilustração considere-se o problema de obter a estimativa da frequência alélica do alelo recessivo, utilizando marcadores dominantes em uma geração F_2 , para a qual as classes fenotípicas com suas respectivas frequências esperadas estão apresentadas na Tabela 6.

Tabela 6. Classes fenotípicas e frequências esperadas em uma geração F_2

Classes	Frequência esperada
AA	n_{AA}
Aa	n_{Aa}
aa	n_{aa}
Total	n

O estimador de máxima verossimilhança, da frequência alélica do alelo a , nesta situação, é dado por:

$$\hat{p}_a = \frac{1}{2n} [n_{Aa} + 2n_{aa}] :$$

Ao se utilizarem marcadores dominantes, não é possível distinguir as classes AA e Aa, portanto, não há como conhecer n_{Aa} , sendo necessário o emprego do algoritmo EM para obtenção da estimativa de máxima verossimilhança de p_a . O algoritmo consiste em:

1. Obter uma estimativa inicial para p_a , $p_a^{(0)}$;
2. Usar $p_a^{(0)}$ para estimar n_{Aa}^* , $n_{Aa}^* = 2p_a^{(0)}(1 - p_a^{(0)})n$ (Passo E)

3. Obter a estimativa de máxima verossimilhança de $p_a^{(1)}$; $p_a^{(1)} = \frac{1}{2n} [n_{Aa}^{(1)} + 2n_{aa}]$
(Passo M)
4. Usar o $p_a^{(1)}$ como estimativa inicial para o algoritmo e repetir iterativamente até a convergência

Supondo que os valores observados fossem $n_{Aa} = 75$ e $n_{aa} = 25$, o algoritmo seria utilizado do seguinte modo:

Primeira iteração

1. $p_a^{(0)} = \frac{2n_{aa}}{2n} = \frac{50}{200} = 0;25$
2. $n_{Aa}^{(0)} = 2p_a^{(0)}(1 - p_a^{(0)})n = 2(0;25)(1 - (0;25))100 = 37;5$
3. $p_a^{(1)} = \frac{1}{2n} [n_{Aa}^{(0)} + 2n_{aa}] = \frac{1}{2(100)} [37;5 + 2(25)] = 0;4375$

Segunda iteração

1. $p_a^{(1)} = 0;4375$
2. $n_{Aa}^{(1)} = 2p_a^{(1)}(1 - p_a^{(1)})n = 2(0;4375)(1 - (0;4375))100 = 49;219$
3. $p_a^{(2)} = \frac{1}{2n} [n_{Aa}^{(1)} + 2n_{aa}] = \frac{1}{2(100)} [49;219 + 2(25)] = 0;4961$

Terceira iteração

1. $p_a^{(2)} = 0;4961$
2. $n_{Aa}^{(2)} = 2p_a^{(2)}(1 - p_a^{(2)})n = 2(0;4961)(1 - (0;4961))100 = 49;997$
3. $p_a^{(3)} = \frac{1}{2n} [n_{Aa}^{(2)} + 2n_{aa}] = \frac{1}{2(100)} [49;997 + 2(25)] = 0;49999$

Quarta iteração

1. $p_a^{(3)} = 0;49999$
2. $n_{Aa}^{(3)} = 2p_a^{(3)}(1 - p_a^{(3)})n = 2(0;49999)(1 - (0;49999))100 = 50;0$

$$3. p_a^{(4)} = \frac{1}{2n} [n_{Aa}^{(4)} + 2n_{aa}] = \frac{1}{2(100)} [50 + 2(25)] = 0,50$$

Quinta iteração

$$1. p_a^{(4)} = 0,50$$

$$2. n_{Aa}^{(4)} = 2p_a^{(4)}(1 - p_a^{(4)})n = 2(0,50)(1 - (0,50))100 = 50; 0$$

$$3. p_a^{(5)} = \frac{1}{2n} [n_{Aa}^{(4)} + 2n_{aa}] = \frac{1}{2(100)} [50 + 2(25)] = 0,50$$

Como não houve nenhuma alteração na estimativa de p_a , da quarta para a quinta iteração, diz-se que o algoritmo convergiu com quatro iterações e a estimativa de máxima verossimilhança de p_a é 0,5.

O algoritmo ECM (Meng & Rubin, 1993) é uma alteração do algoritmo EM, que consiste em fixar um valor de c , e, então, obter as estimativas de máxima verossimilhança para os demais parâmetros do modelo, via algoritmo EM. O nome ECM deriva do fato de a maximização ser realizada condicionada a um fator fixo de c .

2.3.2 Mapeamento por Intervalo

O mapeamento por intervalo, apresentado por Lander & Botstein (1989), é uma extensão do método de máxima verossimilhança para marcas simples, anteriormente apresentado, sendo baseado nas frequências conjuntas de um par de marcas adjacentes e um QTL pontual, o que é baseado por essas duas marcas, como esquematizado na figura 5.

Deste modo, existirá, em uma geração F_2 , um total de 27 classes fenotípicas, como pode ser observado na Tabela 7, em vez de de 9, como ocorre na análise de marcas simples.

Figura 5 - Esquema da localização de um QTL entre dois marcadores. As marcas M_1 e M_2 estão ligadas com uma frequência de recombinação c , e o QTL está localizado entre estas com uma frequência c_1 de recombinação com a marca M_1 e c_2 com M_2

Método da máxima verossimilhança

Assumindo que as 27 classes genotípicas apresentadas na Tabela 7 possuem a mesma variância $\frac{3}{4}^2$ e que os valores fenotípicos do caráter em questão seguem a distribuição normal, a função de verossimilhança para o mapeamento por intervalo será

$$L(y) = \prod_{i=1}^n \left[\sum_{k=1}^g N\left(1_{Q_i}; \frac{3}{4}^2\right) P(Q_k = M_j) \right] \quad (24)$$

em que:

M_j é o genótipo j dos marcadores ($M_1M_1M_2M_2$, $M_1M_1M_2m_2$, ..., $m_1m_1m_2m_2$);

Q_k é o genótipo k do QTL Q (QQ , Qq ou qq), no caso de uma geração F_2 ;

$N(\cdot)$ é a densidade normal;

g é número de classes genotípicas, que, para uma geração F_2 , é igual a 27;

$P(Q_k = M_j)$ é a probabilidade condicionada do genótipo k do QTL ao genótipo j dos marcadores, que pode ser obtida com base na Tabela 7

Tabela 7. Frequências esperadas dos genótipos na geração F_2 ; considerando 3 locos ligados, com uma frequência de recombinação iguais a c_1 entre M_1 e Q , c_2 entre Q e M_2 e c entre M_1 e M_2 :

Genótipo dos marcadores	Genótipo do QTL		
	QQ	Qq	qq
$M_1M_1M_2M_2$	$\frac{1}{4}(1-c_1)^2(1-c_2)^2$	$\frac{1}{2}(1-c_1)c_2c_1(1-c_2)$	$\frac{1}{4}c_1^2c_2^2$
$M_1M_1M_2m_2$	$\frac{1}{4}(1-c_1)^2(1-c_2)c_2$	$\frac{1}{4}(1-c_1)c_2^2c_1$	$\frac{1}{4}c_1^2(1-c_2)c_2$
$M_1M_1m_2m_2$	$\frac{1}{4}(1-c_1)^2c_2^2$	$\frac{1}{2}(1-c_1)c_2c_1(1-c_2)$	$\frac{1}{4}c_1^2(1-c_2)^2$
$M_1m_1M_2M_2$	$\frac{1}{4}(1-c_1)(1-c_2)^2c_1$	$\frac{1}{4}c_1^2(1-c_2)c_2$	$\frac{1}{4}(1-c_1)c_2^2c_1$
$M_1m_1M_2m_2$	$\frac{1}{4}(1-c_1)c_2c_1(1-c_2)$	$\frac{1}{4}c_1^2(1-c_2)^2$	$\frac{1}{4}(1-c_1)c_2c_1(1-c_2)$
$M_1m_1m_2m_2$	$\frac{1}{4}(1-c_1)c_2^2c_1$	$\frac{1}{4}c_1^2(1-c_2)c_2$	$\frac{1}{4}(1-c_1)(1-c_2)^2c_1$
$m_1m_1M_2M_2$	$\frac{1}{4}c_1^2(1-c_2)^2$	$\frac{1}{2}(1-c_1)c_2c_1(1-c_2)$	$\frac{1}{4}(1-c_1)^2c_2^2$
$m_1m_1M_2m_2$	$\frac{1}{4}c_1^2(1-c_2)c_2$	$\frac{1}{4}(1-c_1)c_2^2c_1$	$\frac{1}{4}(1-c_1)^2(1-c_2)c_2$
$m_1m_1m_2m_2$	$\frac{1}{4}c_1^2c_2^2$	$\frac{1}{2}(1-c_1)c_2c_1(1-c_2)$	$\frac{1}{4}(1-c_1)^2(1-c_2)^2$

Pode-se observar que o mapeamento por intervalo é computacionalmente mais trabalhoso do que a análise de marcas simples, pois tomando como exemplo, um indivíduo com genótipo $M_1M_1M_2M_2$, tem-se que sua densidade será:

$$f(y_i=M_1M_1M_2M_2) = \frac{(1-c_1)^2(1-c_2)^2}{(1-c)^2}N(1_{QQ}; \frac{3}{4}^2) + \frac{2(1-c_1)c_2c_1(1-c_2)}{(1-c)^2}N(1_{Qq}; \frac{3}{4}^2) + \frac{c_1^2c_2^2}{(1-c)^2}N(1_{qq}; \frac{3}{4}^2):$$

Aproximação de Haley & Knott (1992)

Haley & Knott (1992) propuseram a utilização de uma regressão linear múltipla para mapear QTLs anqueados por duas marcas. O modelo proposto assume um QTL, Q , localizado entre dois marcadores codominantes (M_1 e M_2), que a variância dentro das três classes genotípicas do QTL é a mesma e segue a distribuição normal. Aqui será apresentado o método utilizando uma população F_2 , muito embora o mesmo seja facilmente estendido para outros tipos de populações segregantes.

Considere-se a geração F_2 apresentada na Tabela 7. Os efeitos genotípicos dos três possíveis genótipos do QTL na geração F_2 serão $m+a$, $m+d$ e $m-a$, para os genótipos QQ , Qq e qq , respectivamente, em que m é a média dos pais (média dos homozigotos), a é o efeito aditivo e d o efeito de dominância.

Os valores genotípicos esperados, em termos do QTL, para cada classe genotípica dos marcadores na geração F_2 , (Tabela 8), são obtidos com base nas frequências esperadas vistas na Tabela 7, do seguinte modo: considere-se o genótipo homozigoto para os marcadores $M_1M_1M_2M_2$; cuja frequência esperada é igual a $(1-c)^2=4$ e as frequências esperadas para os três possíveis genótipos do QTL, são $\frac{1}{4}(1-c_1)^2(1-c_2)^2$, $\frac{1}{2}(1-c_1)c_2c_1(1-c_2)$ e $\frac{1}{4}c_1^2c_2^2$, para os genótipos QQ , Qq e qq , respectivamente, então, o valor fenotípico esperado para essa classe de indivíduos será:

$$m + a[(1-c_1)^2(1-c_2)^2 - c_1^2c_2^2] = (1-c)^2 + d[2(1-c_1)(1-c_2)c_1c_2] = (1-c)^2$$

Tabela 8. Coeficientes de a e d , para uma geração F_2 ; a serem utilizados no modelo Haley & Knott (1992)

Genótipo	Esperança em termos de	
	a	d
$M_1M_1M_2M_2$	$[(1-c_1)^2(1-c_2)^2 - c_1^2c_2^2] = (1-c)^2$	$[2c_1(1-c_1)c_2(1-c_2)] = (1-c)^2$
$M_1M_1M_2m_2$	$[(1-c_1)^2c_2(1-c_2) - c_1^2c_2(1-c_2)] = c(1-c)$	$[c_1(1-c_1)(1-c_2)^2 + c_1(1-c_1)c_2^2] = c(1-c)$
$M_1M_1m_2m_2$	$[(1-c_1)^2c_2^2 - c_1^2(1-c_2)^2] = c^2$	$[2c_1(1-c_1)c_2(1-c_2)] = c^2$
$M_1m_1M_2M_2$	$[c_1(1-c_1)(1-c_2) - c_1(1-c_1)c_2^2] = c(1-c)$	$[(1-c_1)^2c_2(1-c_2) + c_1^2c_2(1-c_2)] = c(1-c)$
$M_1m_1M_2m_2$	0	$[c_1^2c_2^2 + c_1^2(1-c_2)^2 + (1-c_1)^2c_2^2 + (1-c_1)^2(1-c_2)^2] = [c^2 + (1-c)^2]$
$M_1m_1m_2m_2$	$[c_1(1-c_1)c_2^2 - c_1(1-c_1)(1-c_2)^2] = c(1-c)$	$[(1-c_1)^2c_2(1-c_2) + c_1^2c_2(1-c_2)] = c(1-c)$
$m_1m_1M_2M_2$	$[c_1^2(1-c_2)^2 - (1-c_1)^2c_2^2] = c^2$	$[2c_1(1-c_1)c_2(1-c_2)] = c^2$
$m_1m_1M_2m_2$	$[c_1^2c_2(1-c_2) - (1-c_1)^2c_2(1-c_2)] = c(1-c)$	$[c_1(1-c_1)(1-c_2)^2 + c_1(1-c_1)c_2^2] = c(1-c)$
$m_1m_1m_2m_2$	$[c_1^2c_2^2 - (1-c_1)^2(1-c_2)^2] = (1-c)^2$	$[2c_1(1-c_1)c_2(1-c_2)] = (1-c)^2$

Os valores apresentados na Tabela 8 são utilizados para ajustar a e b ,

considerando o seguinte modelo de uma regressao linear múltipla:

$$y_i | (M_j M_{j+1}) = m + aX_{ai} + dX_{di} + \epsilon_i; \quad (25)$$

em que:

$M_j M_{j+1}$ é o genótipo do i -ésimo indivíduo para o intervalo considerado;

a é o efeito aditivo do QTL;

d é o efeito de dominância;

X_{ai} e X_{di} são os coeficientes dos valores de a e d apresentados na Tabela

8 e

ϵ_i é o erro aleatório, $\epsilon_i \sim N(0; \frac{1}{4})$.

Para um intervalo entre duas marcas, valores numéricos são atribuídos à posição do QTL entre as marcas (c, c_1), e os valores dos coeficientes de a e d para cada genótipo marcador são, então, calculados. Uma regressão múltipla é utilizada, para ajustar m , a e d , para cada posição do QTL, utilizando os valores numéricos como coeficientes de a e d . Este procedimento fornece as estimativas para o efeito aditivo a e de dominância d , bem como somas de quadrados e quadrados médios de regressão e desvios, permitindo, assim, a realização de testes de hipóteses sobre a e d . A posição que gerar o melhor ajuste do modelo (menor soma de quadrados de resíduos) fornece a posição mais provável do QTL e as melhores estimativas dos seus efeitos.

A título de exemplo, considerem-se duas marcas, com uma frequência de recombinação entre si igual a 0,10 ($c=0,10$). Tem-se que, para $c_1 = 0,01$, c_2 será igual a 0,09692. Substituindo esses valores na Tabela 8, obtêm-se os coeficientes de a e d , apresentados na Tabela 9. Tais valores serão utilizados para ajustar o modelo, considerando que o QTL apresenta uma frequência de recombinação de 0,01 com a marca M_1 ; isto é, está localizado a 1 cM da marca M_1 . Procede-se dessa maneira sucessivamente para cada valor de c_1 desejado, ajustando um modelo a cada um dos passos (valor de c_1).

Segundo Haley & Knott (1992), assumindo que os fenótipos seguem

Tabela 9. Coeficientes de a e d a serem utilizados no modelo de Haley & Knott (1992), para uma geração F_2 , considerando $c = 0,10$ e $c_1 = 0,01$

Genótipo	Esperança em termos de	
	a	d
$M_1M_1M_2M_2$	0; 9868	$2; 1395 \times 10^{-3}$
$M_1M_1M_2m_2$	0; 7719	$7; 3503 \times 10^{-2}$
$M_1M_1m_2m_2$	0; 9125	0; 1733
$M_1m_1M_2M_2$	$7; 9627 \times 10^{-2}$	0; 7214
$M_1m_1M_2m_2$	0,0000	0; 9860
$M_1m_1m_2m_2$	$-7; 1829 \times 10^{-2}$	0; 7214
$m_1m_1M_2M_2$	-0; 9125	0; 1733
$m_1m_1M_2m_2$	-0; 7719	$7; 3503 \times 10^{-2}$
$m_1m_1m_2m_2$	-0; 9868	$2; 1395 \times 10^{-3}$

a distribuição normal, e se o QTL estiver completamente ligado a um dos marcadores ($c_1 = 0$ ou $c_1 = c$) os resíduos da regressão, obtidos pela equação 25, têm distribuição normal. Nesse caso, as estimativas obtidas pelo método da regressão (Mínimos Quadrados) são também estimativas de máxima verossimilhança, e o teste da razão de verossimilhança pode ser obtido por:

$$\chi^2 = n \ln \left(\frac{SQTot}{SQRes} \right) = -n \ln(1 - r^2); \quad (26)$$

em que:

$SQTot$ é a soma de quadrados total;

$SQRes$ é a soma de quadrados de resíduos da regressão e

r^2 é o coeficiente de determinação do modelo.

Caso o QTL não esteja completamente ligado a um dos marcadores, a distribuição dos resíduos será uma mistura de normais, mesmo nessa situação, Haley

& Knott (1992) demonstraram que a aproximação fornecida pela equação 26 será muito boa.

2.3.3 Mapeamento por Intervalo Composto

As técnicas de mapeamento por intervalo são, indubitavelmente, mais precisas na localização da posição do QTL do que as baseadas em marcas simples, porém ainda apresentam o problema de identificação de "QTLs fantasmas", isto é, de localizar um QTL insistente, como ilustrado na Figura 6, que representa os resultados do teste da razão de verossimilhança em um mapeamento por intervalo, no qual se pode observar, claramente, o surgimento de um "QTL fantasma", localizado entre as marcas M2 e M3, em decorrência da presença de dois QTLs ligados (Q1 e Q2), localizados nos intervalos adjacentes ao intervalo mapeado (M2 a M3).

Figura 6 - Apresentação gráfica dos resultados do teste da razão de verossimilhança, em um mapeamento por intervalo, com o surgimento de um QTL fantasma entre as marcas M2 e M3, devido à influência dos QTLs Q1 e Q2.

Jansen (1992,1993) e Zeng (1993, 1994), independentemente, propuseram a utilização de um modelo de regressão múltipla para eliminação desses efeitos (

QTLs fantasmas).

Um modelo de regressão múltipla é dado por:

$$y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki} + \epsilon_i = \beta_0 + \sum_{j=1}^k \beta_j X_{ji} + \epsilon_i$$

em que:

y_i é o valor observado para o i -ésimo indivíduo;

X_{ji} são variáveis "dummy" representativas dos genótipos do indivíduo i , para o marcador j ;

β_j são coeficientes de regressão parciais, que representam a modificação esperada na variável y devido a variação de uma unidade em X_j estando as demais variáveis X fixas, isto é, medem como y varia em função de X_j após eliminar a influência linear das demais variáveis, como demonstrado por Ho®man & Vieira (1998) para o caso de uma regressão múltipla com duas variáveis exploratórias.

O método de mapeamento por intervalo composto (CIM \Composite Interval Mapping") combina o mapeamento por intervalo com regressão linear múltipla, no intuito de eliminar a interferência de QTLs adjacentes ao intervalo que está sendo mapeado. Considerando o segmento cromossômico esquematizado na Figura 7, o enfoque CIM impede que o QTL Q_2 interfira nos scores dentro do intervalo $(j, j+1)$.

Figura 7 - Esquema de um cromossomo mostrando dois QTLs, Q_1 e Q_2 , localizado nos intervalos °anqueados pelos marcadores $M_j;M_{j+1}$ e $M_{j+2};M_{j+3}$

Porém, esse método não impede que um QTL localizado entre $j+1$ e $j+2$ interfira na estimativa dentro do intervalo $(j, j+1)$. Mas com a utilização de

mapas densos (pequenas distâncias entre marcas adjacentes), a probabilidade de ocorrência de QTLs adjacentes torna-se muito baixa.

Método de máxima verossimilhança

Zeng (1994) propõe o mapeamento por intervalo baseado na razão de verossimilhança do seguinte modo:

$$y_i|(M_j M_{j+1}) = \mu_0 + \mu_j X_{ji} + \sum_{l=j+1}^k \mu_l X_{li} + \epsilon_i;$$

em que:

y_i é o valor observado para o i -ésimo indivíduo;

$(M_j M_{j+1})$ é o genótipo do i -ésimo indivíduo para as marcas j e $j+1$;

μ_j é o efeito do QTL localizado entre as marcas j e $j+1$;

X_{ji} é uma variável "dummy" que pode assumir os valores 1, 0 ou -1, para uma geração F_2 , com probabilidades apresentadas na Tabela 7;

μ_l são os coeficientes de regressão parciais para o caráter em questão, na marca l ;

X_{li} são variáveis "dummy" representativas do genótipo do i -ésimo indivíduo, para a marca l , e

ϵ_i é o erro aleatório associado a i -ésima observação, $\epsilon_i \sim N(0, \sigma^2)$.

A função de verossimilhança é dada por:

$$f[y_i|(M_j M_{j+1})] = P(X_{ji} = 1)f(\mu_{QQ}) + P(X_{ji} = 0)f(\mu_{Qq}) + P(X_{ji} = -1)f(\mu_{qq});$$

em que:

$f(\mu_{QQ})$, $f(\mu_{Qq})$ e $f(\mu_{qq})$ são densidades normais, com médias μ_{QQ} , μ_{Qq} e μ_{qq} , respectivamente, sendo:

$$\mu_{QQ} = \mu_0 + \mu_j + \sum_{l=j+1}^k \mu_l X_{li}$$

$$\mu_{Qq} = \mu_0 + \sum_{l=j+1}^k \mu_l X_{li}$$

$$1_{qq} = \bar{1}_0 - \bar{1}_j + \sum_{l \in j:j+1}^k \bar{1}_l X_{li}$$

Para obtenção dos estimadores de máxima verossimilhança, Zeng (1994) sugere a utilização do algoritmo ECM ("Expectation Conditional Maximization"), ou seja, achar a posição de um eventual QTL dentro do intervalo e, em seguida, estimar os demais parâmetros.

Enfoque de Haley & Knott (1992)

O mapeamento por intervalo composto pode ser realizado com base no enfoque de Haley & Knott (1992), utilizando o seguinte modelo:

$$y_i | (M_j M_{j+1}) = [m + aX_{ai} + dX_{di}] + \sum_{l \in j:j+1}^k \bar{1}_l X_{li} + \epsilon_i;$$

em que a parte entre colchetes é o modelo de Haley e Knott para o mapeamento por intervalo e $\sum_{l \in j:j+1}^k \bar{1}_l X_{li}$ são as marcas fora do intervalo.

A aproximação de Haley e Knott é, computacionalmente, muito eficiente e viabiliza a utilização de técnicas de bootstrap e "data shuffling" (Churchill & Doerge, 1994), o que é computacionalmente inviável utilizando os métodos de máxima verossimilhança.

2.3.4 Mapeamento Por Intervalo Múltiplo

A técnica de mapeamento por intervalo múltiplo ("Multiple Interval Mapping" - MIM) foi apresentada por Kao et al. (1999), sendo baseada no modelo de Cocherham (1954) para interpretação de parâmetros genéticos e no método de máxima verossimilhança para estimação de parâmetros genéticos. Com o MIM, pode-se obter um aumento na precisão e no poder da detecção de QTLs, além de permitir a estimação e análise dos efeitos de interação entre QTLs (epistasia).

Para entender o modelo genético adotado no MIM, considere-se t QTLs, $Q_1; Q_2; \dots; Q_t$, em um retrocruzamento, no qual são possíveis dois genótipos

para cada QTL, $Q_j Q_j$ e $Q_j q_j$; cada qual com probabilidade igual a meio. Considerando t QTLs conjuntamente, existem 2^t diferentes genótipos possíveis na população. O modelo genético de Cockerham, usado para definir os parâmetros genéticos e modelar a relação entre os valores genotípicos e os parâmetros genéticos, é dado por:

$$G_i = \mu + \sum_{j=1}^t a_j x_{ij} + \sum_{j < k}^t w_{jk} (x_{ij} x_{ik}); \quad i = 1; 2; \dots; t \quad (27)$$

em que:

x_{ij} assume valores $\frac{1}{2}$ ou $-\frac{1}{2}$ se o genótipo de Q_j for $Q_j Q_j$ ou $Q_j q_j$ respectivamente;

a_j é o efeito principal do QTL Q_j ;

w_{jk} é o efeito da interação entre os QTLs Q_j e Q_k .

Matricialmente, esse modelo é

$$G = \mu + DE$$

$$\begin{bmatrix} G_{11::11} \\ G_{11::10} \\ G_{11::01} \\ G_{11::00} \\ \vdots \\ G_{00::11} \\ G_{00::10} \\ G_{00::01} \\ G_{00::00} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ \vdots \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \mu + \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & \cdots & \frac{1}{2} & \frac{1}{2} & \cdots & \cdots & \frac{1}{4} \\ \frac{1}{2} & \frac{1}{2} & \cdots & \frac{1}{2} & -\frac{1}{2} & \cdots & \cdots & -\frac{1}{4} \\ \frac{1}{2} & \frac{1}{2} & \cdots & -\frac{1}{2} & \frac{1}{2} & \cdots & \cdots & -\frac{1}{4} \\ \frac{1}{2} & \frac{1}{2} & \cdots & -\frac{1}{2} & -\frac{1}{2} & \cdots & \cdots & \frac{1}{4} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ -\frac{1}{2} & -\frac{1}{2} & \cdots & \frac{1}{2} & \frac{1}{2} & \cdots & \cdots & \frac{1}{4} \\ -\frac{1}{2} & -\frac{1}{2} & \cdots & \frac{1}{2} & -\frac{1}{2} & \cdots & \cdots & -\frac{1}{4} \\ -\frac{1}{2} & -\frac{1}{2} & \cdots & -\frac{1}{2} & \frac{1}{2} & \cdots & \cdots & -\frac{1}{4} \\ -\frac{1}{2} & -\frac{1}{2} & \cdots & -\frac{1}{2} & -\frac{1}{2} & \cdots & \cdots & \frac{1}{4} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \\ w_{12} \\ w_{13} \\ \vdots \\ w_{(m_i-1)m} \end{bmatrix}$$

em que,

G contém os valores genotípicos dos 2^t genótipos possíveis, os subscritos 1 e 0 em G denotam os genótipos homozigotos e heterozigotos para o QTL em ordem do primeiro, segundo,..., t -ésimo QTL, respectivamente;

D é a matriz do delineamento genético, na qual as m primeiras colunas apresentam os coeficientes dos efeitos principais dos QTLs, e as últimas $m(m-1)/2$

os coeficientes dos efeitos epistáticos;

E é o vetor que contém os efeitos principais e epistáticos dos QTLs.

Para o mapeamento por intervalo múltiplo, assume-se que os t QTLs, $Q_1; Q_2; \dots; Q_t$ estão localizados nas posições $p_1; p_2; \dots; p_t$, em t diferentes intervalos entre marcas, $I_1; I_2; \dots; I_t$, ao longo do genoma, controlando um caráter quantitativo, y , sendo o modelo que relaciona o caráter y aos t QTLs dado por:

$$y_i = \mu + \sum_{j=1}^t a_j x_{ij} + \sum_{j \neq k}^t \pm_{jk} (w_{jk} x_{ij} x_{ik}) + \epsilon_i \quad (28)$$

em que:

y_i é o valor observado do caráter quantitativo em questão;

μ é a média do caráter;

x_{ij} é a variável codificada representativa do genótipo do j -ésimo QTL;

a_j e w_{jk} tem as mesmas definições do modelo genético 27;

$\pm_{jk}$ é uma variável indicadora para o efeito de epistasia entre os QTLs j e k , assumindo o valor 1, se existir a epistasia e zero, caso contrário;

ϵ_i é o erro aleatório, $\epsilon_i \sim N(0; \sigma^2)$;

Esse modelo é denominado MIM, porque múltiplos (t) intervalos são simultaneamente utilizados para mapear múltiplos QTLs simultaneamente.

No modelo MIM, o genótipo de cada QTL, Q_j , no intervalo I_j , é desconhecido, mas sua distribuição pode ser inferida a partir dos marcadores "âncoras", baseado na frequência de recombinação entre os mesmos. Para inferir a distribuição de probabilidade condicional conjunta do genótipo dos t QTLs, utiliza-se da propriedade de que, se não houver interferência, isto é, a ocorrência de um "crossing over" não alterar a probabilidade de ocorrência de outros "crossing overs", as distribuições condicionais dos genótipos dos QTLs individuais, dado o genótipo dos marcadores "âncoras", são independentes. Desse modo, tem-se que:

$$P(Q_1; Q_2; \dots; Q_t | I_1; I_2; \dots; I_t) = \prod_{i=1}^t P(Q_i | I_i);$$

ou seja, a probabilidade condicional conjunta dos t QTLs é o produto das probabilidades condicionais marginais de cada QTL.

Portanto, dada uma amostra de tamanho n , a função de verossimilhança para o modelo MIM é:

$$L(\mu | Y; X) = \prod_{i=1}^n \left[\sum_{j=1}^{2^t} P_{ij} \hat{A} \left(\frac{y_i - 1_{ij}}{\sqrt{\sigma^2}} \right) \right];$$

em que:

$\mu = (p_1; p_2; \dots; p_t; a_1; a_2; \dots; a_t; w_{12}; w_{13}; \dots; w_{jk}; \sigma^2)$, p_i é a posição do QTL;

$\hat{A}(\cdot)$ é a densidade normal;

1_{ij} corresponde ao genótipo dos 2^t diferentes QTLs eq(27); ou seja, $1_{ij} = m + D_j E$;

P_{ij} é a probabilidade condicional conjunta correspondente ao genótipo 1_{ij} ; cujos valores para uma geração F_2 encontram-se apresentados Tabela 7, e podem ser obtidos para outras populações segregantes, tais como retrocruzamentos, S_1 ; etc:

Considerando dois QTLs em um retrocruzamento, a verossimilhança para o MIM será uma mistura de quatro normais, e seis parâmetros deverão ser estimados, $(p_1; p_2; a_1; a_2; w_{12}; \sigma^2)$; com três QTLs, a verossimilhança será uma mistura de nove normais, e dez parâmetros serão estimados $(p_1; p_2; p_3; a_1; a_2; a_3; w_{12}; w_{13}; w_{23}; \sigma^2)$: Verifica-se, assim, que o número de distribuições envolvidas na verossimilhança e de parâmetros a serem estimados aumenta drasticamente com o aumento do número de QTLs considerados, sendo o número de distribuições dados por 2^t , e o número de parâmetros por $2^t + 2$. Deste modo, resulta que dois problemas devem ser solucionados para a implementação do MIM. O primeiro é relativo à estimação dos parâmetros da mistura finita de normais, porque à medida que t aumenta, a derivação dos estimadores de máxima verossimilhança para os efeitos dos QTLs e posições torna-se rapidamente impraticável. Para con-

tornar esse problema, Kao et al. (1999) sugerem utilizar as expressões derivadas por Kao & Zeng (1997) para obtenção dos EMV.

O segundo é como encontrar QTLs para ajustar o modelo. Kao et al. (1999) propõem o uso do procedimento "stepwise" para a seleção das marcas que entrarão no modelo, embora existam outras opções a este método, como o "backward", a análise de marcas simples, utilizando o procedimento de Bonferroni, ou FDR, para definir o nível de significância, dentre outras, ou pode-se, ainda, realizar o mapeamento por intervalo ou intervalo composto. O problema desses procedimentos reside no fato de serem procedimentos estatísticos de análise e, portanto, estarem sujeitos aos erros Tipo I e Tipo II. Cometer um erro Tipo I implica assumir a existência de um QTL insistente, o que aumentará, desnecessariamente, o número de parâmetros a serem estimados no procedimento MIM. Já ao cometer um erro Tipo II, estar-se-á eliminando um QTL do procedimento MIM. Deste modo, são necessários estudos mais aprofundados sobre metodologias de determinação de marcas e, consequentemente, QTLs, que serão incluídos no procedimento MIM, de modo a balancear o poder de detecção de QTLs e o número de falsos QTLs detectados, no intuito de não causar grandes perdas no procedimento MIM, nem tampouco torná-lo computacionalmente inviável.

2.3.5 Mapeamento de Múltiplos Caracteres

Jiang & Zeng (1995) apresentaram uma metodologia para o mapeamento de QTLs, considerando mais de um caráter simultaneamente, baseada no método de mapeamento por intervalo composto. Os autores argumentam que, considerando a estrutura de correlação entre múltiplos caracteres, um aumento no poder de detecção de QTLs e na precisão do mapeamento é esperado, quanto comparado às análises realizadas com caracteres isolados.

Bearzoti (2000) ressalta, ainda, que outras vantagens potências dessa técnica são: a possibilidade de estudar caracteres biológicos gerais, que são baseados

em vários caracteres, como o peso total, que é função do peso de vários órgãos; avaliar o mesmo caráter em diferentes fases da vida do indivíduo, quantificando os efeitos dos QTLs em cada fase (análise de medidas repetidas); estudar a natureza da correlação genética entre caracteres, se devido a pleiotropia (um QTL afetando mais de um caráter) ou a ligação (os caracteres são controlados por diferentes QTLs fortemente ligados).

O modelo utilizado para o mapeamento de múltiplos caracteres apresentado por Jiang & Zeng (1995) é essencialmente uma extensão do modelo utilizado para o mapeamento por intervalo composto.

Desta forma, considere-se que n indivíduos de uma geração F_2 sejam avaliados em relação a m caracteres e genotipados para um determinado número de marcadores moleculares. Assim sendo, uma extensão do modelo de mapeamento por intervalo composto, para o mapeamento de múltiplos caracteres é dado por:

$$\begin{aligned} y_{i1} | (M_j M_{j+1}) &= \mu_{01} + \mu_{j1} X_{ji} + \sum_{l \in j:j+1}^k \mu_{l1} X_{li} + \epsilon_{i1} \\ y_{i2} | (M_j M_{j+1}) &= \mu_{02} + \mu_{j2} X_{ji} + \sum_{l \in j:j+1}^k \mu_{l2} X_{li} + \epsilon_{i2} \\ &\vdots \\ y_{im} | (M_j M_{j+1}) &= \mu_{0m} + \mu_{jm} X_{ji} + \sum_{l \in j:j+1}^k \mu_{lm} X_{li} + \epsilon_{im} \end{aligned}$$

em que:

y_{im} é o valor observado para o i -ésimo indivíduo quanto ao m -ésimo caráter;

$(M_j M_{j+1})$ é o genótipo do i -ésimo indivíduo para as marcas j e $j+1$;

μ_{jm} é o efeito do QTL localizado entre as marcas j e $j+1$ sobre o caráter m ;

X_{ji} é uma variável "dummy", que pode assumir os valores 1, 0 ou -1, para uma geração F_2 , com probabilidades apresentadas na Tabela 7;

β_{im} são os coeficientes de regressão parciais para o caráter m , na marca i ;

X_{ij} são variáveis "dummy" representativas do genótipo o i -ésimo indivíduo, para a marca i , e

ϵ_{im} é o erro aleatório associado a i -ésima observação do m -ésimo caráter.

Assume-se que os resíduos têm distribuição normal multivariada, ou seja:

$$\begin{bmatrix} \epsilon_{i1} \\ \epsilon_{i2} \\ \vdots \\ \epsilon_{im} \end{bmatrix} \sim N \left(\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{11}^2 & \sigma_{12} & \cdots & \sigma_{1m} \\ \sigma_{12} & \sigma_{22}^2 & \cdots & \sigma_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1m} & \sigma_{2m} & \cdots & \sigma_{mm}^2 \end{bmatrix} \right)$$

A função de verossimilhança é dada por:

$$L = \prod_{i=1}^n [P(QQ|M_j M_{j+1})f_{QQ}(y_i) + P(Qq|M_j M_{j+1})f_{Qq}(y_i) + P(qq|M_j M_{j+1})f_{qq}(y_i)];$$

em que:

$f_{QQ}(y_i)$, $f_{Qq}(y_i)$ e $f_{qq}(y_i)$ são densidades normais multivariadas para os genótipos QQ, Qq e qq do QTL.

2.4 Testes Múltiplos e seleção de marcas associadas a QTL's

O mapeamento de QTLs difere dos demais tipos de experimentos conduzidos em genética, por se tratar basicamente de um procedimento de testes múltiplos, uma vez que consiste em distribuir marcadores moleculares por todo o genoma e, em seguida, realizar uma varredura ou "screening" dessas marcas, no intuito de verificar qual(is) está(ão) ligada(s) a QTLs. O problema encontrado nesse procedimento refere-se ao nível de significância conjunto do teste, também chamado nível de significância genômico, visto que o nível de significância individual (de cada teste) é definido pelo pesquisador.

Nível de significância é definido, em estatística, como a probabilidade de se cometer um erro Tipo I (rejeitar uma hipótese nula verdadeira). No presente caso, a hipótese de nulidade testada é a de não existência de QTL associado a marca. No mapeamento de QTLs, cometer um erro Tipo I implica admitir a existência de um QTL inexistente (QTL fantasma).

Suponha-se que em um determinado genoma sejam distribuídas aleatoriamente t marcas. Assim, ao realizar a varredura dessas marcas a procura de QTLs, serão realizados t testes de marcas simples (um para cada marca), adotando-se um nível de significância (α) para cada teste. Tem-se que o nível de significância conjunto do teste, ou genômico (α^g), considerando os t testes independentes será:

α^g = probabilidade de rejeitar pelo menos uma hipótese nula verdadeira, ou seja,

α^g = 1- probabilidade de não rejeitar nenhuma hipótese nula verdadeira, isto é,

$$\alpha^g = 1 - (1 - \alpha)^t \quad (29)$$

Deste modo, verifica-se, na equação 36 e na figura 10, que o nível de significância conjunto aumenta, a medida que aumenta o número de testes realizados. Por exemplo, realizando-se 2 testes, adotando-se $\alpha = 0,05$ em cada um, o nível de significância conjunto será $\alpha^g = 0,0975$, com 10 testes, esse valor será de 0,4012, com 50 testes 0,9230. Uma alternativa para contornar esse problema é utilizar a chamada correção de Bonferroni, que, de acordo com Province (1999), consiste em determinar o valor do nível de significância individual (α), que proporcionará o nível de significância conjunto (α^g) desejado, dada por:

$$\alpha = -\exp\left(\frac{\ln(1 - \alpha^g)}{t}\right) + 1 \quad (30)$$

A figura 11 apresenta os resultados obtidos pela equação 37, considerando um nível de significância conjunto (α^g) igual a 0,05.

Figura 8 - Nível de significância conjunto em função do número de testes realizados (t), considerando um nível de significância individual de 0,05.

O problema de usar a correção de Bonferroni para obtenção do nível de significância individual, que proporcionaria o nível de significância conjunto desejado, é que essa correção é obtida considerando t testes independentes, o que não ocorre no mapeamento de QTLs, pois as marcas podem estar localizadas em um mesmo cromossomo. O que torna esse critério um pouco conservativo, ou seja, o verdadeiro valor de α é um pouco menor que o nominal, o que aumenta a dificuldade em rejeitar a hipótese H_0 (não existência de QTL ligado a marca). No intuito de atenuar esse problema, Churchill & Doerge (1994) propuseram o uso de um procedimento de aleatorização ("Data Shuffling") para obtenção da distribuição empírica do testador, em lugar das distribuições teóricas (F , t , χ^2), habitualmente utilizadas.

O procedimento proposto por Churchill & Doerge (1994) fundamenta-se no fato de que, se os dados obtidos indicam a existência de efeitos de QTL, pode-se destruir essa associação entre os valores observados e os marcadores, embaralhando-se os valores observados, isto é, redistribuindo aleatoriamente os valores fenotípicos observados entre os indivíduos, porém conservando os seus valores genotípicos quanto ao marcador. Por outro lado, se não houver QTLs ligados a regiões específicas do

Figura 9 - Nível de significância individual, a ser utilizado, para obtenção de um nível de significância conjunto de 0,05, em função do número de testes realizados (t), obtido, utilizando-se a correção de Bonferroni.

genoma, o processo de aleatorização não afetará a distribuição da estatística usada no teste. Decorre então, que cada amostra obtida no processo de aleatorização é essencialmente uma amostra da distribuição sob H_0 : Retirando-se N amostras e calculando-se o valor da estatística teste para cada amostra, obtém-se a distribuição empírica dessa estatística teste sob H_0 . Os autores comentam que o processo de aleatorização pode ser utilizado para a construção de dois tipos de teste, os testes individuais (pontuais) ou teste conjunto (todo o genoma).

Para determinação do ponto de corte ("threshold"), que garantirá um nível de significância α em cada um dos testes individuais, basta ordenar os N valores da estatística teste, obtidos para cada ponto a ser testado, e tomar o $100(1-\alpha)$ percentil desses valores como ponto de corte do teste, para aquela determinada marca. Já o ponto de corte, que proporcionará um nível de significância genômico α^* pode ser obtido tomando o valor máximo da estatística teste obtida dentre todos os pontos do genoma, em cada uma das N aleatorizações. O $100(1-\alpha^*)$ percentil desses valores será, assim, o ponto de corte para cada teste pontual, que proporcionará um nível

de significância genômico α .

Para contornar alguns dos problemas encontrados nos testes de comparações múltiplas, Benjamin & Hocheberg (1995) propuseram controlar a "razão de falsas descobertas" (FDR "false discovery rate"), definida como sendo a proporção de hipóteses nulas (H_0) verdadeiras, entre as hipóteses nulas rejeitadas, ou seja, a proporção de erros devidos à falsa rejeição de H_0 , também chamada proporção de falsos positivos.

Para melhor compreensão desse teste, considere-se que sejam testadas t hipóteses H_0 , das quais um determinado número (t_0) seja verdadeira, e que R das t hipóteses foram rejeitadas. A tabela 11 resume a situação apresentada

Nessa tabela, R e t são variáveis observáveis, U, V, S, W e t_0 são variáveis aleatórias não observáveis. Em termos dessas variáveis aleatórias, o nível de significância genômico, ou FWER ("familywise error rate"), como definido pelos autores, é $P(V \geq 1)$.

A proporção de erros devidos à falsa rejeição de H_0 (proporção de falsos positivos) é dada pela variável aleatória $Q = \frac{V}{V+S} = \frac{V}{R}$; naturalmente, define-se $Q = 0$, quando $R = 0$. Q é uma variável aleatória não observável. Assim, define-se a FDR, Q_e ; como sendo a esperança matemática de Q , isto é, $Q_e = E\left[\frac{V}{V+S}\right] = E\left[\frac{V}{R}\right]$

De acordo com Benjamin e Hocheberg (1995), duas propriedades dessa razão de erros decorrem imediatamente:

1. Se todas as hipóteses H_0 forem verdadeiras, a FDR é equivalente a FWER,

Tabela 10. Número de erros cometidos ao se testarem t hipóteses

	Nao Rejeitadas	Rejeitadas	Total
H_0 verdadeiras	U	V	t_0
H_0 falsas	W	S	$t-t_0$
Total	$t-R$	R	t

nesse caso, $S = 0$ e $V = R$. Portanto, se $V = 0 \Rightarrow Q = 0$ e se $V > 0 \Rightarrow Q = 1$, ocasionando $P(V \geq 1) = E(Q) = Q_e$. Desse modo, controlar a FDR implica, a grosso modo, controlar a FWER.

2. Quando $t_0 < t$; a FDR é menor do que ou igual a FWER; nesse caso, se $V > 0$ decorre que $V=R \leq 1$; ocasionando $\hat{A}_{(V, 1)} \geq Q$: Aplicando o operador Esperança em ambos os lados tem-se: $P(V \geq 1) \geq Q_e$. Resulta daí que qualquer procedimento que controle a FWER também controla a FDR. No entanto, se um procedimento controlar apenas a FDR, ele poderá ser menos restritivo, e um ganho em poder deverá ser esperado. Em particular, quanto maior o número de hipóteses falsas, maior tende a ser S , e, conseqüentemente, a diferença entre as razões de erros (FDR e FWER). Decorre, pois, que o potencial de aumento do poder é tanto maior quando maior for o número de hipóteses falsas. Em análise de QTLs, isso equivale a dizer que quanto maior o número de QTLs segregando, maior será o poder.

Controlar a variável aleatória Q em cada teste é muito desejável. Porém, isso é impossível, pois, se $t_0 = t$ e se ao menos uma única hipótese for rejeitada, $V/R = 1$ e Q não poderá ser controlada. Controlar $V=R \mid R > 0$ apresenta o mesmo problema, ou seja, é igual a um, na mesma situação; conseqüentemente, $E(V=R \mid R > 0)$ não pode ser controlada. A FDR é, então, $P(R > 0)E(V=R \mid R > 0)$, que pode ser controlado, como demonstrado por Benjamin & Hocheberg (1995).

O procedimento para determinar o ponto de corte em testes múltiplos, controlando a FDR, pode ser realizado do seguinte modo (Benjamin & Hocheberg, 1995): para cada uma das hipóteses a serem testadas $H_{01}; H_{02}; \dots; H_{0t}$, obter o valor da estatística teste e o correspondente valor de p (probabilidade sob a hipótese H_0 ; de obter um valor igual ou superior ao obtido para estatística teste), $P_1, P_2, \dots, P_t$. Em seguida, ordenar os valores P_i : Seja $P_{(1)}, P_{(2)}, \dots, P_{(t)}$ os valores de P ordenados e $H_{0(i)}$ a hipótese correspondente. Definindo $q = \frac{tP_{(i)}}{i}$; a FDR pode ser controlada

em um nível q^α , determinando o maior i , para o qual:

$$q^\alpha \geq \frac{tP_{(i)}}{i} \quad (31)$$

A título de ilustração do procedimento, considere-se o exemplo apresentado por Benjamin & Hocheberg (1995), em que foram realizados 15 testes de hipóteses, e que os $P_{(i)}$ s, ordenados, tenham sido: 0,0001; 0,0004; 0,0019; 0,0095; 0,0201; 0,0278; 0,0298; 0,0344; 0,0495; 0,3240; 0,4262; 0,5719; 0,6528; 0,7590; 1,0000.

Considere-se, ainda, que se deseje obter um nível de significância conjunto, $\alpha = 0,05$, que equivale a uma FDR de 5%, caso todas as hipóteses sejam verdadeiras, tem-se que:

$$q^\alpha = 0,05 \Rightarrow 0,05 \geq \frac{15P_{(i)}}{i}$$

Assim, para $P_{(4)} = 0,0095 \Rightarrow 0,05 \geq 0,0356$; $P_{(5)} = 0,0201 \Rightarrow 0,05 \leq 0,0603$; portanto, devem-se rejeitar todas as hipóteses com p-valores menores ou iguais a 0,0095.

Weller et al. (1996) compararam a utilização da FDR versus o controle do nível e significância individual e conjunto, utilizando a metodologia de aleatorização proposta por Churchill & Doerge (1994). Utilizaram dados referentes à produção de 1400 vacas US Holstein, filhas de 17 touros, genotipadas para 26 microssatélites, e concluíram que, ao se controlar a FDR, o número de hipóteses rejeitadas, isto é, o número de QTLs detectados, é uma função do número atual de QTLs segregando na população, o que não ocorre quando se controlam os níveis de significância conjunto ou individual. A FDR pode ser controlada mesmo para testes correlacionados e pode ser, também, controlada no mesmo nível, quer para todo o genoma ou parte deste. Os autores sugerem que a FDR seja utilizada para uma varredura inicial do genoma à procura de QTLs.

Em análises de ligação gênica, o nível de significância para os testes é escolhido de modo que seja baixa a proporção de falsos positivos, dentre os resulta-

dos significativos. Soutthey & Fernando(1996) defendem a utilização dessa mesma técnica, para detecção de QTLs, e afirmam que uma vantagem dessa técnica é o fato de que o nível de significância e o poder do teste independem do número de testes realizados (t), como demonstrado a seguir, utilizando um modelo de regressão múltipla em um retrocruzamento (Soutthey & Fernando, 1996)

Considere-se um retrocruzamento obtido a partir de duas linhagens, A e B. Os dados obtidos nesse retrocruzamento podem ser analisados pelo seguinte modelo(Zeng, 1993):

$$y_{ijkl} = \mu_0 + X_{Li} \mu_L + X_{Cj} \mu_C + X_{Rk} \mu_R + e_{ijkl};$$

em que:

y_{ijkl} é o valor observado;

μ_0 é o intercepto;

μ_L ; μ_C ; μ_R são os coeficientes de regressão para as três marcas;

X_{Li} ; X_{Cj} e X_{Rk} são variáveis dummy, assumindo os valores 0 ou 1, dependendo se o genótipo do marcador for heterozigoto ou homozigoto;

C é o marcador central, L e R são os marcadores flancuadores, situados à esquerda e direita de C, respectivamente;

e_{ijkl} é o erro associado à observação y_{ijkl} ;

Como demonstrado por Zeng (1993), μ_C é diferente de zero, somente se existir um QTL entre L e R, portanto, testar a hipótese de não existência de QTL, no intervalo flancuado por L e R, equivale a testar $H_0 : \mu_C = 0$. Dada a localização e a magnitude de um QTL, o poder para este teste é calculado por $P(Z < Z_{\frac{\alpha}{2}} - \frac{\mu_C}{\sqrt{\frac{1}{4}}})$, em que Z é a variável normal padronizada, $Z_{\frac{\alpha}{2}}$ é o valor crítico para um teste bilateral no nível de significância α , e $\sqrt{\frac{1}{4}}$ é o desvio padrão da estimativa de μ_C :

De acordo com Morton (1955), citado por Soutthey & Fernando (1996), a proporção de falsos positivos entre resultados significativos para um intervalo $i(p_i)$ é dada por:

$$p_i = \frac{\alpha_i \Pr(H_{0i})}{\alpha_i \Pr(H_{0i}) + \Pr(H_{ai})P_i} \quad (32)$$

em que:

α_i é o nível de significância,

$\Pr(H_{0i})$ é a probabilidade a priori da hipótese de nulidade,

$\Pr(H_{ai})$ é a probabilidade a priori da hipótese alternativa, e

P_i é o poder médio do teste sobre a localização do QTL no intervalo.

Se t intervalos forem testados em um "screening" genômico, a proporção genômica de falsos positivos entre os resultados significativos será:

$$p = \frac{\sum_{i=1}^t \alpha_i \Pr(H_{0i})}{\sum_{i=1}^t (\alpha_i \Pr(H_{0i}) + \Pr(H_{ai})P_i)} \quad (33)$$

Assumindo a mesma informação a priori, nível de significância e poder médio para cada intervalo, p reduz-se a:

$$p = \frac{t\alpha \Pr(H_0)}{t(\alpha \Pr(H_0) + \Pr(H_a)P)} \quad (34)$$

que claramente independe do número de intervalos testados (t).

Southey & Fernando (1996) obtiveram os valores de p em diferentes situações, utilizando simulação computacional de dados. Para tanto, assumiram um genoma de 30 Morgans, o qual foi dividido em k intervalos igualmente espaçados, nos quais QTLs foram aleatoriamente distribuídos, sob a restrição de que um intervalo não poderia conter mais de um QTL. A $\Pr(H_a)$ foi calculada como sendo o número de QTLs dividido por k , e $\Pr(H_0) = 1 - \Pr(H_a)$. No cálculo de P , os autores assumiram a ausência de interferência, QTLs de iguais tamanhos, e todas as localizações de um QTL, em um intervalo equiprováveis. Variando a herdabilidade do caráter, número de progenies e nível de significância, concluíram que tanto o erro tipo I quanto o tipo II devem ser baixos, para manter o valor de p em níveis aceitáveis

e manter um poder de detecção de QTLs adequado. Não foi encontrado um nível de significância único que fosse apropriado a todas as situações, devido à redução no poder, principalmente, nas situações de baixa herdabilidade e pequeno número de progênies avaliadas. Para um número grande de progênies, muitas das pressuposições tornaram-se sem importância, incluindo o número de QTLs e a magnitude dos efeitos individuais dos QTLs. A grande vantagem de controlar a proporção de falsos positivos é que o mesmo nível de significância pode ser usado para os testes simples ou múltiplos, pois o poder do teste independe do número de testes realizados, o que não ocorre quando se controla a probabilidade de erro tipo I.

Além do nível de significância, outro aspecto que deve ser considerado, na análise de QTLs, refere-se ao método biométrico a ser utilizado, pois a associação marcador-QTL pode ser testada por meio de um teste t ou análise de variância, para cada marca individualmente, ou via um modelo de regressão múltipla, incluindo todas as marcas, utilizando de um procedimento de eliminação de variáveis, como o "forward", o "backward" ou o "stepwise", para o descarte das marcas não informativas.

O procedimento "forward" inicia-se considerando o modelo só como o intercepto, em seguida, procura-se o "melhor" submodelo pela inserção de regressores (marcas) ao modelo, uma de cada vez. A primeira marca a ser adicionada ao modelo será aquela que apresentar maior correlação simples com a variável resposta (y), por exemplo, a marca1 (m1). A segunda marca candidata a ser incluída no modelo será aquela que apresentar maior correlação com y, ajustado para o efeito da primeira marca ($y = x^-$), denominada correlação parcial ($r_{x,y:m1}$): Suponha-se que, no segundo passo, a marca que apresentou maior correlação parcial com y fosse a marca2 (m2). Essa marca será incluída no modelo se o valor da estatística f:

$$f = \frac{SQRes_{duo}(m2 | m1)}{QMRes_{duo}(m1; m2)} \quad (35)$$

exceder um valor f a um determinado nível de significância. Em seguida, repete-se o processo até que nenhuma outra marca seja significativa.

Uma desvantagem desse procedimento é que, após uma marca ser incluída no modelo, ela não pode ser retirada, e, muitas das vezes, após a entrada de outra marca, sua informação torna-se redundante, devido à alta correlação entre as mesmas.

O procedimento "forward" inicia-se com um modelo sem nenhuma marca, e vai inserindo marcas a cada passo, já o "backward" atua em sentido oposto, isto é, inicia-se com um modelo que inclui todas as t marcas candidatas. A cada passo, a estatística f parcial é calculada para cada marca, como se ela tivesse sido a última incluída no modelo pelo processo "forward". O menor desses valores de f parcial é então comparado com o valor $f_{\text{saída}}$, e, se o valor de f parcial for menor que $f_{\text{saída}}$, a marca associada é eliminada do modelo. Em seguida, um modelo com as $t-1$ variáveis restantes é ajustado e repete-se o processo até que nenhuma marca seja mais eliminada.

O procedimento de eliminação de variáveis "backward" é considerado um procedimento muito bom, sendo defendido por pesquisadores que desejem verificar o efeito da inclusão de todas as marcas no modelo. Porém, como o procedimento inicia-se com o modelo completo, nos seus primeiros passos, podem existir poucos graus de liberdade para o denominador da estatística f , o que não é estatisticamente desejável.

O procedimento "stepwise" é uma modificação do "forward", no qual, a cada passo, todas as marcas previamente incluídas no modelo são testadas via a estatística f parcial, para sua permanência. Deste modo, um marcador previamente adicionado ao modelo pode ser eliminado em um passo subsequente, por ter-se tornado redundante devido à inclusão de outro(s) marcador(es). Verifica-se, portanto, que, no procedimento "stepwise", é necessária a definição de dois f , um para entrada de marcas no modelo, f_{entrada} ; e outro para sua permanência, $f_{\text{saída}}$. Manipulando os níveis de significância para entrada de variáveis no modelo e para sua permanência, o pesquisador possui um maior controle sobre o método. Por meio da escolha adequada dos níveis de significância para entrada e saída de variáveis do modelo, o

procedimento "stepwise" pode conduzir ao mesmo modelo obtido pelo "backward", muito embora todos os métodos descritos tendam a conduzir a "melhores" modelos diferentes.

Kao et al. (1999), ao apresentarem o método de mapeamento por intervalo ("multiple interval mapping"), advogam o uso da metodologia "stepwise", utilizando o teste da razão de verossimilhança para escolha das marcas que entrarão no processo de mapeamento, uma vez que é computacionalmente inviável fazer o mapeamento múltiplo com um número elevado de marcas.

Maiores informações sobre procedimentos para seleção de variáveis em regressão múltiplas são dadas por Montgomery & Peck (1992)

2.5 Interação QTLs x Ambiente

Segundo Lynch & Walsh (1998), outro problema que permanece não resolvido na análise de QTLs é a ocorrência da interação QTLs x ambientes. Duas metodologias têm sido utilizadas no estudo dessas interações: a consistência das associações marcador-caráter através dos ambientes e métodos de ANOVA, incorporando termos específicos para a referida interação. A primeira é muito incipiente, pois consiste simplesmente em verificar se uma associação marca-caráter é detectada em todos os ambientes avaliados. Em caso afirmativo, isto é tomado como evidência de não existência de interação marcadores x ambientes, enquanto que a detecção de associação em apenas alguns dos ambientes é sempre considerada como evidência de interação, embora um QTL possa ter efeito significativo em todos os ambientes mesmo na presença de forte interação. De forma semelhante, baixos poderes de detecção podem resultar em que um QTL venha a ser detectado em apenas algumas repetições do experimento, mesmo que seus efeitos sejam idênticos em todos os ambientes. Fortalecendo essa explicação, Koester et al. (1993) observaram que QTLs com pequenos efeitos eram menos propensos de serem detectados através dos locais, do que QTLs de pronunciado efeito.

Lynch & Walsh (1998) afirmam que, pelo teste explícito da interação

marcadores x ambientes, os métodos de ANOVA provêm uma medida mais sensível dos efeitos dessa interação. Atuando nessa linha, Cockerham & Zeng (1996) estendem a análise de variância do delineamento III de Comstock & Robinson (1952) para incluir os efeitos de ligação entre marcadores moleculares e QTLs, epistasia entre dois locos e utilização de gerações F_3 . Além de desenvolverem uma metodologia baseada em contrastes ortogonais, para estimação e teste de hipóteses sobre esses efeitos bem como suas interações com o ambiente.

Algumas metodologias de abordagem do problema da interação QTLs x ambientes no mapeamento de QTLs têm sido propostas, dentre essas pode-se destacar as apresentadas por:

Wang et al. (1999), que sugerem a utilização da metodologia de modelos mistos para o mapeamento de QTLs, envolvendo os efeitos de epistasia e interação QTLs x ambientes, argumentando que a grande vantagem dessa abordagem consiste na possibilidade do desdobramento da interação QTLs x ambientes em seus componentes devidos à interação de efeitos aditivos x ambientes e epistasia x ambientes.

Korol et al. (1998), que propõem uma metodologia de mapeamento, considerando a interação QTLs x ambientes, na qual o número de parâmetros envolvidos é pouco influenciado pelo aumento do número de ambientes. Para tanto, sugerem expressar o valor da substituição alélica no QTL por meio de uma regressão polinomial da média dos genótipos através dos ambientes. Desse modo, obtêm-se os valores fenotípicos estimados para os indivíduos que serão mapeados em cada ambiente.

Metodologia semelhante, porém utilizando a abordagem AMMI, ("Additive Main Effects and Multiplicative Interaction"), é apresentada por Romagosa et al. (1996), que sugerem a decomposição dos efeitos da interação em seus componentes multiplicativos via análise AMMI, e a realização do mapeamento, utilizando como valores fenotípicos, os valores preditos dos genótipos obtidos via análise AMMI.

Jansen et al. (1995) sugerem que o mapeamento seja realizado em duas etapas; na primeira, utiliza-se um modelo de regressão múltipla, incluindo todas

as marcas e efeitos de interações marcas x ambientes, e, por meio de um processo de seleção de variáveis, como o "backward", eliminam-se as marcas e interações não significativas. Em seguida, realiza-se o mapeamento por intervalo composto, utilizando todas as marcas selecionadas no procedimento "backward", como cofatores, ignorando-se as interações. No caso das regiões cromossômicas que apresentaram interação QTLs x ambientes, realiza-se o mapeamento para cada ambiente separadamente. Um problema dessa metodologia é que o número de parâmetros a serem incluídos no modelo inicial aumenta, substancialmente, com o aumento do número de ambientes avaliados.

As metodologias encontradas na literatura, para abordagem da interação QTLs x ambientes, consideram-na como um complicador da análise de QTLs e atuam no sentido de eliminar seu efeito sobre o seu mapeamento. Mas, nem sempre, a interação genótipos x ambientes é um complicador para o trabalho do melhorista, pois, quantificando-se o seu efeito, podem-se selecionar genótipos que sejam altamente adaptados a um determinado ambiente, captando, assim, o seu efeito. Ou, então, selecionar os genótipos que não apresentem essa interação, isto é, apresentam um comportamento estável nos ambientes em que são cultivados. Esse entendimento pode ser transportado para o nível de QTLs, ou seja, mediante a quantificação dos efeitos da interação QTLs x ambientes, é possível identificar aqueles mais estáveis, que podem ser muito úteis em programas de seleção assistida por marcadores moleculares, visando à produção de materiais com alta estabilidade, e os QTLs que apresentam alta interação, passíveis de serem utilizados em programas de seleção assistida visando à obtenção de genótipos específicos para um determinado tipo de ambiente.

Deste modo, um dos objetivos deste trabalho consiste em obter um estimador da proporção da variância da interação genótipos x ambientes, devida à interação QTLs x ambientes, bem como avaliar algumas metodologias para detecção de QTLs, quanto ao poder de identificação dos QTLs mais estáveis e dos que apresentam maior interação com os ambientes.

3 TESTES MÚLTIPLOS E PODER DE DETECÇÃO DE QTL's

Resumo

Avaliou-se, via simulação computacional de dados, o poder de detecção QTLs da análise de marcas simples, realizada por meio de regressão linear múltipla, empregando o procedimento "stepwise" para seleção das marcas, com procedimentos baseados em testes individuais, utilizando os critérios FDR e de Bonferroni para determinação do nível de significância conjunto (α). Os resultados mostraram que, para detecção de QTL's, utilizando a metodologia de análise de marcas simples, o procedimento baseado em regressão múltipla, empregando o procedimento "stepwise" para seleção de variáveis, foi mais poderoso, em identificar as marcas associadas a QTL's, do que os procedimentos baseados em testes individuais, utilizando tanto o critério FDR, quanto o de Bonferroni. Mesmo nos casos em que esse procedimento apresentou poder ligeiramente inferior aos demais, mostrou a grande vantagem de selecionar apenas as marcas mais fortemente ligadas ao QTL, devendo, portanto, ser preferido para seleção das marcas a serem utilizadas como covariáveis no processo de mapeamento por intervalo múltiplo. Dentre os critérios FDR e de Bonferroni, o primeiro mostrou-se, em geral mais poderoso, devendo ser adotado nos procedimentos de mapeamento por intervalo.

Summary

In general terms, QTL mapping differs from other research activities in genetics. Being basically a multiple test procedure, problems arise which are related to the joint level of significance of the analysis, and consequently, to its power. Using computational simulation of data, the power of simple marker analysis, carried out through multiple linear regression using stepwise procedures to select the markers was obtained. Procedures based on single tests, using both the FDR and the Bonferroni criteria to determinate the joint level of significance were also used. Results showed that the procedure based on multiple regression, using the stepwise technique, was the most powerful in identifying markers associated to QTL's. However, in cases where its power was smaller, its advantage was the ability to detect only markers strongly associates with QTL's. In comparision with the Bonferroni method, the FDR criterion was in general more powerful, and should be adopted in the interval mapping procedures.

3.1 Introdução

O mapeamento de QTLs difere dos demais tipos de experimentos conduzidos em genética, por tratar-se, basicamente, de um procedimento de testes múltiplos, uma vez que consiste em distribuir marcadores moleculares por todo o genoma e, em seguida, realizar uma varredura ou "screening" dessas marcas, no intuito de verificar qual(is) está(ão) ligada(s) a QTLs, no caso de estar-se utilizando a análise de marcas simples, ou de posições relativas no intervalo entre duas marcas ao utilizar-se o mapeamento por intervalo (Lande & Bostein, 1989). Já no caso no mapeamento por intervalo composto (Jansen, 1992,1993 e Zeng, 1993,1994), primeiramente realiza-se o "screening" das marcas, no intuito de se identificar quais serão usadas para controlar o "back ground" genético, durante o mapeamento de um

dado intervalo, e, em seguida procede-se ao mapeamento do referido intervalo.

O problema encontrado nesses procedimentos refere-se ao nível de significância conjunto do teste, também chamado nível de significância genômico, visto que o nível de significância individual (de cada teste) é definido pelo pesquisador.

Nível de significância é definido, em estatística, como a probabilidade de se cometer um erro Tipo I (rejeitar uma hipótese nula verdadeira). No presente caso, a hipótese de nulidade testada é a de não existência de QTL associado a marca. No mapeamento de QTLs, cometer um erro Tipo I implica admitir a existência de um QTL inexistente (QTL fantasma).

Suponha-se que, em um determinado genoma, sejam distribuídas aleatoriamente t marcas. Assim, ao proceder a varredura dessas marcas a procura de QTLs, serão realizados t testes de marcas simples (um para cada marca), adotando-se um nível de significância (α) para cada teste. Tem-se que o nível de significância conjunto do teste, ou genômico (α^c), considerando os t testes independentes será:

α^c = probabilidade de rejeitar pelo menos uma hipótese nula verdadeira, ou seja,

α^c = 1- probabilidade de não rejeitar nenhuma hipótese nula verdadeira, isto é,

$$\alpha^c = 1 - (1 - \alpha)^t \quad (36)$$

Desse modo, verifica-se, na equação 36 e na figura 10, que o nível de significância conjunto cresce, à medida que aumenta o número de testes realizados. Por exemplo, realizando-se 2 testes, adotando-se $\alpha = 0,05$ em cada um, o nível de significância conjunto será $\alpha^c = 0,0975$, com 10 testes, esse valor será de 0,4012, com 50 testes 0,9230. Uma alternativa para contornar esse problema é utilizar a chamada correção de Bonferroni, que, de acordo com Province (1999), consiste em determinar o valor do nível de significância individual (α), que proporcionará o nível de significância conjunto (α^c) desejado, dado por:

Figura 10 - Nível de significância conjunto em função do número de testes realizados (t), considerando um nível de significância individual de 0,05.

$$\alpha = -\exp\left(\frac{\ln(1 - \alpha^t)}{t}\right) + 1: \quad (37)$$

A Figura 11 apresenta os resultados obtidos pela equação 37, considerando um nível de significância conjunto (α) igual a 0,05.

O problema de usar a correção de Bonferroni, para obtenção do nível de significância individual, que proporcionaria o nível de significância conjunto desejado, é que essa correção é obtida considerando t testes independentes, o que não ocorre no mapeamento de QTLs, pois as marcas podem estar localizadas em um mesmo cromossomo. Isso torna esse critério um pouco conservativo, ou seja, o verdadeiro valor de α é um pouco menor que o nominal, o que aumenta a dificuldade em rejeitar a hipótese H_0 (não existência de QTL ligado a marca). No intuito de atenuar esse problema, Churchill & Doerge (1994) propuseram o uso de um procedimento de aleatorização ("Data Shuffling") para obter a distribuição empírica do testador, em lugar das distribuições teóricas (F , t , χ^2), habitualmente utilizadas.

O procedimento proposto por Churchill & Doerge (1994) fundamenta-se no fato de que, se os dados obtidos indicam a existência de efeitos de QTL, pode-se

Figura 11 - Nível de significância individual, a ser utilizado, considerando a correção de Bonferroni para obtenção de um nível de significância conjunto de 0,05, em função do número de testes realizados (t).

destruir tal associação entre os valores observados e os marcadores, embaralhando-se os valores observados, isto é, redistribuindo aleatoriamente os valores fenotípicos observados entre os indivíduos, porém conservando os seus valores genotípicos quanto ao marcador. Por outro lado, se não houver QTLs ligados a regiões específicas do genoma, o processo de aleatorização não afetará a distribuição da estatística usada no teste. Decorre, então, que cada amostra obtida no processo de aleatorização é essencialmente uma amostra da distribuição sob H_0 . Retirando-se N amostras e calculando-se o valor da estatística teste para cada amostra, obtém-se a distribuição empírica dessa estatística teste sob H_0 . Os autores comentam que o processo de aleatorização pode ser utilizado para a construção de dois tipos de teste, os testes individuais (pontuais) ou o teste conjunto (todo o genoma).

Para determinação do ponto de corte ("threshold"), que garantirá um nível de significância α em cada um dos testes individuais, basta ordenar os N valores da estatística teste, obtidos para cada ponto a ser testado, e tomar o $100(1-\alpha)$ percentil desses valores como ponto de corte do teste, para aquela determinada marca.

Se o ponto de corte, que proporcionará um nível de significância genômico α , pode ser obtido tomando o valor máximo da estatística teste obtida dentre todos os pontos do genoma, em cada uma das N aleatorizações. O $100(1 - \alpha)$ percentil desses valores será, assim, o ponto de corte para cada teste pontual, que proporcionará um nível de significância genômico α .

Para contornar alguns dos problemas encontrados nos testes de comparações múltiplas, Benjamin & Hochberg (1995) propuseram controlar a "razão de falsas descobertas" (FDR "false discovery rate"), definida como sendo a proporção de hipóteses nulas (H_0) verdadeiras, entre as hipóteses nulas rejeitadas, ou seja, a proporção de erros devidos à falsa rejeição de H_0 , também chamada proporção de falsos positivos.

Para melhor compreensão desse teste, considere-se que sejam testadas t hipóteses H_0 , das quais um determinado número (t_0) sejam verdadeiras, e que R das t hipóteses foram rejeitadas. A tabela 11 resume a situação apresentada

Nessa tabela, R e t são variáveis observáveis, U, V, S, W e t_0 são variáveis aleatórias não observáveis. Em termos dessas variáveis aleatórias, o nível de significância genômico, ou FWER ("familywise error rate"), como definido pelos autores, é $P(V \geq 1)$.

A proporção de erros devidos à falsa rejeição de H_0 (proporção de falsos positivos) é dada pela variável aleatória $Q = \frac{V}{V+S} = \frac{V}{R}$; naturalmente, define-se $Q = 0$, quando $R = 0$. Q é uma variável aleatória não observável. Assim, define-se a

Tabela 11. Número de erros cometidos ao se testarem t hipóteses

	Não Rejeitadas	Rejeitadas	Total
H_0 verdadeiras	U	V	t_0
H_0 falsas	W	S	$t - t_0$
Total	$t - R$	R	t

FDR, Q_e ; como sendo a esperança matemática de Q , isto é, $Q_e = E \left[\frac{V}{V+S} \right] = E \left[\frac{V}{R} \right]$

De acordo com Benjamin e Hocheberg (1995), duas propriedades dessa razão de erros decorrem imediatamente:

1. Se todas as hipóteses H_0 forem verdadeiras, a FDR é equivalente a FWER, nesse caso, $S = 0$ e $V = R$. Portanto, se $V = 0 \Rightarrow Q = 0$ e se $V > 0 \Rightarrow Q = 1$, ocasionando $P(V \geq 1) = E(Q) = Q_e$. Desse modo, controlar a FDR implica, a grosso modo, controlar a FWER.
2. Quando $t_0 < t$; a FDR é menor do que ou igual a FWER; neste caso, se $V > 0$ decorre que $V=R \leq 1$; ocasionando $\hat{A}_{(V, 1)} \geq Q$: Aplicando o operador Esperança em ambos os lados, tem-se: $P(V \geq 1) \geq Q_e$: Resulta daí que qualquer procedimento que controle a FWER também controlará a FDR. No entanto, se um procedimento controlar apenas a FDR, ele poderá ser menos restritivo, e um ganho em poder deverá ser esperado. Em particular, quanto maior o número de hipóteses falsas, maior tende a ser S , e, conseqüentemente, a diferença entre as razões de erros (FDR e FWER). Decorre, pois, que o potencial de aumento do poder é tanto maior quanto maior for o número de hipóteses falsas. Em análise de QTLs, isso equivale a dizer que quanto maior o número de QTLs segregando, maior será o poder.

Controlar a variável aleatória Q em cada teste é muito desejável. Porém isso é impossível, pois, se $t_0 = t$ e se ao menos uma única hipótese for rejeitada, $V/R = 1$ e Q não poderá ser controlada. Controlar $V=R \mid R > 0$ apresenta o mesmo problema, ou seja, é igual a um, na mesma situação; conseqüentemente, $E(V=R \mid R > 0)$ não pode ser controlada. A FDR é, então, $P(R > 0)E(V=R \mid R > 0)$, que pode ser controlado, como demonstrado por Benjamin & Hocheberg (1995).

O procedimento para determinar o ponto de corte em testes múltiplos, controlando a FDR, pode ser realizado do seguinte modo (Benjamin & Hocheberg, 1995): para cada uma das hipóteses a serem testadas $H_{01}; H_{02}; \dots; H_{0t}$, obter o valor

da estatística teste e o correspondente valor de p (probabilidade sob a hipótese H_0 ; de obter um valor igual ou superior ao encontrado para estatística teste), $P_1, P_2, \dots, P_t$. Em seguida, ordenar os valores P_i : Seja $P_{(1)}, P_{(2)}, \dots, P_{(t)}$ os valores de P ordenados e $H_{0(i)}$ a hipótese correspondente. Definindo $q = \frac{tP_{(i)}}{i}$; a FDR pode ser controlada em um nível q^α , determinando o maior i , para o qual:

$$q^\alpha \geq \frac{tP_{(i)}}{i} \quad (38)$$

A título de ilustração do procedimento, considere-se o exemplo apresentado por Benjamin & Hocheberg (1995), em que foram realizados 15 testes de hipóteses, e que os $P_{(i)}$ s, ordenados, tenham sido: 0,0001; 0,0004; 0,0019; 0,0095; 0,0201; 0,0278; 0,0298; 0,0344; 0,0495; 0,3240; 0,4262; 0,5719; 0,6528; 0,7590; 1,0000.

Considere-se, ainda, que se deseje obter um nível de significância conjunto, $\alpha = 0,05$, que equivale a uma FDR de 5%, caso todas as hipóteses sejam verdadeiras; tem-se que:

$$q^\alpha = 0,05 \Rightarrow 0,05 \geq \frac{15P_{(i)}}{i}.$$

Assim, para $P_{(4)} = 0,0095 \Rightarrow 0,05 \geq 0,356$; $P_{(5)} = 0,0201 \Rightarrow 0,05 \leq 0,603$; portanto, devem-se rejeitar todas as hipóteses com p -valores menores ou iguais a 0,0095.

Weller et al. (1996) compararam a utilização da FDR versus o controle do nível e significância individual e conjunto, utilizando a metodologia de aleatorização proposta por Churchill & Doerge (1994). Utilizaram dados referentes à produção de 1400 vacas US Holstein, filhas de 17 touros, genotipadas para 26 microssatélites, e concluíram que, ao se controlar a FDR, o número de hipóteses rejeitadas, isto é, o número de QTLs detectados, é uma função do número atual de QTLs segregando na população, o que não ocorre quando se controlam os níveis de significância conjunto ou individual. A FDR pode ser controlada mesmo para testes correlacionados e pode ser, também, controlada no mesmo nível, quer para todo o

genoma ou parte deste. Os autores sugerem que a FDR seja utilizada para uma varredura inicial do genoma à procura de QTLs.

Em análises de ligação gênica, o nível de significância dos testes é escolhido de modo que seja baixa a proporção de falsos positivos, dentre os resultados significativos. Soutthey & Fernando(1996) defendem a utilização dessa mesma técnica, para detecção de QTLs, e afirmam que uma vantagem dessa técnica é o fato de que o nível de significância e o poder do teste independem do número de testes realizados (t), como demonstrado a seguir, utilizando um modelo de regressão múltipla em um retrocruzamento (Soutthey & Fernando, 1996)

Considere-se um retrocruzamento obtido a partir de duas linhagens, A e B. Cujos resultados podem ser analisados pelo seguinte modelo(Zeng, 1993):

$$y_i = \mu_0 + X_{Li}\mu_L + X_{Ci}\mu_C + X_{Ri}\mu_R + e_i;$$

em que:

y_i é o valor observado do i -ésimo indivíduo, quanto ao caráter em questão;

μ_0 é o intercepto;

μ_L ; μ_C ; μ_R são os coeficientes de regressão para as três marcas;

X_{Li} ; X_{Ci} e X_{Ri} são variáveis dummy, assumindo os valores 0 ou 1, dependendo se o genótipo do marcador for heterozigótico ou homozigótico;

C é o marcador central, L e R são os marcadores âncoradores, situados à esquerda e direita de C, respectivamente;

e_i é o erro associado à observação y_i ;

Como demonstrado por Zeng (1993), μ_C é diferente de zero, somente se existir um QTL entre L e R; portanto, testar a hipótese de não existência de QTL, no intervalo âncorado por L e R, equivale a testar $H_0 : \mu_C = 0$. Dada a localização e a magnitude de um QTL, o poder para este teste é calculado por $P(Z < Z_{\frac{\alpha}{2}} - \frac{\mu_C}{\frac{1}{4} - \frac{1}{4}})$, em que Z é a variável normal padronizada, $Z_{\frac{\alpha}{2}}$ é o valor crítico para um teste bilateral no nível de significância α , e $\frac{1}{4} - \frac{1}{4}$ é o desvio padrão da estimativa de μ_C :

De acordo com Morton (1955), citado por Soutthey & Fernando (1996), a proporção de falsos positivos entre resultados significativos para um intervalo i (p_i) é dada por:

$$p_i = \frac{\alpha_i \Pr(H_{0i})}{\alpha_i \Pr(H_{0i}) + \Pr(H_{ai})P_i} \quad (39)$$

em que:

α_i é o nível de significância,

$\Pr(H_{0i})$ é a probabilidade a priori da hipótese de nulidade,

$\Pr(H_{ai})$ é a probabilidade a priori da hipótese alternativa, e

P_i é o poder médio do teste sobre a localização do QTL no intervalo.

Se t intervalos forem testados em um "screening" genômico, a proporção genômica de falsos positivos entre os resultados significativos será:

$$p = \frac{\sum_{i=1}^t \alpha_i \Pr(H_{0i})}{\sum_{i=1}^t (\alpha_i \Pr(H_{0i}) + \Pr(H_{ai})P_i)} \quad (40)$$

Assumindo a mesma informação a priori, nível de significância e poder médio para cada intervalo, p reduz-se a:

$$p = \frac{t\alpha \Pr(H_0)}{t(\alpha \Pr(H_0) + \Pr(H_a)P)} \quad (41)$$

que claramente independe do número de intervalos testados (t).

Soutthey & Fernando (1996) obtiveram os valores de p em diferentes situações, utilizando simulação computacional de dados. Para tanto, assumiram um genoma de 30 Morgans, o qual foi dividido em k intervalos igualmente espaçados, nos quais QTLs foram aleatoriamente distribuídos, sob a restrição de que um intervalo não poderia conter mais de um QTL. A $\Pr(H_a)$ foi calculada como sendo o número de QTLs dividido por k , e $\Pr(H_0) = 1 - \Pr(H_a)$. No cálculo de P , os autores assumiram

a ausência de interferência, QTLs de iguais tamanhos, e todas as localizações de um QTL, em um intervalo equiprováveis. Variando a herdabilidade do caráter, número de progênies e nível de significância, concluíram que tanto o erro tipo I quanto o tipo II devem ser baixos, para manter o valor de p em níveis aceitáveis e manter um poder de detecção de QTLs adequado. Não foi encontrado um nível de significância único que fosse apropriado a todas as situações, devido à redução no poder, principalmente, nos casos em que herdabilidade é baixa, e utilizou-se um pequeno número de progênies. Para um número grande de progênies, muitas das pressuposições tornaram-se sem importância, incluindo o número de QTLs e a magnitude dos efeitos individuais dos QTLs. A grande vantagem de controlar a proporção de falsos positivos é que o mesmo nível de significância pode ser usado para os testes simples ou múltiplos, pois o poder do teste independe do número de testes realizados, o que não ocorre quando se controla a probabilidade de erro tipo I.

Além do nível de significância, outro aspecto que deve ser considerado, na análise de QTLs, refere-se ao método biométrico a ser utilizado, pois a associação marcador-QTL pode ser testada por meio de um teste t ou análise de variância, para cada marca individualmente, ou via um modelo de regressão múltipla, incluindo todas as marcas, empregando um procedimento de eliminação de variáveis, como o "forward", o "backward" ou o "stepwise", para o descarte das marcas não informativas.

O procedimento "forward" é iniciado considerando o modelo só como o intercepto, em seguida, procura-se o "melhor" submodelo pela inserção de regressores (marcas) ao modelo, uma de cada vez. A primeira marca a ser adicionada ao modelo será aquela que apresentar maior correlação simples com a variável resposta (y), por exemplo, a marca 1 (m_1). A segunda marca candidata a ser incluída no modelo será aquela que apresentar maior correlação com y , ajustado para o efeito da primeira marca ($\hat{y} = x_1$), denominada correlação parcial ($r_{x_2:y:m_1}$): Suponha-se que, no segundo passo, a marca que apresentou maior correlação parcial com y foi a marca 2 (m_2). Esta marca será incluída no modelo se o valor da estatística F :

$$f = \frac{SQRes_{\text{duo}}(m2 | m1)}{QMRes_{\text{duo}}(m1; m2)} \quad (42)$$

exceder a um valor f a um determinado nível de significância. Em seguida, repete-se o processo até que nenhuma outra marca seja significativa.

Uma desvantagem desse procedimento é que, após uma marca ser incluída no modelo, ela não pode ser retirada, e, muitas vezes, após a entrada de outra marca, sua informação torna-se redundante, devido à alta correlação entre as mesmas.

O procedimento "forward" é iniciado com um modelo sem nenhuma marca, inserindo-se marcas a cada passo, já o "backward" atua em sentido oposto, isto é, inicia-se com um modelo que inclui todas as t marcas candidatas. A cada passo, a estatística f parcial é calculada para cada marca, como se ela tivesse sido a última incluída no modelo pelo processo "forward". O menor desses valores de f parcial é, então, comparado com o valor $f_{\text{saída}}$, e, se o valor de f parcial for menor que $f_{\text{saída}}$, a marca associada é eliminada do modelo. Em seguida, um modelo com as $t-1$ variáveis restantes é ajustado, e repete-se o processo até que nenhuma marca seja mais eliminada.

O procedimento de eliminação de variáveis "backward" é considerado muito bom, sendo defendido por pesquisadores que desejam verificar o efeito da inclusão de todas as marcas no modelo. Porém, como o procedimento inicia-se com o modelo completo, nos seus primeiros passos, podem existir poucos graus de liberdade para o denominador da estatística f , o que não é estatisticamente desejável.

O procedimento "stepwise" é uma modificação do "forward", no qual, a cada passo, todas as marcas previamente incluídas no modelo são testadas via a estatística F parcial, para sua permanência. Desse modo, um marcador previamente adicionado ao modelo pode ser eliminado em um passo subsequente, por ter-se tornado redundante devido à inclusão de outro(s) marcador(es). Verifica-se, portanto, que, no procedimento "stepwise", é necessária a definição de dois F , um para entrada de marcas no modelo, F_{entrada} ; e outro para sua permanência, $F_{\text{saída}}$. Manipulando os

níveis de significância para entrada de variáveis no modelo e para sua permanência, o pesquisador possui um maior controle sobre o método. Por meio da escolha adequada dos níveis de significância para entrada e saída de variáveis do modelo, o procedimento "stepwise" pode conduzir ao mesmo modelo obtido pelo "backward", muito embora todos os métodos descritos tendam a conduzir a "melhores" modelos diferentes.

Kao et al. (1999), ao apresentarem o método de mapeamento por intervalo ("multiple interval mapping"), advogam o uso da metodologia "stepwise", utilizando o teste da razão de verossimilhança para escolha das marcas que entrarão no processo de mapeamento, uma vez que é computacionalmente inviável fazer o mapeamento múltiplo com um número elevado de marcas.

Maiores informações sobre procedimentos para seleção de variáveis em regressão múltiplas são dadas por Montgomery & Peck (1992)

3.2 Metodologia

Para avaliar o poder de detecção de QTLs, da análise de marcas simples realizada por meio de regressão linear múltipla, empregando o procedimento "stepwise" para seleção das marcas, com procedimentos baseados em testes múltiplos, utilizando os critérios FDR e de Bonferroni para determinação nível de significância conjunto (α), simulou-se uma espécie hipotética com quatro cromossomos de um Morgan cada. Em cada cromossomo, foram distribuídas, sistematicamente, a partir da origem, 11 marcas espaçadas de 10 cM entre si. De acordo com Bearzoti (1997), a distribuição sistemática das marcas, ao invés de aleatória, justifica-se pelo fato de, em uma situação real, os marcadores moleculares serem, em geral, escolhidos dentre os disponíveis nos mapas de ligação, cobrindo todo o genoma da maneira mais homogênea possível.

O espaçamento de 10 cM entre as marcas foi adotado com base em Darvasi et al. (1993), que avaliaram, via simulação computacional de dados, o poder

de detecção de QTLs do teste t para marcas simples e do mapeamento por intervalo, em função do espaçamento entre as marcas utilizadas e concluíram não haver diferença para espaçamentos inferiores a 10 cM e que ocorreu apenas um pequeno decréscimo no poder, quando o espaçamento saltou de 10 cM para 20 cM.

O delineamento genético utilizado, no presente trabalho, foi uma geração F_2 , obtida a partir do cruzamento de duas linhagens homozigóticas completamente contrastantes. Para obtenção dos genótipos das progênes F_2 , considerou-se que: as permutas gênicas (crossing over) ocorrem ao acaso e não há interferência. Desse modo, como cada cromossomo possui um Morgan, o número de permutas gênicas por cromossomo segue uma distribuição Poisson com média um. Esse número de permutas gênicas (N_c) foi obtido por:

$$N_c = P^{-1}(U); \quad (43)$$

em que:

U é um número aleatório com distribuição uniforme [0,1]

P^{-1} é a inversa da distribuição Poisson com média um.

A posição na qual ocorreu o crossing over foi dada por $L = 100U$:

O modelo utilizado para simulação dos valores fenotípicos dos indivíduos (y_i) foi:

$$y_i = \mu + \sum_{t=1}^g a_t x_{t(i)} + \sum_{t=1}^g d_t (1 - x_{t(i)}^2) + e_i; \quad (44)$$

em que:

y_i é o fenotípico do i -ésimo indivíduo, $i = 1, 2, \dots, n$;

μ é a média geral;

g é o número de QTLs controlando o caráter;

a_t é o efeito aditivo do t -ésimo QTL;

$x_{t(i)}$ é uma variável codificadora do genótipo do t -ésimo QTL, assumindo os valores -1, 0 e 1, para os genótipos qq, Qq e QQ, respectivamente;

d_t é o desvio de dominância do t -ésimo QTL;

e_i é o erro aleatório.

Os erros foram simulados com distribuição normal de média $\bar{A} = 0$ e variância $\frac{3}{4}$. Esses efeitos foram obtidos por:

$$e_i = \Phi^{-1} U_i \frac{3}{4}; \quad (45)$$

em que:

Φ^{-1} representa a distribuição normal acumulada inversa;

U_i é um número aleatório com distribuição uniforme [0,1].

Na simulação do caráter quantitativo, consideraram-se as seguintes situações quanto ao controle genético deste:

1. Caráter controlado por um único QTL, com efeito aditivo $a = 1$ e dominância $d = 0$, localizado a 20 cM do início do cromossomo 1, estando completamente ligado a uma marca;
2. Caráter controlado por três QTLs independentes com efeitos aditivos $a = 1$ e dominância $d = 0$. localizados a 20 cM do início dos cromossomos 1, 2 e 3, estando cada um completamente ligado a uma marca;
3. Caráter controlado por 8 QTLs, com efeitos aditivos $a = 1$ e dominância $d = 0$, estando dois localizados no cromossomo 1, a 20 cM e 40 cM da origem deste, dois no cromossomo 2, a 20 cM e 50 cM do início do mesmo, 2 no terceiro cromossomo, a 20 cM e 60 cM do início e dois no quarto, a 20 cM e 70 cM do início do mesmo, estando cada um dos oito QTLs completamente ligados a uma determinada marca, e que se encontravam ligados em fase de associação;
4. Caráter controlado por 8 QTLs, localizados nas mesmas posições dos QTLs do item anterior, porém com QTLs localizados no mesmo cromossomo estando em fase de repulsão.

Para cada situação, consideraram-se ainda quatro valores de herdabilidade, $h^2 = 0; 10$, $h^2 = 0; 25$, $h^2 = 0; 50$ e $h^2 = 0; 75$ e três tamanhos de experimentos,

a saber: experimentos constituídos por 50, 100 e 200 indivíduos F_2 . Para cada uma dessas 48 condições, simularam-se 500 experimentos, os quais foram submetidos a análise de marcas simples, utilizando os procedimentos "stepwise", FDR e de Bonferroni.

O procedimento "stepwise" foi realizado considerando todo o genoma simultaneamente, tendo sido adotados os níveis de significância de 0,15 para entrada de uma variável (marca) no modelo e 0,05 para sua permanência.

Para os critérios FDR e Bonferroni, realizou-se uma análise de regressão linear simples, com o teste da hipótese $H_0 : \beta_1 = 0$ para cada marca, totalizando 44 testes por experimento. Para decisão quanto à rejeição da hipótese H_0 , adotou-se, no caso do critério FDR, uma frequência de falsos positivos igual a 0,05, e, no de Bonferroni, um nível de significância conjunto de 0,05, que foi obtido, considerando-se o nível de significância individual (por teste) de 0,00165.

O poder de um teste é a probabilidade que o teste apresenta de rejeitar uma hipótese H_0 falsa, deste modo, o poder de cada metodologia foi calculado como sendo a frequência relativa de hipóteses H_0 rejeitadas na posição em que os QTLs se localizavam.

3.3 Resultados e Discussão

Na Figura 12, encontram-se apresentadas as frequências relativas de hipóteses H_0 rejeitadas, pelos procedimentos "stepwise", FDR e Bonferroni em 500 experimentos, em função da herdabilidade do caráter e número de indivíduos avaliados, para um caráter controlado por um único QTL. Lembrando que a hipótese H_0 , aqui testada, refere-se a ausência de QTL ligado à marca em questão, e que, na simulação desse caráter, considerou-se este sendo controlado por um único QTL, localizado a 20 cM da origem do cromossomo 1, essa frequência relativa é, então, nesta posição do cromossomo, o poder do teste, e, nas demais posições, a frequência de falsos positivos ou probabilidade de identificação de QTLs fantasmas.

Figura 12 - Frequência relativa de hipóteses rejeitadas, pelos procedimentos "stepwise", FDR e Bonferroni em 500 experimentos, em função da herdabilidade do caráter e número de indivíduos avaliados, para um caráter controlado por um único QTL. Q representa a localização do QTL.

Nessa Figura, verifica-se que, quando se utilizou uma geração F_2 , para análise de um caráter com herdabilidade igual a 0,10, o procedimento "stepwise" mostrou-se mais poderoso que os procedimentos baseados em testes múltiplos, utilizando tanto o critério FDR, quanto o de Bonferroni, que não diferiram entre si quanto ao poder de detecção do QTL. Essa superioridade, em poder, do procedimento "stepwise" é tanto maior quanto menor o número de indivíduos avaliados, uma vez que, nos experimentos com 50 indivíduos o poder médio observado foi de 0,33 para o "stepwise", 0,12 para o FDR e 0,14 para o Bonferroni, mostrando uma superioridade, em poder, de 0,20 do primeiro procedimento em relação à média dos demais. Nos experimentos com 100 indivíduos, a superioridade do procedimento "stepwise" sobre os demais foi de 0,19, visto que o poder de detecção de QTLs foi de 0,58 para o "stepwise", 0,40 para FDR e 0,38 para o critério de Bonferroni. Essa

superioridade caiu para 0,16 nos experimentos com 200 indivíduos.

Outro fato digno de nota refere-se ao ganho em poder, ao se aumentar o número de indivíduos de 50 para 100, que foi de 0,24 no procedimento "stepwise", 0,15 no caso do critério FDR e 0,10 no Bonferroni. Já, ao passar de 100 para 200 indivíduos, os ganhos não foram tão expressivos, sendo de apenas 0,005 para o procedimento "stepwise", 0,01 para o FDR e 0,02 para o critério de Bonferroni. Verifica-se, nesta situação, ($h^2 = 0,10$), que o procedimento de seleção de variáveis "stepwise" deve ser preferido em relação aos testes múltiplos, utilizando tanto o critério FDR, quanto o de Bonferroni, para a detecção de QTLs, e que o aumento de 100 para 200 indivíduos avaliados não promoverá um ganho significativo no poder de detecção de QTLs, que é baixo (0,58).

Para um caráter com herdabilidade igual a 0,25, as três metodologias apresentaram-se equivalentes, com poder igual a 0,72, no caso de serem avaliados 50 indivíduos, 0,96 com 100 indivíduos e 1,00 com 200 indivíduos. Para um caráter com herdabilidade de 0,50, ou 0,75, não foi observada nenhuma diferença, entre as três metodologias estudadas, quanto ao poder de detecção do QTL, que foi igual a um, mesmo ao se avaliarem apenas 50 indivíduos. Porém o poder dos critérios FDR e de Bonferroni foi maior, para as marcas mais distantes da real posição do QTL. Pois, enquanto para as duas marcas adjacentes à posição do QTL, localizadas, respectivamente, a 10 cM a direita e a 10 cM à esquerda do QTL, o poder do procedimento "stepwise" caiu para próximo de 0,05, esse nível só foi atingido pelos procedimentos FDR e Bonferroni na última marca do cromossomo, situada a 80 cM do QTL, sendo que, para o caráter com herdabilidade igual a 0,75, o nível mínimo atingido pelo critério FDR foi em torno de 0,10 nos experimentos com 50 indivíduos e 0,20 nos com 100 indivíduos e 200 indivíduos. Essa discordância do procedimento "stepwise" é inerente à sua concepção, pois esse procedimento tem por princípio descartar todas marcas que sejam redundantes. Como sabidamente só existe um QTL nesse genoma, uma única marca (a mais próxima do QTL) é suficiente para descrever seu comportamento; em razão disso, é que o poder do procedimento "stepwise" cai bruscamente

para as marcas adjacentes a uma marca a qual o QTL se encontra completamente ligado, o que é uma grande vantagem dessa metodologia em relação aos demais critérios estudados.

Como pode ser verificado na Figura 13, essas mesmas considerações são válidas quando o caráter for controlado por 3 QTLs independentes e de iguais efeitos, com a ressalva de que, neste caso, cada QTL contribui com $\frac{1}{3}$ da variância genética do caráter e, conseqüentemente, com $\frac{1}{3}$ da herdabilidade do mesmo. Deste modo, tem-se que, para um caráter com herdabilidade 0,10, controlado por 3 QTLs independentes, a herdabilidade associada a cada QTL é de 0,033.

Figura 13 - Frequência relativa de hipóteses rejeitadas, pelos procedimentos "stepwise", FDR e Bonferroni em 500 experimentos, em função da herdabilidade do caráter e número de indivíduos avaliados, para um caráter controlado por três QTLs independentes. Q representa a localização do QTL.

Para o caráter com herdabilidade 0,10, verifica-se, na Figura 13, que o poder de detecção é muito baixo, mesmo avaliando-se 200 indivíduos, sendo que,

nesta situação, o maior poder foi de apenas 0,45, alcançado pelo procedimento "stepwise", mostrando que, em mais de metade dos experimentos, não se conseguiu detectar esse QTL. Porém esse poder aumentou consideravelmente com o aumento da herdabilidade, passando para 0,93, quando a herdabilidade foi de 0,25, tendo sido esse valor obtido pelo critério FDR, vindo o procedimento "stepwise" logo em seguida com poder igual 0,81. Apesar de ter apresentado um poder ligeiramente inferior ao do FDR, o procedimento "stepwise" continua a apresentar a grande vantagem de identificar apenas a marca mais próxima ao QTL. Esses resultados mostram a necessidade incontestável da utilização de um número elevado de indivíduos, pelo menos 200.

No caso de um caráter controlado por oito QTLs, de iguais efeitos, ligados dois a dois, em fase de associação, observa-se, na Figura 14, que o procedimento FDR, apresentou-se mais poderoso que os demais, principalmente quando se considerou um caráter de baixa herdabilidade ($h^2 = 0; 10$); porém, não foi capaz de diferenciar os dois QTLs, mesmo quando estes estavam separados por uma distância maior (50 cM), o que não ocorreu com o procedimento "stepwise". Este último, apesar de menos poderoso, foi capaz de distinguir os dois QTLs, independentemente da distância entre os mesmos.

Verifica-se, na Figura 15, que, considerando os mesmos QTL's do caso anterior, porém ligados em fase de repulsão, os critérios FDR e Bonferroni apresentaram poder de detecção muito inferiores ao procedimento "stepwise", sendo que as duas primeiras metodologias só se aproximaram da última, em termos de poder, quando o caráter sob análise apresentou herdabilidade igual a 0,75, e, mesmo assim, utilizando-se 200 indivíduos e estando os QTLs bem distantes entre si, isto é, a uma distância maior ou igual a 40 cM.

No caso dos caracteres de baixa herdabilidade ($h^2 = 0; 10$ e $h^2 = 0; 25$), o poder dos critérios FDR e Bonferroni foi desprezível, mesmo ao se utilizarem 200 indivíduos, sendo terminantemente desaconselhável o uso dessas duas metodologias para detecção de QTLs, nesta situação, o mesmo ocorrendo com o procedimento

Figura 14 - Frequência relativa de hipóteses rejeitadas, pelos procedimentos "stepwise", FDR e Bonferroni em 500 experimentos, em função da herdabilidade do caráter e número de indivíduos avaliados, para um caráter controlado por 8 QTLs de iguais efeitos, ligados dois a dois em fase de associação. Q representa a localização do QTL.

"stepwise", nos experimentos com 50 e 100 indivíduos. Já ao se utilizarem 200 indivíduos, o procedimento "stepwise" apresentou um ganho em poder, que foi de 0,16 para os QTLs mais próximos e 0,31 para os mais distantes, o que poderia indicar uma possibilidade de maiores ganhos, com o aumento do número de indivíduos. Porém, aumentando o número de indivíduos avaliados para 400, o poder do procedimento "stepwise" para os dois QTL's mais próximos entre si foi de 0,15. Verifica-se, assim, que o aumento do número de indivíduos não alterou o poder.

Para o caráter com herdabilidade 0,50, o poder de detecção de QTLs do procedimento "stepwise" foi de 0,41, para os QTL's mais próximos e 0,72 para os mais distantes, quando se avaliaram 200 indivíduos, e bem inferiores a esses valores, ao utilizar-se um número menor de indivíduos. Considerando um caráter com her-

Figura 15 - Frequência relativa de hipóteses rejeitadas, pelos procedimentos "stepwise", FDR e Bonferroni em 500 experimentos, em função da herdabilidade do caráter e número de indivíduos avaliados, para um caráter controlado por 8 QTLs, de iguais efeitos, ligados dois a dois em fase de repulsão. Q representa a localização do QTL.

dabilidade igual a 0,75, o poder deste procedimento atingiu níveis elevados, mesmo se empregando 100 indivíduos, sendo que, com a utilização de 200 indivíduos, esse poder foi de 0,93 para os dois QTL's mais próximos e 1 para os demais.

A literatura indica que os caracteres quantitativos são controlados por um grande número de QTLs (Goldman et al., 1994; Berke & Rocheford, 1995), dos quais, alguns pertencem a um mesmo grupo de ligação. Pode-se, portanto, concluir que, sendo esses caracteres de baixa herdabilidade, é desaconselhável a utilização da técnica de análise de marcas simples, tanto pelo procedimento "stepwise", quanto pelos critérios FDR e de Bonferroni, para a detecção de QTLs. Essa afirmação é fortalecida pelo fato de nas simulações realizadas terem sido considerados QTLs completamente ligados a uma determinada marca, situação que é a mais favorável

possível para o mapeamento, do ponto de vista de distância entre o QTL e a marca estudada. Além, conforme mostrado por Darvasi et al. (1993), a medida que o QTL se localiza mais longe da marca, ocorre uma redução no poder de detecção, redução que é máxima, quando o QTL se localiza exatamente no meio do intervalo entre duas marcas subsequentes.

Uma alternativa para o mapeamento de QTL's, envolvidos no controle de caracteres de baixa herdabilidade, é a utilização de outras metodologias, como o mapeamento por intervalo (Lande & Botstein, 1989). Porém Darvasi et al. (1993) verificaram que o método de mapeamento por intervalo apresentou apenas uma pequena vantagem, em termos de poder de detecção de QTLs, em relação à análise de marcas simples, sendo que a grande vantagem desta, em relação à análise de marcas simples refere-se à precisão na localização do QTL, no intervalo entre duas marcas adjacentes. Portanto, na situação aqui estudada, também não se espera nenhuma vantagem desta metodologia, principalmente, devido ao fato dos QTLs estarem completamente ligados a uma determinada marca. Pois, nesta situação, as duas metodologias são análogas, uma vez que, considerando uma frequência de recombinação nula entre o QTL e uma das marcas, ou seja, fazendo $c_1 = 0$ nos modelos para o mapeamento por intervalo de Lande & Botstein (1989) ou de Haley & Knott (1992), estes se reduzem aos modelos de análise marcas simples, considerando os métodos de estimação de máxima verossimilhança no primeiro caso, e de mínimos quadrados no último. O mesmo sendo válido para o mapeamento por intervalo composto. (Jansen, 1992;1993 e Zeng, 1993;1994).

Verifica-se, então, que os baixos poderes apresentados pelos métodos de mapeamento são decorrentes da baixa herdabilidade dos caracteres sob análise. Definindo a herdabilidade como a razão entre a variância genética e a variância fenotípica entre as unidades de seleção, isto é:

$$h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \frac{\sigma_e^2}{r}} \quad (46)$$

em que:

σ_g^2 é a variância genética;

$\frac{3}{4}a^2$ é a variância ambiental e r é o número de repetições, na qual as unidades de seleção foram avaliadas.

Verifica-se, em 46, que aumentando-se o número de repetições (r), aumenta-se a herdabilidade e, conseqüentemente, o poder de detecção de QTL's. A título de exemplo, considere-se o caráter controlado por um único QTL, com efeito aditivo igual a um. Nesse caso, tem-se que: $\frac{3}{4}a^2 \simeq \frac{1}{2}a^2 = 0,5$. Para obtenção de $h^2 = 0,10$, considerou-se, na simulação, $\frac{3}{4}a^2 = 4,5$. Nesta situação, caso os indivíduos sejam avaliados em experimentos com repetições, e considerando-se como unidade de seleção/mapeamento o valor fenotípico médio de cada indivíduo, os valores da herdabilidade serão: $h^2 = 0,25$, ao se utilizarem três repetições e $h^2 = 0,5$, se utilizarem nove. Nesse caso, pode-se observar, na Figura 12 que o poder de detecção, no caso da avaliação de 200 indivíduos, passará de 0,60, quando não se utilizarem repetições, para 0,98, ao se utilizarem três repetições e um ao se utilizarem nove.

Como alertado anteriormente, os métodos de mapeamento por intervalo constituem-se de um procedimento de testes múltiplos, nos quais, em um dado intervalo entre duas marcas, vários testes de hipóteses são realizados. Nesse caso, das três metodologias avaliadas, apenas os critérios de FDR e de Bonferroni são aplicáveis. Uma observação das Figuras 12 a 15 mostra que, em geral, o critério FDR apresentou-se mais poderoso que o de Bonferroni, devendo, portanto, ser preferido, para o controle do nível de significância conjunto ao se realizar o mapeamento de QTL, utilizando métodos de mapeamento por intervalo.

3.4 Conclusões

- i. O procedimento baseado em regressão múltipla, utilizando o procedimento "step-wise" para seleção de variáveis, foi mais poderoso, em identificar as marcas associadas a QTL's, do que os procedimentos baseados em testes individuais, utilizando tanto o critério FDR, quanto o de Bonferroni para o controle do nível de significância conjunto.

- ii. Dentre os critérios FDR e de Bonferroni, que são aplicáveis às metodologias de mapeamento por intervalo, o primeiro apresentou-se em geral mais poderoso, devendo, portanto, ser preferido para o controle do nível de significância conjunto.
- iii. O poder de detecção de QTL's é função direta da herdabilidade do caráter, podendo ser aumentado, por meio do uso de um maior número de repetições nos experimentos para avaliação fenotípica do caráter sob análise.

4 PARTIÇÃO DA VARIÂNCIA DA INTERAÇÃO GENÓTIPOS X AMBIENTES EM EFEITOS EXPLICADOS PELOS MARCADORES E DESVIOS

Resumo

Outro problema encontrado na análise de QTL's refere-se à abordagem da interação QTL's x ambientes. Neste contexto, apresentou-se uma partição da variância da interação genótipos x ambientes em efeitos explicados pelos marcadores e desvios, a partir da qual obtiveram-se os estimadores da proporção da variância genética (p_m), e da variância da interação genótipos x ambientes (p_{ms}), explicadas pelos marcadores moleculares. Estes estimadores independem de desvios das frequências alélicas dos marcadores em relação às esperadas (1:2:1 em uma geração F_2 , 1:1 em um retrocruzamento, etc.), porém, apresentam uma alta probabilidade de obtenção de estimativas fora do intervalo paramétrico, principalmente para valores elevados destas proporções. Contudo, estas probabilidades podem ser reduzidas com o aumento do número de repetições e/ou de ambientes nos quais as progênies são avaliadas.

Summary

A problem found in the QTL analysis refer to the QTL x environment interaction. We consider this aspect by partitioning the genotype x environment interaction variance in components explained by the molecular markers and deviations. This allowed estimating the proportion of the genetic variance (p_m), and genotype x environment variance (p_{ms}), explained by the markers. These estimators do not depend on the deviations of allelic frequencies of the markers in relation to the expected values (1:2:1 in a F2 generation, 1:1 in a backcross, etc). However, there is a high probability of obtaining estimates out of the parametric range, specially for high values of these proportion. Nevertheless, these probabilities can be reduced by increasing the number of replications and/or environments where the progenies are evaluated.

4.1 Introdução

As modernas técnicas de obtenção de marcadores moleculares permitem aos geneticistas construir mapas genéticos de alta densidade para qualquer espécie vegetal de interesse. Associados a técnicas estatísticas, esses mapas podem ser utilizados para detecção, localização e estimação dos efeitos de locos controladores de caracteres quantitativos (QTLs). Desse modo, os marcadores moleculares ligados a QTLs podem ser utilizados para aumentar a eficiência dos programas de seleção, por meio da técnica denominada seleção assistida por marcadores moleculares (SAM). Lande & Thompson (1990) propuseram a realização da SAM, utilizando um índice que combina as informações fenotípicas e moleculares, quantificadas por meio de um escore molecular, em proporções apropriadas. Neste procedimento, o escore molecular pode ser obtido a partir dos valores preditos por um modelo de regressão múltipla, tendo como variáveis preditoras o número de "alelos" para cada loco marcador. O peso atribuído a esse escore, no índice de seleção, uma função da

proporção da variância genética explicada pelos marcadores moleculares.

O coeficiente de determinação (r^2) do modelo de regressão múltipla, utilizado na descrição genotípica do caráter, pode ser tomado como estimador da proporção da variância genética explicada pelos marcadores moleculares. Porém Bearzoti & Vencovsky (1998) argumentam que, devido ao fato de a esperança matemática da soma de quadrados de regressão envolver componentes de variâncias genéticas e ambientais, este estimador pode não ser muito adequado e apresentam um outro estimador, baseado na partição da variância genética em efeitos explicados pelos marcadores e desvios.

De acordo com Wang et al. (1999), a interação genótipos x ambientes tem apresentado um importante papel para os melhoristas e geneticistas quantitativos, e que, com a utilização de marcadores moleculares e delineamento experimentais adequados, essa interação pode ser dissecada em seus componentes devido à interação QTLs x ambientes, que são de grande importância para os procedimentos de SAM. Desse modo, propõe-se, no presente trabalho, uma extensão do estimador da proporção variância genética explicada pelos marcadores moleculares apresentado por Bearzoti & Vencovsky (1998), aplicável a experimentos conduzidos em um único ambiente, para o caso de experimentos conduzidos em mais de um ambiente, bem como uma partição da variância da interação genótipos x ambientes em efeitos explicados pelos marcadores moleculares e os seus desvios.

4.2 Metodologia

Apresenta-se uma extensão da partição do componente de variância genética (σ_g^2), nos componentes explicados pelos marcadores moleculares (σ_m^2) e desvios (σ_h^2), ou seja, $\sigma_g^2 = \sigma_m^2 + \sigma_h^2$, apresentada por Bearzoti & Vencovsky (1998), que é aplicável a experimentos conduzidos em um único ambiente, à situação na qual os genótipos são avaliados em mais de um ambiente. Além deste desdobramento de σ_g^2 , é apresentada a correspondente partição do componente de variância da interação

genótipos x ambientes, isto é, $\sigma_{gs}^2 = \sigma_{ms}^2 + \sigma_{hs}^2$.

Para tanto, considerou-se uma partição ortogonal, via regressão linear múltipla, da soma de quadrados de genótipos em soma de quadrados de marcadores e desvios, bem como da soma de quadrados da interação genótipos x ambientes em soma de quadrados da interação marcadores x ambientes e desvios x ambientes, como esquematizado na tabela 12.

Tabela 12. Esquema da análise de variância, com a partição das somas de quadrados de genótipos e da interação em soma de quadrados de marcadores, desvios, interação marcadores x ambientes e desvios x ambientes

FV	GL	SQ	QM
Ambientes	I-1	SQAmb.	QMAmb.
Genótipos	g-1	SQGen.	QMGen
Marcadores	k	SQReg.	QMReg.
Desvios	g-k-1	SQDesv.	QMDesv.
Gen. x Amb.	(g-1)(I-1)	SQGen x Amb.	QMGen. x Amb.
Marc. x Amb.	k(I-1)	SQReg. x Amb.	QMReg. x Amb.
Desv. x Amb.	(g-k-1)(I-1)	SQDesv. x Amb.	QMDesv. x Amb.
Resíduo	lg(r-1)	SQResíduo	QMResíduo
Total	Irg-1		

Os estimadores dos componentes de variância foram obtidos pelo método dos momentos, equacionando-se os quadrados médios, apresentados na tabela 12, às suas respectivas esperanças matemáticas, as quais foram obtidas considerando o modelo aleatório, com erros independentes e distribuição normal.

Pressupondo-se distribuição normal aos valores fenotípicos, e lançando mão dos seguintes teoremas, apresentados por Searle et al. (1992), identificaram-se as distribuições associadas às somas de quadrados, a partir das quais foram obtidas

as estatísticas apropriadas aos testes das hipóteses de nulidade dos componentes de variância σ_m^2 , σ_h^2 , σ_{ms}^2 e σ_{hs}^2 .

Teorema 1 (T1) Se $y \sim N(\mu; V)$, então, uma condição necessária e suficiente para que a forma quadrática $y^0 A y \sim \hat{A}_{(r[A]; \frac{1}{2}\mu^0 A \mu)}^2$ é que AV seja idempotente, em que $r[A]$ é o rank ou posto de A .

Teorema 2 (T2) Sejam $\hat{A} \sim \hat{A}_n^2$ e $\sigma \sim \hat{A}_m^2$ duas variáveis independentes. Nestas condições, a estatística $f = \frac{\hat{A}}{\sigma} \sim \mathcal{F}_{n,m}$

Teorema 3 (T3) Uma condição necessária e suficiente para que as formas quadráticas $y^0 A y$ e $y^0 B y$ sejam independentes é que $AVB = \hat{A}$

Apresentou-se um estimador pontual da proporção da variância genética explicada pelos marcadores moleculares, para o caso de experimentos conduzidos em mais de um ambiente (p_m), passível de ser utilizado na estimação do coeficiente associado ao escore molecular (b_m), no índice de seleção de Lande & Thompson (1990), quando este for obtido, considerando-se o desempenho médio dos genótipos. Neste contexto, apresentou-se um estimador da proporção da variância da interação genótipos x ambientes, explicada pelos marcadores moleculares, de grande utilidade na compreensão da interação genótipos x ambientes.

Em virtude da inviabilidade de expressar as esperanças e variâncias de p_m e p_{ms} explicitamente, derivaram-se expressões aproximadas para esses parâmetros, utilizando-se a expansão em série de Taylor da razão entre duas variáveis aleatórias (X e Y), conforme apresentado por Mood et al. (1974), cujas expressões são:

$$E \left[\frac{X}{Y} \right] = \frac{E[X]}{E[Y]} - \frac{\text{Cov}[X; Y]}{E^2[Y]} + \frac{E[X]}{E^3[Y]} V[Y] \quad (47)$$

$$V \left[\frac{X}{Y} \right] = \frac{E^2[Y] V[X] + E^2[X] V[Y] - 2E[X]E[Y]\text{Cov}[X; Y]}{E^4[Y]} \quad (48)$$

Dado que as variáveis aleatórias em questão são combinações lineares de quadrados médios, suas variâncias foram obtidas por:

$$V[QM] = \frac{2[E[QM]]^2}{v} \quad (\text{Searle et al., 1992}) \quad (49)$$

em que v são os graus de liberdade associados ao quadrado médio

Um estimador não tendencioso, desta variância (49), apresentado por Searle et al. (1992), é:

$$\hat{V}[QM] = \frac{2(QM)^2}{v + 2} \quad (50)$$

4.3 Resultados e discussão

4.3.1 Modelo matemático e partição das somas de quadrados

Considere-se um experimento no qual são avaliadas g progênes em r repetições em l ambientes. Então, um modelo matemático para a análise desses dados é:

$$y_{ij\sim} = \mu + s_{\sim} + t_i + (ts)_{i\sim} + e_{ij\sim} \quad (51)$$

em que

- $y_{ij\sim}$ é o valor observado da progênie i , $i = 1; 2; \dots; g$, na repetição j , $j = 1; 2; \dots; r$ do ambiente $\sim$, $\sim = 1; 2; \dots; l$;
- μ é uma constante inerente a todas as observações;
- $s_{\sim}$ é o efeito do ambiente $\sim$;
- t_i é o efeito da progênie i ;
- $(ts)_{i\sim}$ é o efeito da interação entre a progênie i e o ambiente $\sim$;
- $e_{ij\sim}$ é o erro aleatório associado à observação $y_{ij\sim}$.

Na forma matricial, o modelo 94 é

$$y = X\mu + \epsilon \quad (52)$$

sendo:

y o vetor de valores fenotípicos observados;

X a matriz do delineamento, e

ϵ o vetor de erros aleatórios.

- Somas de quadrados total, de parâmetros e resíduo

Sabe-se que:

$$SQ_{Total} = SQ_{Parâmetros} + SQ_{Resíduo};$$

ou seja:

$$y^0 I y = y^0 P y + y^0 (I - P) y \quad (53)$$

com $P = X X^+ = X (X^0 X)^G X^0$, sendo X^+ a inversa de Moore-Penrose de X , $(X^0 X)^G$ qualquer inversa generalizada de $X^0 X$; e I uma matriz identidade de dimensões $n \times n$, ($n = \text{grl}$).

As somas de quadrados Total, de Parâmetros e Resíduo, estão associadas, respectivamente, a $r[I] = n$, $r[P] = l + g + lg + 1$ e $r[I - P] = lg(r - 1)$ graus de liberdade.

- Soma de quadrados de ambientes

A porção da $SQ_{Parâmetros}$, devida ao efeito de ambientes, denominada Soma de Quadrados de Ambientes, SQL , é dada por:

$$SQL = y^0 (P_L - P_1) y; \quad (54)$$

em que:

$P_L = X_L X_L^+$ e $P_1 = J J^+$, com X_L sendo uma submatriz de X , com dimensões $n \times (l + 1)$ referente aos efeitos de ambientes, e J uma matriz de dimensões $n \times 1$, de uns.

Esta soma de quadrados está associada a $r[P_L - P_1] = l - 1$ graus de liberdade.

- Soma de quadrados de progênes

A porção da SQParâmetros devida ao efeito de progênes, denominada Soma de Quadrados de Progênes, SQP, é dada por:

$$SQP = y^0(P_p - P_1)y; \quad (55)$$

em que:

$P_p = X_p X_p^+$ com X_p sendo uma submatriz de X , com dimensões $n \times (g + 1)$ referente aos efeitos de progênes.

Esta associa-se a $r[P_p - P_1] = g - 1$ graus de liberdade.

- Soma de quadrados da interação

A porção da SQParâmetros devida aos efeitos de interação progênes x ambientes, denominada Soma de Quadrados da Interação, SQPxL é dada por:

$$SQP \times L = y^0(P - P_L - P_p + P_1)y; \quad (56)$$

estando associada a $r[P - P_L - P_p + P_1] = (g - 1)(l - 1)$ graus de liberdade.

- Partição das somas de quadrados de progênes e interação progênes x ambientes em efeitos explicados pelos marcadores e desvios.

Quando se possuem informações de marcadores moleculares para as progênes, o efeito de progênes (t_i) pode ser desdobrado, por meio de regressão linear, em efeitos explicados pelos marcadores e desvios, o mesmo ocorrendo para a interação, do seguinte modo:

Adotando-se o modelo 94, e sendo t_i modelado por:

$$t_i = \sum_{m=1}^k \beta_m Z_{im} + h_i; \quad (57)$$

em que:

β_m é o coeficiente de regressão da marca m ;

Z_{im} é uma variável dummy codificadora do genótipo da i -ésima progênie, quanto à marca t ;

h_i é a porção de g_i não explicada pelas marcas (desvio)

Desse modo, a soma de quadrados de progênies pode ser desdobrada em soma de quadrados de regressão e desvios, sendo:

$$SQ_{Reg} = y'(P_R - P_1)y; \quad (58)$$

com $P_R = ZZ^+$, sendo Z uma matriz de dimensões $n \times k + 1$.

$$SQ_{Desv} = y'(P_p - P_r)y; \quad (59)$$

Estas têm, respectivamente, $r[P_R - P_1] = k$ e $r[P_p - P_r] = g - k - 1$ graus de liberdade.

A soma de quadrados da interação progênies x ambientes também pode ser particionada em soma de quadrados de interação regressão x ambientes e soma de quadrados de interação desvios x ambientes, por meio das seguintes relações:

Sabe-se que:

$$SQ_{P \times L} = \sum_l SQ_{P^l} - SQ_P; \quad (60)$$

em que:

SQ_{P^l} é a soma de quadrados de progênies no ambiente l ;

SQ_P é a soma de quadrados de progênies na análise conjunta, considerando os l ambientes.

Aplicando-se a relação 60 às somas de quadrados de regressão e desvios, tem-se:

$$SQ_{Reg} \times Amb = \sum_i SQ_{Reg,i} - SQ_{Reg}; \quad (61)$$

$$SQ_{Desv} \times Amb = \sum_i SQ_{Desv,i} - SQ_{Desv}; \quad (62)$$

4.3.2 Esperanças matemáticas das somas de quadrados e dos quadrados médios.

Assumindo-se que:

$$s_i \sim N(0; \frac{1}{4} \sigma_s^2);$$

$$t_i \sim N(0; \frac{1}{4} \sigma_g^2);$$

$$(ts)_i \sim N(0; \frac{1}{4} \sigma_{gs}^2);$$

$$e_{ij} \sim N(0; \frac{1}{4} \sigma^2);$$

s_i ; t_i ; $(ts)_i$ e e_{ij} são independentes;

$cov(t_i; t_{i^0}) = 0$; o mesmo sendo válido para s e ts ;

decorre que

$$E(y_{ij}) = 1;$$

$$V(y_{ij}) = \frac{1}{4} \sigma^2 + \frac{1}{4} \sigma_{gs}^2 + \frac{1}{4} \sigma_g^2 + \frac{1}{4} \sigma_s^2;$$

$$cov(y_{ij}; y_{ij^0}) = \frac{1}{4} \sigma_{gs}^2 + \frac{1}{4} \sigma_g^2 + \frac{1}{4} \sigma_s^2;$$

$$cov(y_{ij}; y_{ij^0}) = \frac{1}{4} \sigma_g^2;$$

$$cov(y_{ij}; y_{ij^0}) = \frac{1}{4} \sigma_g^2;$$

$$cov(y_{ij}; y_{i^0j}) = \frac{1}{4} \sigma_s^2;$$

$$cov(y_{ij}; y_{i^0j^0}) = \frac{1}{4} \sigma_s^2;$$

Sabendo-se que a esperança matemática de uma forma quadrática $y^0 A y$ é dada por:

$$E[y^0 A y] = \text{tr}[A V] + E[y]^0 A E[y] \quad \text{Searle (1971);}$$

com V sendo a matriz de variâncias e covariâncias de y . Tem-se que:

$$\begin{aligned} \bullet E[SQL] &= E[y^0 (P_L - P_1) y] = \text{tr}[(P_L - P_1) V] + E[y]^0 (P_L - P_1) E[y] \\ &= \text{tr}[(P_L - P_1) V] + \mu^0 X^0 (P_L - P_1) X \mu \\ &= (l - 1)\frac{3}{4}^2 + (l - 1)(r\frac{3}{4}_{gs}^2 + rg\frac{3}{4}_l^2) \end{aligned}$$

consequentemente,

$$E[QML] = E \left[\frac{1}{l - 1} SQL \right] = \frac{3}{4}^2 + r\frac{3}{4}_{gs}^2 + rg\frac{3}{4}_l^2: \quad (63)$$

$$\begin{aligned} \bullet E[SQP] &= E[y^0 (P_P - P_1) y] = \text{tr}[(P_P - P_1) V] + E[y]^0 (P_P - P_1) E[y] \\ &= \text{tr}[(P_P - P_1) V] + \mu^0 X^0 (P_P - P_1) X \mu \\ &= (g - 1)\frac{3}{4}^2 + (g - 1)(r\frac{3}{4}_{gs}^2 + rl\frac{3}{4}_g^2) \end{aligned}$$

consequentemente,

$$E[QMP] = E \left[\frac{1}{g - 1} SQP \right] = \frac{3}{4}^2 + r\frac{3}{4}_{gs}^2 + rl\frac{3}{4}_g^2: \quad (64)$$

$$\begin{aligned} \bullet E[SQP \times L] &= E[y^0 (P - P_L - P_P + P_1) y] = \text{tr}[(P - P_L - P_P + P_1) V] + \\ &E[y]^0 (P - P_L - P_P + P_1) E[y] \\ &= \text{tr}[(P - P_L - P_P + P_1) V] + \mu^0 X^0 (P - P_L - P_P + P_1) X \mu \\ &= (l - 1)(g - 1)\frac{3}{4}^2 + (l - 1)(g - 1)r\frac{3}{4}_{gs}^2 \end{aligned}$$

consequentemente,

$$E[QMP \times L] = E \left[\frac{1}{(l - 1)(t - 1)} SQP \right] = \frac{3}{4}^2 + r\frac{3}{4}_{gs}^2: \quad (65)$$

$$\begin{aligned} \bullet E[SQResidual] &= E[y^0 (I - P) y] = \text{tr}[(I - P) V] + E[y]^0 (I - P) E[y] \\ &= \text{tr}[(I - P) V] + \mu^0 X^0 (I - P) X \mu \\ &= lg(r - 1)\frac{3}{4}^2 \end{aligned}$$

consequentemente,

$$E[QMResíduo] = E \left[\frac{1}{lg(r-1)} SQResíduo \right] = \frac{\sigma^2}{4}; \quad (66)$$

Para obtenção das esperanças matemáticas dos quadrados médios de regressão, desvios, interação regressão x ambientes e desvios x ambientes assume-se ainda que:

$$\frac{\sigma_g^2}{4} = \frac{\sigma_m^2}{4} + \frac{\sigma_h^2}{4};$$

$$\frac{\sigma_{gs}^2}{4} = \frac{\sigma_{ms}^2}{4} + \frac{\sigma_{hs}^2}{4}$$

$$Cov(y_{ij\cdot}; y_{ij\cdot} = m_i) = \frac{\sigma^2}{4} + \frac{\sigma_h^2}{4} + \frac{\sigma_{hs}^2}{4}$$

$$Cov(y_{ij\cdot}; y_{ij\cdot} = m_i) = \frac{\sigma_h^2}{4} + \frac{\sigma_{hs}^2}{4}$$

$$Cov(y_{ij\cdot}; y_{ij\cdot} = m_i) = \frac{\sigma_h^2}{4}$$

Deste modo, a matriz de variâncias e covariâncias dados os genótipos das progênies, quanto aos marcadores moleculares é:

$$V^{\alpha} = \begin{bmatrix} \frac{\sigma^2}{4} + \frac{\sigma_h^2}{4} + \frac{\sigma_{hs}^2}{4} & \frac{\sigma_h^2}{4} + \frac{\sigma_{hs}^2}{4} & \cdots & 0 & 0 & 0 & 0 \\ \frac{\sigma_h^2}{4} + \frac{\sigma_{hs}^2}{4} & \frac{\sigma^2}{4} + \frac{\sigma_h^2}{4} + \frac{\sigma_{hs}^2}{4} & \cdots & 0 & 0 & 0 & 0 \\ 0 & 0 & \cdots & \frac{\sigma_h^2}{4} & \frac{\sigma_h^2}{4} & 0 & 0 \\ 0 & 0 & \cdots & \frac{\sigma_h^2}{4} & \frac{\sigma_h^2}{4} & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 & \frac{\sigma_h^2}{4} & \frac{\sigma_h^2}{4} \\ 0 & 0 & \cdots & 0 & 0 & \frac{\sigma_h^2}{4} & \frac{\sigma_h^2}{4} \\ \frac{\sigma_h^2}{4} & \frac{\sigma_h^2}{4} & \cdots & 0 & 0 & 0 & 0 \\ \frac{\sigma_h^2}{4} & \frac{\sigma_h^2}{4} & \cdots & 0 & 0 & 0 & 0 \\ 0 & 0 & \cdots & \frac{\sigma^2}{4} + \frac{\sigma_h^2}{4} + \frac{\sigma_{hs}^2}{4} & \frac{\sigma_h^2}{4} + \frac{\sigma_{hs}^2}{4} & 0 & 0 \\ 0 & 0 & \cdots & \frac{\sigma_h^2}{4} + \frac{\sigma_{hs}^2}{4} & \frac{\sigma^2}{4} + \frac{\sigma_h^2}{4} + \frac{\sigma_{hs}^2}{4} & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 & \frac{\sigma^2}{4} + \frac{\sigma_h^2}{4} + \frac{\sigma_{hs}^2}{4} & \frac{\sigma_h^2}{4} + \frac{\sigma_{hs}^2}{4} \\ 0 & 0 & \cdots & 0 & 0 & \frac{\sigma_h^2}{4} + \frac{\sigma_{hs}^2}{4} & \frac{\sigma^2}{4} + \frac{\sigma_h^2}{4} + \frac{\sigma_{hs}^2}{4} \end{bmatrix}$$

E as expressões das esperanças matemáticas dos quadrados médios, de regressão, desvios, interação regressão x ambientes e interação desvios x ambientes são:

$$\begin{aligned}
\bullet E[\text{SQDesv:}] &= E[y^0(P_P - P_R)y] = \text{tr}[(P_P - P_R)V^{\text{u}}] + E[y]^0(P_P - P_R)E[y] \\
&= \text{tr}[(P_P - P_R)V^{\text{u}}] + {}^{-0}Z^0(P_P - P_R)Z^{-} \\
&= (g - k - 1)(\mathbb{A}^2 + r\mathbb{A}_{\text{hs}}^2 + r\text{I}\mathbb{A}_{\text{h}}^2)
\end{aligned}$$

consequentemente,

$$E[\text{QMDesv}] = E\left[\frac{1}{g - k - 1}\text{SQDesv:}\right] = \mathbb{A}^2 + r\mathbb{A}_{\text{hs}}^2 + r\text{I}\mathbb{A}_{\text{h}}^2$$

$$\begin{aligned}
\bullet E[\text{SQReg:}] &= E[y^0(P_R - P_1)y] = \text{tr}[(P_R - P_1)V^{\text{u}}] + E[y]^0(P_R - P_1)E[y] \\
&= \text{tr}[(P_R - P_1)V^{\text{u}}] + {}^{-0}Z^0(P_R - P_1)Z^{-} \\
&= k(\mathbb{A}^2 + r\mathbb{A}_{\text{hs}}^2 + r\text{I}\mathbb{A}_{\text{h}}^2) + {}^{-0}Z^0(P_R - P_1)Z^{-}
\end{aligned}$$

Mas SQP = SQReg. + SQDesv.

$$y^0(P_P - P_1)y = y^0(P_R - P_1)y + y^0(P_P - P_R)y$$

$$E[y^0(P_P - P_1)y] = E[y^0(P_R - P_1)y] + E[y^0(P_P - P_R)y]$$

$$(g - 1)\mathbb{A}^2 + (g - 1)(r\mathbb{A}_{\text{gs}}^2 + r\text{I}\mathbb{A}_{\text{g}}^2) = k(\mathbb{A}^2 + r\mathbb{A}_{\text{hs}}^2 + r\text{I}\mathbb{A}_{\text{h}}^2) + {}^{-0}Z^0(P_P - P_1)Z^{-} + (g - k - 1)(\mathbb{A}^2 + r\mathbb{A}_{\text{hs}}^2 + r\text{I}\mathbb{A}_{\text{h}}^2)$$

$$(g - 1)(r\mathbb{A}_{\text{gs}}^2 + r\text{I}\mathbb{A}_{\text{g}}^2) = {}^{-0}Z^0(P_P - P_1)Z^{-} + (g - 1)(r\mathbb{A}_{\text{hs}}^2 + r\text{I}\mathbb{A}_{\text{h}}^2)$$

$$\text{Lembrando que: } \mathbb{A}_{\text{g}}^2 = \mathbb{A}_{\text{h}}^2 + \mathbb{A}_{\text{m}}^2, \mathbb{A}_{\text{gs}}^2 = \mathbb{A}_{\text{hs}}^2 + \mathbb{A}_{\text{ms}}^2$$

$$(g - 1)[r(\mathbb{A}_{\text{hs}}^2 + \mathbb{A}_{\text{ms}}^2) + r\text{I}(\mathbb{A}_{\text{h}}^2 + \mathbb{A}_{\text{m}}^2)] = {}^{-0}Z^0(P_P - P_1)Z^{-} + (g - 1)(r\mathbb{A}_{\text{hs}}^2 + r\text{I}\mathbb{A}_{\text{h}}^2)$$

$$(g - 1)(r\mathbb{A}_{\text{ms}}^2 + r\text{I}\mathbb{A}_{\text{m}}^2) = {}^{-0}Z^0(P_P - P_1)Z^{-}$$

portanto,

$$E[\text{SQReg:}] = k(\mathbb{A}^2 + r\mathbb{A}_{\text{hs}}^2 + r\text{I}\mathbb{A}_{\text{h}}^2) + {}^{-0}Z^0(P_P - P_1)Z^{-}$$

$$= k(\mathbb{A}^2 + r\mathbb{A}_{\text{hs}}^2 + r\text{I}\mathbb{A}_{\text{h}}^2) + (g - 1)(r\mathbb{A}_{\text{ms}}^2 + r\text{I}\mathbb{A}_{\text{m}}^2)$$

consequentemente,

$$E[\text{QMReg}] = E \left[\frac{1}{k} \text{SQReg:} \right] = \frac{1}{k} \left(\frac{1}{k} + r \frac{1}{k} + r l \frac{1}{k} + \frac{(g-1)}{k} (r \frac{1}{k} + r l \frac{1}{k}) \right) \quad (67)$$

$$\begin{aligned} & \bullet E[\text{SQReg:} \times \text{Amb:}] = E[\text{SQReg:}_1] + E[\text{SQReg:}_2] + \dots + E[\text{SQReg:}_l] - \\ & E[\text{SQReg:}_{\text{conj}}] = \\ & \sum_{s=1}^l [k(\frac{1}{k} + r \frac{1}{k}) + (g-1)r \frac{1}{k}] - [k(\frac{1}{k} + r \frac{1}{k} + r l \frac{1}{k}) + (g-1)(r \frac{1}{k} + \\ & r l \frac{1}{k})] = \\ & k \sum_{s=1}^l \frac{1}{k} + r k \sum_{s=1}^l \frac{1}{k} + (g-1)r \sum_{s=1}^l \frac{1}{k} - k(\frac{1}{k} + r \frac{1}{k} + r l \frac{1}{k}) - \\ & (g-1)(r \frac{1}{k} + r l \frac{1}{k}) = \\ & \text{mas } \sum_{s=1}^l \frac{1}{k} = l \frac{1}{k}; \sum_{s=1}^l \frac{1}{k} = l(\frac{1}{k} + \frac{1}{k}); \sum_{s=1}^l \frac{1}{k} = l(\frac{1}{k} + \frac{1}{k}) \\ & k l \frac{1}{k} + r k l (\frac{1}{k} + \frac{1}{k}) + (g-1) r l (\frac{1}{k} + \frac{1}{k}) - k(\frac{1}{k} + r \frac{1}{k} + r l \frac{1}{k}) - (g-1) \\ & (r \frac{1}{k} + r l \frac{1}{k}) = \\ & k(l-1) \frac{1}{k} + r k(l-1) \frac{1}{k} + (g-1)(l-1) r \frac{1}{k} \end{aligned}$$

consequentemente,

$$E[\text{QMReg:} \times \text{Amb:}] = E \left[\frac{1}{k(l-1)} \text{SQReg:} \times \text{Amb:} \right] = \frac{1}{k(l-1)} \left(\frac{1}{k} + r \frac{1}{k} + \frac{(g-1)}{k} r \frac{1}{k} \right) \quad (68)$$

$$\begin{aligned} & \bullet E[\text{SQDesv:} \times \text{Amb:}] = E[\text{SQDesv:}_1] + E[\text{SQDesv:}_2] + \dots + \\ & E[\text{SQDesv:}_l] - E[\text{SQDesv:}_{\text{conj}}] = \\ & \sum_{s=1}^l (g-k-1)(\frac{1}{k} + r \frac{1}{k}) - (g-k-1)(\frac{1}{k} + r \frac{1}{k} + r l \frac{1}{k}) = \\ & (g-k-1) \sum_{s=1}^l \frac{1}{k} + r(g-k-1) \sum_{s=1}^l \frac{1}{k} - (g-k-1)(\frac{1}{k} + r \frac{1}{k} + r l \frac{1}{k}) = \\ & (g-k-1) l \frac{1}{k} + r l (g-k-1) (\frac{1}{k} + \frac{1}{k}) - (g-k-1)(\frac{1}{k} + r \frac{1}{k} + r l \frac{1}{k}) = \\ & (g-k-1)(l-1) \frac{1}{k} + r l (g-k-1) \frac{1}{k} - (g-k-1) r \frac{1}{k} = \\ & (g-k-1)(l-1) \frac{1}{k} + r(g-k-1)(l-1) \frac{1}{k} \end{aligned}$$

consequentemente,

$$E[\text{QMDesv:} \times \text{Amb:}] = E \left[\frac{1}{(g-k-1)(l-1)} \text{SQDesv:} \times \text{Amb:} \right] = \frac{1}{(g-k-1)(l-1)} \left(\frac{1}{k} + r \frac{1}{k} \right) \quad (69)$$

Assim, pode-se esquematizar o quadro de análise de variância (ANAVA) apresentado na tabela 13.

Tabela 13. Esquema da análise de variância conjunta com o desdobramento da soma de quadrados de tratamentos e interação, em efeitos explicados pelos marcadores moleculares e desvios, com as respectivas esperanças matemáticas dos quadrados médios.

FV	GL	SQ	E(QM)
Ambientes	$l-1$	$y^0(P_L - P_1)y$	$\frac{1}{4}^2 + r\frac{1}{4}_{gs}^2 + rg\frac{1}{4}_l^2$
Progenies	$g-1$	$y^0(P_P - P_1)y$	$\frac{1}{4}^2 + r\frac{1}{4}_{gs}^2 + rl\frac{1}{4}_g^2$
Reg.	k	$y^0(P_R - P_1)y$	$\frac{1}{4}^2 + r\frac{1}{4}_{hs}^2 + rl\frac{1}{4}_h^2 + \frac{(g_l-1)r}{k}(\frac{1}{4}_{ms}^2 + l\frac{1}{4}_m^2)$
Desv.	$g-k-1$	$y^0(P_P - P_R)y$	$\frac{1}{4}^2 + r\frac{1}{4}_{hs}^2 + rl\frac{1}{4}_h^2$
Prog. x Amb.	$(g-1)(l-1)$	$y^0(P - P_L - P_P + P_1)y$	$\frac{1}{4}^2 + r\frac{1}{4}_{gs}^2$
Reg. x Amb.	$k(l-1)$	$\sum_{s=1}^l SQR_s - y^0(P_R - P_1)y$	$\frac{1}{4}^2 + r\frac{1}{4}_{hs}^2 + \frac{(g_l-1)r}{k}r\frac{1}{4}_{ms}^2$
Desv. x Amb.	$(g-k-1)(l-1)$	$\sum_{s=1}^l SQDR_s - y^0(P_P - P_R)y$	$\frac{1}{4}^2 + r\frac{1}{4}_{hs}^2$
Resíduo	$lg(r-1)$	$y^0(I - P)y$	$\frac{1}{4}^2$
Total	$lrg-1$		

$P_R = ZZ^+$; SQR_s = Soma de quadrados de regressão no ambiente s ; $SQDR_s$ = Soma de quadrados de desvios no ambiente s ;

Vale a pena ressaltar que essas expressões das esperanças matemáticas dos quadrados médios independem das proporções genótípicas da população segregante utilizada, quanto ao loco marcador, bem como de desvios dessas proporções em relação às proporções esperadas (1:2:1 em uma geração F_2 , 1:1 em RC, etc.).

4.3.3 Estimadores dos componentes de variância

Com base nas expressões das esperanças matemáticas dos quadrados médios, apresentadas na tabela 13, obtiveram-se os estimadores das componentes de variância residual ($\frac{1}{4}^2$), genética ($\frac{1}{4}_g^2$), de ambientes ($\frac{1}{4}_s^2$), interação progenies x ambientes ($\frac{1}{4}_{gs}^2$), devida ao efeito de interação desvios x ambientes ($\frac{1}{4}_{hs}^2$), interação

marcas x ambientes (χ^2_{ms}), desvios (χ^2_h) e marcas (χ^2_m), dados por:

$$\chi^2_{hs} = \frac{QMResíduo}{r}; \quad (70)$$

$$\chi^2_{ms} = \frac{k(QMReg: \times Amb: - QMDesv: \times Amb:)}{r(g-1)}; \quad (71)$$

$$\chi^2_h = \frac{QMDesv: - QMDesv: \times Amb:}{rl}; \quad (72)$$

$$\chi^2_m = \frac{k[(QMReg: - QMDesv:) - (QMReg: \times Amb: - QMDesv: \times Amb:)]}{rl(g-1)}; \quad (73)$$

$$\chi^2_{gs} = \frac{QMProg: \times Amb: - QMResíduo}{r};$$

$$\chi^2_g = \frac{QMProg: - QMProg: \times Amb:}{rl};$$

$$\chi^2_s = \frac{QMAmb: - QMProg: \times Amb:}{rg};$$

4.3.4 Testes de hipóteses.

A maneira usual de realizar um teste de hipótese, dispondo de um quadro de análise de variância, é mediante uma estatística f , obtida pela razão entre dois quadrados médios, a qual, sob determinadas condições, está associada a uma determinada distribuição de probabilidade $\mathcal{F}$. Assim, conhecendo as esperanças matemáticas dos quadrados médios, pode-se facilmente identificar quais deles deverão compor a referida estatística f . Para obtenção das estatísticas f , apropriadas aos testes de hipóteses, sobre os parâmetros χ^2_m , χ^2_h , χ^2_{ms} e χ^2_{hs} , é necessário conhecer as distribuições das somas de quadrados de regressão, desvios, interação regressão x ambientes, interação desvios x ambientes e dos resíduos, bem como a demonstração da independência entre estas.

Distribuições e independência das formas quadráticas

Assumindo que $y \sim N(\mu; V)$, com $\mu = [1; 1; \dots; 1]^0$ e V já definido, pode-se obter as distribuições associadas às somas de quadrados de regressão, desvios

interação regressão x ambientes, interação desvios x ambientes e resíduo, dadas por:

- Soma de quadrados de regressão.

Dado que:

$$SQReg = y^0(P_R - P_1)y;$$

$$\frac{1}{\frac{3}{4}^2 + r\frac{3}{4}_{hs}^2 + r l \frac{3}{4}_h^2 + \frac{g_i - 1}{k} (r\frac{3}{4}_{ms}^2 + r l \frac{3}{4}_m^2)} (P_R - P_1)V \quad \text{é idempotente;}$$

$$\mu^0(P_R - P_1)\mu = \hat{A} \text{ e}$$

$$r[(P_R - P_1)] = k.$$

Portanto, pelo teorema T1,

$$\frac{SQReg}{\frac{3}{4}^2 + r\frac{3}{4}_{hs}^2 + r l \frac{3}{4}_h^2 + \frac{g_i - 1}{k} (r\frac{3}{4}_{ms}^2 + r l \frac{3}{4}_m^2)} \sim \hat{A}_{(k)}^2, \text{ central.}$$

- Soma de quadrados de desvios.

Dado que:

$$SQDesv = y^0(P_P - P_R)y;$$

$$\frac{1}{\frac{3}{4}^2 + r\frac{3}{4}_{hs}^2 + r l \frac{3}{4}_h^2} (P_P - P_R)V \quad \text{é idempotente;}$$

$$\mu^0(P_P - P_R)\mu = \hat{A} \text{ e}$$

$$r[(P_P - P_R)] = g - k - 1.$$

Então, pelo teorema T1,

$$\frac{SQDesv}{\frac{3}{4}^2 + r\frac{3}{4}_{hs}^2 + r l \frac{3}{4}_h^2} \sim \hat{A}_{(g-k-1)}^2, \text{ central.}$$

- Soma de quadrados da interação regressão x ambientes.

Dado que:

$$SQReg. \text{ x Amb.} = \sum_{i=1}^I SQR_i = y^0(P_R - P_1)y$$

$$= \sum_{i=1}^I y^0(P_{R(i)} - P_{1(i)})y = y^0(P_R - P_1)y$$

$$= y^0[(P_{R(1)} - P_{1(1)}) \oplus \cdots \oplus (P_{R(I)} - P_{1(I)})]y = y^0(P_R - P_1)y$$

$$= y^0\{[(P_{R(1)} - P_{1(1)}) \oplus \cdots \oplus (P_{R(I)} - P_{1(I)})] - (P_R - P_1)\}y,$$

$$\frac{1}{\frac{3}{4}^2 + r\frac{3}{4}_{hs}^2 + \frac{g_i - 1}{k} r\frac{3}{4}_{ms}^2} \{[(P_{R(1)} - P_{1(1)}) \oplus \cdots \oplus (P_{R(I)} - P_{1(I)})] - (P_R - P_1)\}V$$

é idempotente;

$\mu^0\{[(P_{R(1)} - P_{1(1)}) \oplus \cdots \oplus (P_{R(l)} - P_{1(l)})](P_R - P_1)\}\mu = \hat{A}$ e
 $r\{[(P_{R(1)} - P_{1(1)}) \oplus \cdots \oplus (P_{R(l)} - P_{1(l)})](P_R - P_1)\} = k(l-1)$, com $y_{\cdot}$,
 sendo o vetor de valores observados no ambiente $\cdot$, e $(P_{R(\cdot)} - P_{1(\cdot)})$, a matriz núcleo
 da forma quadrática referente à soma de quadrados de regressão, no ambiente $\cdot$;

Assim, pelo teorema T1,

$$\frac{SQReg: \times Amb:}{\frac{3}{4}^2 + r\frac{3}{4}_{hs}^2 + \frac{g_i-1}{k}r\frac{3}{4}_{ms}^2} \sim \hat{A}_{k(l_i-1)}^2, \text{ central:}$$

• Soma de quadrados da interação desvios x ambientes.

Dado que:

$$\begin{aligned}
 SQDesv. \times Amb. &= \sum_{\cdot=1}^l SQDesv. - y^0(P_P - P_R)y \\
 &= \sum_{\cdot=1}^l y^0(P_{P(\cdot)} - P_{R(\cdot)})y_{\cdot} - y^0(P_P - P_R)y \\
 &= y^0[(P_{P(1)} - P_{R(1)}) \oplus \cdots \oplus (P_{P(l)} - P_{R(l)})]y - y^0(P_P - P_R)y \\
 &= y^0\{[(P_{P(1)} - P_{R(1)}) \oplus \cdots \oplus (P_{P(l)} - P_{R(l)})] - (P_P - P_R)\}y; \\
 &\frac{1}{\frac{3}{4}^2 + r\frac{3}{4}_{hs}^2} \{[(P_{P(1)} - P_{R(1)}) \oplus \cdots \oplus (P_{P(l)} - P_{R(l)})] - (P_P - P_R)\}V \text{ é idem-} \\
 &\text{potente;}
 \end{aligned}$$

$\mu^0\{[(P_{P(1)} - P_{R(1)}) \oplus \cdots \oplus (P_{P(l)} - P_{R(l)})] - (P_P - P_R)\}\mu = \hat{A};$
 $r\{[(P_{P(1)} - P_{R(1)}) \oplus \cdots \oplus (P_{P(l)} - P_{R(l)})] - (P_P - P_R)\} = (g - k -$
 $1)(l - 1)$, com $y_{\cdot}$, sendo o vetor de valores observados no ambiente $\cdot$ e $(P_{P(\cdot)} - P_{R(\cdot)})$
 a matriz núcleo da forma quadrática referente à soma de quadrados de desvios, no
 ambiente $\cdot$;

Pelo teorema T1, decorre que:

$$\frac{SQDesv: \times Amb:}{\frac{3}{4}^2 + r\frac{3}{4}_{hs}^2} \sim \hat{A}_{(g-k-1)(l_i-1)}^2, \text{ central.}$$

• Soma de quadrados dos resíduos.

Dado que:

$$\begin{aligned}
 SQRes &= y^0(I - P)y; \\
 &\frac{1}{\frac{3}{4}^2} (I - P)V \text{ é idempotente;}
 \end{aligned}$$

$$\mu^0(I - P)\mu = \hat{A}; \text{ e}$$

$$r[(I - P)] = gl(r - 1).$$

Tem-se, pelo teorema T1, que:

$$\frac{SQRes\text{íduo}}{\frac{3}{4}^2} \sim \hat{A}_{gl(r-1)}^2, \text{ central.}$$

De acordo com o teorema T3, uma condição necessária e suficiente para que as formas quadráticas y^0Ay e y^0By sejam independentes é que $AVB = \hat{A}$. No presente caso, as matrizes A e B são projetores ortogonais, e, portanto, simétricas e idempotentes; V é a matriz de variâncias-covariâncias, que é simétrica e positiva definida. Desse modo, para demonstrar a independência de duas somas de quadrados, basta demonstrar que $AB = \hat{A}$. Então, para as somas de quadrados de interesse tem-se:

- SQDesv. $\times$ Amb. e SQResíduo são independentes

$$SQRes\text{íduo} = y^0(I - P)y;$$

$$SQDesv. \times Amb. = y^0\{[(P_{P(1)} - P_{R(1)}) \oplus \cdots \oplus (P_{P(l)} - P_{R(l)})] - (P_P - P_R)\}y;$$

$$\begin{aligned} & \{[(P_{P(1)} - P_{R(1)}) \oplus \cdots \oplus (P_{P(l)} - P_{R(l)})] - (P_P - P_R)\}(I - P) = \\ & = [(P_{P(1)} - P_{R(1)}) \oplus \cdots \oplus (P_{P(l)} - P_{R(l)})](I - P) - (P_P - P_R)(I - P) \end{aligned}$$

$$\begin{aligned} (P_P - P_R)(I - P) &= (P_P I - P_P P - P_R I + P_R P) \\ &= (P_P - P_P - P_R + P_R) \\ &= \hat{A} \end{aligned}$$

$$\begin{aligned} & [(P_{P(1)} - P_{R(1)}) \oplus \cdots \oplus (P_{P(l)} - P_{R(l)})](I - P) = \\ & = [(P_{P(1)} \oplus \cdots \oplus P_{P(l)}) - (P_{R(1)} \oplus \cdots \oplus P_{R(l)})](I - P) \\ & = [P_P - (P_{R(1)} \oplus \cdots \oplus P_{R(l)})](I - P) \\ & = [P_P - (P_{R(1)} \oplus \cdots \oplus P_{R(l)})] - [P_P P - (P_{R(1)} \oplus \cdots \oplus P_{R(l)})P] \\ & = [P_P - (P_{R(1)} \oplus \cdots \oplus P_{R(l)})] - [P_P - (P_{R(1)} \oplus \cdots \oplus P_{R(l)})] \\ & = \hat{A} \end{aligned}$$

- SQReg. e SQDesv. são independentes

$$\text{SQReg} = y^0(P_R - P_1)y$$

$$\text{SQDesv} = y^0(P_P - P_R)y;$$

$$\begin{aligned}(P_R - P_1)(P_P - P_R) &= \\ &= P_R P_P - P_R P_R - P_1 P_P + P_1 P_R \\ &= P_R - P_R - P_1 + P_1 = \hat{A}\end{aligned}$$

- SQReg. x Amb. e SQDesv. x Amb. sao independentes

$$\text{SQReg: } \times \text{ Amb: } = y^0\{[(P_{R(1)} - P_{1(1)}) \oplus \cdots \oplus (P_{R(l)} - P_{1(l)})] - (P_R - P_1)\}y$$

$$\text{SQDesv: } \times \text{ Amb: } = y^0\{[(P_{P(1)} - P_{R(1)}) \oplus \cdots \oplus (P_{P(l)} - P_{R(l)})] - (P_P - P_R)\}y;$$

$$\begin{aligned}&\{[(P_{R(1)} - P_{1(1)}) \oplus \cdots \oplus (P_{R(l)} - P_{1(l)})] - (P_R - P_1)\}\{[(P_{P(1)} - P_{R(1)}) \oplus \cdots \oplus (P_{P(l)} - P_{R(l)})] - (P_P - P_R)\} = \\ &= [(P_{R(1)} - P_{1(1)}) \oplus \cdots \oplus (P_{R(l)} - P_{1(l)})][(P_{P(1)} - P_{R(1)}) \oplus \cdots \oplus (P_{P(l)} - P_{R(l)})] - [(P_{P(1)} - P_{R(1)}) \oplus \cdots \oplus (P_{P(l)} - P_{R(l)})](P_R - P_1) - [(P_{R(1)} - P_{1(1)}) \oplus \cdots \oplus (P_{R(l)} - P_{1(l)})](P_P - P_R) - (P_R - P_1)(P_P - P_R) \\ &= \hat{A}.\end{aligned}$$

Pois:

$$\begin{aligned}&[(P_{R(1)} - P_{1(1)}) \oplus \cdots \oplus (P_{R(l)} - P_{1(l)})][(P_{P(1)} - P_{R(1)}) \oplus \cdots \oplus (P_{P(l)} - P_{R(l)})] = \\ &= [(P_{R(1)} - P_{1(1)})(P_{P(1)} - P_{R(1)}) \oplus \cdots \oplus (P_{R(l)} - P_{1(l)})(P_{P(l)} - P_{R(l)})] \\ &= [\hat{A} \oplus \cdots \oplus \hat{A}] \\ &= \hat{A};\end{aligned}$$

$$\begin{aligned}&(P_R - P_1)[(P_{P(1)} - P_{R(1)}) \oplus \cdots \oplus (P_{P(l)} - P_{R(l)})] = \\ &= (P_R - P_1)[(P_{P(1)} \oplus \cdots \oplus P_{P(l)}) - (P_{R(1)} \oplus \cdots \oplus P_{R(l)})] \\ &= P_R(P_{P(1)} \oplus \cdots \oplus P_{P(l)}) - P_1(P_{P(1)} \oplus \cdots \oplus P_{P(l)}) - P_R(P_{R(1)} \oplus \cdots \oplus P_{R(l)}) + P_1(P_{R(1)} \oplus \cdots \oplus P_{R(l)}) \\ &= P_R - P_1 - P_R + P_1 \\ &= \hat{A};\end{aligned}$$

$$[(P_{R(1)} - P_{1(1)}) \oplus \cdots \oplus (P_{R(l)} - P_{1(l)})](P_P - P_R) =$$

$$\begin{aligned}
&= [(P_{R(1)} \oplus \cdots \oplus P_{R(l)}) - (P_{1(1)} \oplus \cdots \oplus P_{1(l)})](P_P - P_R) \\
&= (P_{R(1)} \oplus \cdots \oplus P_{R(l)})P_P - (P_{R(1)} \oplus \cdots \oplus P_{R(l)})P_R - \\
&\quad (P_{1(1)} \oplus \cdots \oplus P_{1(l)})P_P + (P_{1(1)} \oplus \cdots \oplus P_{1(l)})P_R \\
&= P_P - P_R - P_P + P_R \\
&= \hat{A}: \\
&(P_R - P_1)(P_P - P_R) = \\
&\quad P_R P_P - P_R P_R - P_1 P_P + P_1 P_R \\
&= P_R - P_R - P_1 + P_1 \\
&= \hat{A}
\end{aligned}$$

• SQDesv. e SQDesv. $\times$ Amb. são independentes

$$SQDesv: = y^0(P_P - P_R)y;$$

$$SQDesv: \times Amb: = y^0\{[(P_{P(1)} - P_{R(1)}) \oplus \cdots \oplus (P_{P(l)} - P_{R(l)})] - (P_P - P_R)\}y;$$

$$\begin{aligned}
&(P_P - P_R)\{[(P_{P(1)} - P_{R(1)}) \oplus \cdots \oplus (P_{P(l)} - P_{R(l)})] - (P_P - P_R)\} = \\
&= (P_P - P_R)[(P_{P(1)} - P_{R(1)}) \oplus \cdots \oplus (P_{P(l)} - P_{R(l)})] - (P_P - P_R)(P_P - \\
&\quad P_R) \\
&= (P_P - P_R)[(P_{P(1)} \oplus \cdots \oplus P_{P(l)}) - (P_{R(1)} \oplus \cdots \oplus -P_{R(l)})] \\
&= P_P(P_{P(1)} \oplus \cdots \oplus P_{P(l)}) - P_P(P_{R(1)} \oplus \cdots \oplus -P_{R(l)}) - \\
&\quad P_R(P_{P(1)} \oplus \cdots \oplus P_{P(l)}) + P_R(P_{R(1)} \oplus \cdots \oplus -P_{R(l)}) - (P_P - P_R) \\
&= P_P - P_R - P_R + P_R - P_P + P_R \\
&= \hat{A}:
\end{aligned}$$

Estadísticas f • Teste para $H_0: \chi_{hs}^2 = 0$ vs $H_a: \chi_{hs}^2 > 0$:

Sob normalidade, tem-se que:

$$\frac{SQDesv: \times Amb:}{\chi_{hs}^2 + r\chi_{hs}^2} \sim \hat{A}_{(g_i - k_i - 1)(l_i - 1)}, \text{ central};$$

$$\frac{SQResiduo}{\chi_{hs}^2} \sim \hat{A}_{gl(r_i - 1)}, \text{ central},$$

SQDesv: $\times$ Amb: e SQResiduo são independentes.

Pelo teorema T2,

$$\begin{aligned} & \frac{\frac{\text{SQDesv: } \mathbb{E} \text{ Amb:}}{\frac{3}{4}^2 + r \frac{3}{4}^2_{\text{hs}}}}{\frac{(g_i - k_i - 1)(l_i - 1)}{\frac{\text{SQResIduo}}{\frac{3}{4}^2}}}}{\frac{1}{gl(r_i - 1)}} \sim \mathcal{F}_{(g_i - k_i - 1)(l_i - 1); gl(r_i - 1)}, \text{ consequentemente,} \\ & \frac{\frac{\text{QMDesv: } \mathbb{E} \text{ Amb:}}{\frac{3}{4}^2 + r \frac{3}{4}^2_{\text{hs}}}}{\frac{\text{QMResIduo}}{\frac{3}{4}^2}} \sim \mathcal{F}_{(g_i - k_i - 1)(l_i - 1); gl(r_i - 1)}, \text{ e,} \\ & \frac{\frac{3}{4}^2 f}{\frac{3}{4}^2 + r \frac{3}{4}^2_{\text{hs}}} \sim \mathcal{F}_{(g_i - k_i - 1)(l_i - 1); gl(r_i - 1)}: \end{aligned} \quad (74)$$

Assim, sob $H_0 : \frac{3}{4}^2_{\text{hs}} = 0$, o lado esquerdo da expressao 74 reduz-se a f , a qual tem uma distribuicao $\mathcal{F}_{(g_i - k_i - 1)(l_i - 1); gl(r_i - 1)}$. E, desse modo, a estatística:

$$f = \frac{\text{QMDesv: } \times \text{ Amb:}}{\text{QMResIduo}}$$

pode ser utilizada para testar a hipótese $H_0 : \frac{3}{4}^2_{\text{hs}} = 0$ vs $H_a : \frac{3}{4}^2_{\text{hs}} > 0$.

- Teste para $H_0 : \frac{3}{4}^2_{\text{ms}} = 0$ vs $H_a : \frac{3}{4}^2_{\text{ms}} > 0$:

Sob normalidade, tem-se que:

$$\begin{aligned} & \frac{\text{SQReg: } \times \text{ Amb:}}{\frac{3}{4}^2 + r \frac{3}{4}^2_{\text{hs}} + \frac{(g_i - 1)}{k} \frac{3}{4}^2_{\text{ms}}} \sim \hat{A}_{k(l_i - 1)}^2, \text{ central;} \\ & \frac{\text{SQDesv: } \times \text{ Amb:}}{\frac{3}{4}^2 + r \frac{3}{4}^2_{\text{hs}}} \sim \hat{A}_{(g_i - k_i - 1)(l_i - 1)}^2, \text{ central;} \\ & \text{SQReg: } \times \text{ Amb: e SQDesv: } \times \text{ Amb: sao independentes.} \end{aligned}$$

Pelo teorema T2, tem-se que

$$\frac{\frac{\text{SQReg: } \bar{E} \text{ Amb:}}{\frac{3/4^2 + r \cdot 3/4_{hs}^2 + \frac{(g_i - 1)}{k} \cdot 3/4_{ms}^2}{k(l_i - 1)}}}{\frac{\text{SQDesv: } \bar{E} \text{ Amb:}}{\frac{3/4^2 + r \cdot 3/4_{hs}^2}{(g_i - k_i - 1)(l_i - 1)}}} \sim \mathcal{F}_{k(l_i - 1); (g_i - k_i - 1)(l_i - 1)}, \text{ consequentemente,}$$

$$\frac{\frac{\text{QMReg: } \bar{E} \text{ Amb:}}{\frac{3/4^2 + r \cdot 3/4_{hs}^2 + \frac{(g_i - 1)}{k} \cdot 3/4_{ms}^2}{k(l_i - 1)}}}{\frac{\text{QMDesv: } \bar{E} \text{ Amb:}}{\frac{3/4^2 + r \cdot 3/4_{hs}^2}{(g_i - k_i - 1)(l_i - 1)}}} \sim \mathcal{F}_{k(l_i - 1); (g_i - k_i - 1)(l_i - 1)}, e,$$

$$\frac{(3/4^2 + r \cdot 3/4_{hs}^2)f}{3/4^2 + r \cdot 3/4_{hs}^2 + \frac{(g_i - 1)}{k} \cdot 3/4_{ms}^2} \sim \mathcal{F}_{k(l_i - 1); (g_i - k_i - 1)(l_i - 1)}: \quad (75)$$

Assim, sob $H_0 : 3/4_{ms}^2 = 0$, o lado esquerdo da expressao 75 reduz-se a f , cuja distribuicao e $\mathcal{F}_{k(l_i - 1); (g_i - k_i - 1)(l_i - 1)}$. E, desse modo, a estatistica:

$$f = \frac{\text{QMReg: } \times \text{ Amb:}}{\text{QMDesv: } \times \text{ Amb:}}$$

pode ser utilizada para testar a hipoteses $H_0 : 3/4_{ms}^2 = 0$ vs $H_a : 3/4_{ms}^2 > 0$.

- Teste para $H_0 : 3/4_n^2 = 0$ vs $H_a : 3/4_n^2 > 0$:

Sob normalidade, tem-se que:

$$\frac{\text{SQDesv:}}{3/4^2 + r \cdot 3/4_{hs}^2 + r \cdot 3/4_n^2} \sim \hat{A}_{(g_i - k_i - 1)}^2, \text{ central;}$$

$$\frac{\text{SQDesv: } \times \text{ Amb:}}{3/4^2 + r \cdot 3/4_{hs}^2} \sim \hat{A}_{(g_i - k_i - 1)(l_i - 1)}^2, \text{ central;}$$

SQDesv: e SQDesv: $\times$ Amb: sao independentes.

Entao, pelo teorema T2,

$$\frac{\frac{\text{SQDesv:}}{\frac{3/4^2 + r \cdot 3/4_{hs}^2 + r \cdot 3/4_n^2}{g_i - k_i - 1}}}{\frac{\text{SQDesv: } \bar{E} \text{ Amb:}}{\frac{3/4^2 + r \cdot 3/4_{hs}^2}{(g_i - k_i - 1)(l_i - 1)}}} \sim \mathcal{F}_{(g_i - k_i - 1); (g_i - k_i - 1)(l_i - 1)}, \text{ consequentemente,}$$

$$\frac{\frac{\text{QMDesv:}}{\frac{3/4^2 + r \cdot 3/4_{hs}^2 + r \cdot 3/4_n^2}{g_i - k_i - 1}}}{\frac{\text{QMDesv: } \bar{E} \text{ Amb:}}{\frac{3/4^2 + r \cdot 3/4_{hs}^2}{(g_i - k_i - 1)(l_i - 1)}}} \sim \mathcal{F}_{(g_i - k_i - 1); (g_i - k_i - 1)(l_i - 1)}, e,$$

$$\frac{(3/4^2 + r \cdot 3/4_{hs}^2)f}{3/4^2 + r \cdot 3/4_{hs}^2 + r \cdot 3/4_n^2} \sim \mathcal{F}_{(g_i - k_i - 1); (g_i - k_i - 1)(l_i - 1)}: \quad (76)$$

Assim, sob $H_0 : \sigma_h^2 = 0$, o lado esquerdo da expressão 76 reduz-se a f , a qual tem uma distribuição $\mathcal{F}_{(g_i - k_i - 1); (g_i - k_i - 1)(l_i - 1)}$. E, desse modo, a estatística:

$$f = \frac{\text{QM Desv:}}{\text{QM Desv:} \times \text{Amb:}}$$

Pode ser utilizada para testar a hipótese $H_0 : \sigma_h^2 = 0$ vs $H_a : \sigma_h^2 > 0$.

- Teste para $H_0 : \sigma_m^2 = 0$ vs $H_a : \sigma_m^2 > 0$:

A estatística f , utilizada para o teste da hipótese de nulidade de um determinado componente de variância, deve ser construída a partir da razão entre dois quadrados médios (Q_1 e Q_2), de modo que apenas Q_1 contenha o componente em questão, e que a razão

$$\frac{E[Q_2]}{E[Q_1]},$$

sob H_0 seja igual a um.

Observando as expressões das esperanças matemáticas dos quadrados médios, apresentadas na tabela 13, nota-se não ser possível obter um teste exato para a hipótese $H_0 : \sigma_m^2 = 0$ vs $H_a : \sigma_m^2 > 0$, sendo necessário, para tanto, lançar mão de combinações de quadrados médios, como mostrado a seguir:

Verifica-se, na tabela 13, que o numerador da estatística f deve ser o quadrado médio de regressão, pois é o único que contém, em sua esperança, o componente σ_m^2 , e, neste caso, o denominador mais apropriado é o quadrado médio de desvios. Assim, tem-se:

$$\frac{E[\text{QM Desv:}]}{E[\text{QM Reg}]} = \frac{\sigma^2 + r\sigma_{hs}^2 + r l \sigma_h^2}{\sigma^2 + r\sigma_{hs}^2 + r l \sigma_h^2 + \frac{(g_i - 1)r}{k} (\sigma_{ms}^2 + \sigma_m^2)}: \quad (77)$$

Para que a expressão 77, sob $H_0 : \sigma_m^2 = 0$ seja igual a um, deve-se utilizar a seguinte combinação de quadrados médios:

$$f = \frac{\text{QM Reg:} + \text{QM Desv:} \times \text{Amb:}}{\text{QM Desv:} + \text{QM Reg:} \times \text{Amb:}} \quad (78)$$

Pois, nesse caso:

$$\frac{E[QMDesv: + QMReg: \times Amb:]}{E[QMReg: + QMDesv: \times Amb:]} = \frac{\frac{3}{4}^2 + r\frac{3}{4}_{hs}^2 + rI\frac{3}{4}_h^2 + \frac{3}{4}^2 + r\frac{3}{4}_{hs}^2 + \frac{(g_i - 1)r}{k}(\frac{3}{4}_{ms}^2 + I\frac{3}{4}_m^2)}{\frac{3}{4}^2 + r\frac{3}{4}_{hs}^2 + rI\frac{3}{4}_h^2 + \frac{(g_i - 1)r}{k}r\frac{3}{4}_{ms}^2 + \frac{3}{4}^2 + r\frac{3}{4}_{hs}^2}, \quad (79)$$

que sob $H_0 : \frac{3}{4}_m^2 = 0$, é igual a um.

Dado que QMReg:, QMDesv: $\times$ Amb:, QMDesv: e QMReg: $\times$ Amb: têm distribuições χ^2 , e são independentes entre si, a estatística 78, sob $H_0 : \frac{3}{4}_m^2 = 0$, tem uma distribuição aproximada $\mathcal{F}$, com v_1 graus de liberdade para o numerador e v_2 graus de liberdade para o denominador, sendo v_1 e v_2 dados pela fórmula de Satterthwaite (1946),

$$v_1 \approx \frac{[QMReg + QMDesv: \times Ambientes]^2}{\frac{[QMReg]^2}{GLReg} + \frac{[QMDesv: \times Ambientes]^2}{GLDesv: \times Ambientes}}$$

$$v_2 \approx \frac{[QMDesv + QMReg: \times Ambientes]^2}{\frac{[QMDesv]^2}{GLDesv} + \frac{[QMReg: \times Ambientes]^2}{GLReg: \times Ambientes}}$$

Esta estatística pode, então, ser utilizada para testar a hipótese $H_0 : \frac{3}{4}_m^2 = 0$ vs $H_a : \frac{3}{4}_m^2 > 0$.

4.3.5 Estimadores das proporções da variância genética e da variância da interação genótipos $\times$ ambientes explicadas pelos marcadores moleculares

Uma extensão natural do estimador 16 apresentado por Bearzoti & Vencovsky (1998) para a situação em que as progênies são avaliadas em vários ambientes é:

$$\hat{p} = \frac{\frac{3}{4}_m^2}{\frac{3}{4}_m^2 + \frac{3}{4}_h^2}; \quad (80)$$

em que o estimador $\mathbb{A}_m^2$ é dado por 84, e o estimador $\mathbb{A}_h^2$, por 83.

Um estimador da proporção da variância da interação genótipos x ambientes, devida à interação marcadores moleculares x ambientes, é dado por:

$$\hat{p}_{ms} = \frac{\mathbb{A}_{ms}^2}{\mathbb{A}_{ms}^2 + \mathbb{A}_{hs}^2}; \quad (81)$$

em que o estimador $\mathbb{A}_{ms}^2$ é dado por 71, e o estimador $\mathbb{A}_{hs}^2$, por 70.

Esperança e Variância de $\hat{p}_m$ e $\hat{p}_{ms}$

Esperança de $\hat{p}_m$

Sendo $\hat{p}_m = \frac{\mathbb{A}_m^2}{\mathbb{A}_m^2 + \mathbb{A}_h^2}$ (80), e fazendo $x = \mathbb{A}_m^2$ e $y = \mathbb{A}_m^2 + \mathbb{A}_h^2$ em 47 decorre que:

$$\begin{aligned} E[\hat{p}_m] &= E\left[\frac{\mathbb{A}_m^2}{\mathbb{A}_m^2 + \mathbb{A}_h^2}\right] = \frac{E[\mathbb{A}_m^2]}{E[\mathbb{A}_m^2 + \mathbb{A}_h^2]} - \frac{\text{Cov}[\mathbb{A}_m^2; \mathbb{A}_m^2 + \mathbb{A}_h^2]}{E^2[\mathbb{A}_m^2 + \mathbb{A}_h^2]} + \frac{E[\mathbb{A}_m^2]}{E^3[\mathbb{A}_m^2 + \mathbb{A}_h^2]} V[\mathbb{A}_m^2 + \mathbb{A}_h^2] \\ &= \frac{\mathbb{A}_m^2}{\mathbb{A}_m^2 + \mathbb{A}_h^2} - \frac{V(\mathbb{A}_m^2) + \text{Cov}[\mathbb{A}_m^2; \mathbb{A}_h^2]}{(\mathbb{A}_m^2 + \mathbb{A}_h^2)^2} + \frac{\mathbb{A}_m^2}{(\mathbb{A}_m^2 + \mathbb{A}_h^2)^3} [V(\mathbb{A}_m^2) + V(\mathbb{A}_h^2) + 2\text{Cov}(\mathbb{A}_m^2; \mathbb{A}_h^2)] \\ &= \frac{\mathbb{A}_m^2}{\mathbb{A}_m^2 + \mathbb{A}_h^2} - \frac{V(\mathbb{A}_m^2) + \text{Cov}[\mathbb{A}_m^2; \mathbb{A}_h^2]}{(\mathbb{A}_m^2 + \mathbb{A}_h^2)^2} + \frac{\mathbb{A}_m^2}{(\mathbb{A}_m^2 + \mathbb{A}_h^2)^3} [V(\mathbb{A}_m^2) + V(\mathbb{A}_h^2) + 2\text{Cov}(\mathbb{A}_m^2; \mathbb{A}_h^2)]; \end{aligned} \quad (82)$$

com,

$$\begin{aligned} V(\mathbb{A}_h^2) &= V\left[\frac{1}{r}(\text{QM Desv.:} - \text{QM Desv.:} \times \text{Amb.:})\right] \\ &= \frac{1}{r^2} [V(\text{QM Desv.:}) + V(\text{QM Desv.:} \times \text{Amb.:}) - 2\text{Cov}(\text{QM Desv.:}; \text{QM Desv.:} \times \text{amb.:})] \\ &= \frac{1}{r^2} \left[\frac{2(\text{QM Desv.:})^2}{g - k - 1} + \frac{2(\text{QM Desv.:} \times \text{Amb.:})^2}{(g - k - 1)(l - 1)} \right] \\ &= \frac{1}{r^2} \left[\frac{2(\mathbb{A}^2 + r\mathbb{A}_{hs}^2 + r\mathbb{A}_h^2)^2}{g - k - 1} + \frac{2(\mathbb{A}^2 + r\mathbb{A}_{hs}^2)^2}{(g - k - 1)(l - 1)} \right]; \end{aligned} \quad (83)$$

$$\begin{aligned}
V(\mathcal{M}_m^2) &= V \left[\frac{k}{r(l(g-1))} [(QMReg: - QMDesv:) - (QMReg: \times Amb: - QMDesv: \times Amb:)] \right] \\
&= \left(\frac{k}{r(l(g-1))} \right)^2 [V(QMReg:) + V(QMDesv:) + V(QMReg: \times Amb:) \\
&\quad + V(QMDesv: \times Amb:) - 2Cov(QMReg:; QMDesv:) \\
&\quad - 2Cov(QMReg:; QMDesv: \times Amb:) + 2Cov(QMReg:; QMReg: \times Amb:) \\
&\quad - 2Cov(QMReg:; QMDesv: \times Amb:) + 2Cov(QMDesv:; QMReg: \times Amb:) \\
&\quad - 2Cov(QMDesv:; QMDesv: \times Amb:) - 2Cov(QMDesv:; QMReg: \times Amb:)] \\
&= \left(\frac{k}{r(g-1)} \right)^2 \left[\frac{2(QMReg:)^2}{k} + \frac{2(QMDesv:)^2}{t-k-1} + \frac{2(QMReg: \times Amb:)^2}{k(l-1)} \right. \\
&\quad \left. + \frac{2(QMDesv: \times Amb:)^2}{(g-k-1)(l-1)} \right] \\
&= \left(\frac{k}{r(g-1)} \right)^2 \left[\frac{2(\mathcal{M}_m^2 + r\mathcal{M}_{hs}^2 + r\mathcal{M}_h^2 + \frac{r(g_i-1)}{k}(\mathcal{M}_{ms}^2 + \mathcal{M}_m^2))^2}{k} + \frac{2(\mathcal{M}_m^2 + r\mathcal{M}_{hs}^2 + r\mathcal{M}_h^2)^2}{g-k-1} \right. \\
&\quad \left. + \frac{2(\mathcal{M}_m^2 + r\mathcal{M}_{hs}^2 + r\mathcal{M}_h^2 + \frac{r(g_i-1)}{k}\mathcal{M}_{ms}^2)^2}{k(l-1)} + \frac{2(\mathcal{M}_m^2 + r\mathcal{M}_{hs}^2)^2}{(g-k-1)(l-1)} \right]; \tag{84}
\end{aligned}$$

$$\begin{aligned}
Cov(\mathcal{M}_m^2; \mathcal{M}_h^2) &= \\
&= Cov \left[\frac{k}{(g-1)(r)} (QMReg: - QMDesv: - QMReg: \times Amb: + QMDesv: \times Amb:); \right. \\
&\quad \left. \frac{1}{r} (QMDesv: - QMDesv: \times Amb:) \right] \\
&= \frac{k}{(r)^2(g-1)} [Cov(QMReg:; QMDesv:) - Cov(QMReg:; QMDesv: \times Amb:) \\
&\quad - Cov(QMDesv:; QMDesv:) + Cov(QMDesv:; QMDesv: \times Amb:) \\
&\quad - Cov(QMReg: \times Amb:; QMDesv:) + Cov(QMReg: \times Amb:; QMDesv: \times Amb:) \\
&\quad + Cov(QMDesv: \times Amb:; QMDesv:) - Cov(QMDesv: \times Amb:; QMDesv: \times Amb:)] \\
&= \frac{k}{(r)^2(g-1)} [V(QMDesv:) - V(QMDesv: \times Amb:)] \\
&= \frac{k}{(r)^2(g-1)} \left[\frac{2(QMDesv:)^2}{g-k-1} - \frac{2(QMDesv: \times Amb:)^2}{(g-k-1)(l-1)} \right] \\
&= \frac{k}{(r)^2(g-1)} \left[\frac{2(\mathcal{M}_m^2 + r\mathcal{M}_{hs}^2 + r\mathcal{M}_h^2)^2}{g-k-1} - \frac{2(\mathcal{M}_m^2 + r\mathcal{M}_{hs}^2)^2}{(g-k-1)(l-1)} \right]; \tag{85}
\end{aligned}$$

Variância de $\hat{\rho}_m$

Fazendo $x = \mathcal{A}_m^2$ e $y = \mathcal{A}_m^2 + \mathcal{A}_h^2$ em 48 tem-se:

$$\begin{aligned} \text{Var}[\hat{\rho}_m] &= \text{Var}\left[\frac{\mathcal{A}_m^2}{\mathcal{A}_m^2 + \mathcal{A}_h^2}\right] = \frac{(\mathcal{A}_m^2 + \mathcal{A}_h^2)^2 V(\mathcal{A}_m^2) + (\mathcal{A}_h^2)^2 V(\mathcal{A}_m^2 + \mathcal{A}_h^2) - 2\mathcal{A}_m^2 (\mathcal{A}_m^2 + \mathcal{A}_h^2) \text{Cov}(\mathcal{A}_m^2; \mathcal{A}_m^2 + \mathcal{A}_h^2)}{(\mathcal{A}_m^2 + \mathcal{A}_h^2)^4} \\ &= \frac{(\mathcal{A}_m^2 + \mathcal{A}_h^2)^2 V(\mathcal{A}_m^2) + \mathcal{A}_m^4 [V(\mathcal{A}_m^2) + V(\mathcal{A}_h^2) + 2\text{Cov}(\mathcal{A}_m^2; \mathcal{A}_h^2)] - 2\mathcal{A}_m^2 (\mathcal{A}_m^2 + \mathcal{A}_h^2) [V(\mathcal{A}_m^2) + \text{Cov}(\mathcal{A}_m^2; \mathcal{A}_h^2)]}{(\mathcal{A}_m^2 + \mathcal{A}_h^2)^4} \\ &= \frac{(\mathcal{A}_m^2 + \mathcal{A}_h^2)^2 V(\mathcal{A}_m^2) + \mathcal{A}_m^4 V(\mathcal{A}_m^2) + \mathcal{A}_m^4 V(\mathcal{A}_h^2) - 2\mathcal{A}_m^2 \mathcal{A}_h^2 V(\mathcal{A}_m^2) - 2\mathcal{A}_m^2 \mathcal{A}_h^2 \text{Cov}(\mathcal{A}_m^2; \mathcal{A}_h^2)}{(\mathcal{A}_m^2 + \mathcal{A}_h^2)^4} \\ &= \frac{(\mathcal{A}_m^4 + 2\mathcal{A}_m^2 \mathcal{A}_h^2 + \mathcal{A}_h^4) V(\mathcal{A}_m^2) + \mathcal{A}_m^4 V(\mathcal{A}_h^2) - 2\mathcal{A}_m^2 \mathcal{A}_h^2 V(\mathcal{A}_m^2) - 2\mathcal{A}_m^2 \mathcal{A}_h^2 \text{Cov}(\mathcal{A}_m^2; \mathcal{A}_h^2)}{(\mathcal{A}_m^2 + \mathcal{A}_h^2)^4} \end{aligned}$$

$$\text{Var}(\hat{\rho}_m) = \text{Var}\left[\frac{\mathcal{A}_m^2}{\mathcal{A}_m^2 + \mathcal{A}_h^2}\right] = \frac{\mathcal{A}_h^4 V(\mathcal{A}_m^2) + \mathcal{A}_m^4 V(\mathcal{A}_h^2) - 2\mathcal{A}_m^2 \mathcal{A}_h^2 \text{Cov}(\mathcal{A}_m^2; \mathcal{A}_h^2)}{(\mathcal{A}_m^2 + \mathcal{A}_h^2)^4}; \quad (86)$$

com,

$V(\mathcal{A}_m^2)$ dado por 84;

$V(\mathcal{A}_h^2)$ dado por 83 e

$\text{Cov}(\mathcal{A}_m^2; \mathcal{A}_h^2)$ dado por 85.

Esperança de $\hat{\rho}_{ms}$

Dado que $\hat{\rho}_{ms} = \frac{\mathcal{A}_{ms}^2}{\mathcal{A}_{ms}^2 + \mathcal{A}_{hs}^2}$ (81), então fazendo $x = \mathcal{A}_{ms}^2$ e $y = \mathcal{A}_{ms}^2 + \mathcal{A}_{hs}^2$ em 47 decorre:

$$\begin{aligned} E[\hat{\rho}_{ms}] &= E\left[\frac{\mathcal{A}_{ms}^2}{\mathcal{A}_{ms}^2 + \mathcal{A}_{hs}^2}\right] = \frac{E[\mathcal{A}_{ms}^2]}{E[\mathcal{A}_{ms}^2 + \mathcal{A}_{hs}^2]} - \frac{\text{Cov}[\mathcal{A}_{ms}^2; \mathcal{A}_{ms}^2 + \mathcal{A}_{hs}^2]}{E^2[\mathcal{A}_{ms}^2 + \mathcal{A}_{hs}^2]} + \frac{E[\mathcal{A}_{ms}^2]}{E^3[\mathcal{A}_{ms}^2 + \mathcal{A}_{hs}^2]} V[\mathcal{A}_{ms}^2 + \mathcal{A}_{hs}^2] \\ &= \frac{\mathcal{A}_{ms}^2}{\mathcal{A}_{ms}^2 + \mathcal{A}_{hs}^2} - \frac{V(\mathcal{A}_{ms}^2) + \text{Cov}[\mathcal{A}_{ms}^2; \mathcal{A}_{hs}^2]}{(\mathcal{A}_{ms}^2 + \mathcal{A}_{hs}^2)^2} + \frac{\mathcal{A}_{ms}^2}{(\mathcal{A}_{ms}^2 + \mathcal{A}_{hs}^2)^3} [V(\mathcal{A}_{ms}^2) + V(\mathcal{A}_{hs}^2) + 2\text{Cov}(\mathcal{A}_{ms}^2; \mathcal{A}_{hs}^2)] \end{aligned}$$

$$E(\hat{\rho}_{ms}) = \frac{\mathcal{A}_{ms}^2}{\mathcal{A}_{ms}^2 + \mathcal{A}_{hs}^2} - \frac{V(\mathcal{A}_{ms}^2) + \text{Cov}[\mathcal{A}_{ms}^2; \mathcal{A}_{hs}^2]}{(\mathcal{A}_{ms}^2 + \mathcal{A}_{hs}^2)^2} + \frac{\mathcal{A}_{ms}^2}{(\mathcal{A}_{ms}^2 + \mathcal{A}_{hs}^2)^3} [V(\mathcal{A}_{ms}^2) + V(\mathcal{A}_{hs}^2) + 2\text{Cov}(\mathcal{A}_{ms}^2; \mathcal{A}_{hs}^2)]; \quad (87)$$

com

$$\begin{aligned}
 V(\mathcal{M}_{ms}^2) &= V \left[\frac{k}{r(g-1)} (\text{QMReg} \times \text{Amb} - \text{QMDesv} \times \text{Amb}) \right] \\
 &= \left(\frac{k}{r(g-1)} \right)^2 [V(\text{QMReg} \times \text{Amb}) + V(\text{QMDesv} \times \text{Amb}) \\
 &\quad - 2\text{Cov}(\text{QMReg} \times \text{Amb}; \text{QMDesv} \times \text{Amb})] \\
 &= \left(\frac{k}{r(g-1)} \right)^2 \left[\frac{2(\text{QMReg} \times \text{Amb})^2}{k(l-1)} + \frac{2(\text{QMDesv} \times \text{Amb})^2}{(g-k-1)(l-1)} \right] \\
 &= \left(\frac{k}{r(g-1)} \right)^2 \left[\frac{2(\mathcal{M}^2 + r\mathcal{M}_{hs}^2 + \frac{r(g-1)}{k}\mathcal{M}_{ms}^2)^2}{k(l-1)} + \frac{2(\mathcal{M}^2 + r\mathcal{M}_{hs}^2)^2}{(g-k-1)(l-1)} \right] : \quad (88)
 \end{aligned}$$

$$\begin{aligned}
 V(\mathcal{M}_{hs}^2) &= V \left[\frac{1}{r} (\text{QMDesv} \times \text{Amb} - \text{QMResíduo}) \right] \\
 &= \frac{1}{r^2} [V(\text{QMDesv} \times \text{Amb}) + V(\text{QMResíduo}) - 2\text{cov}(\text{QMDesv} \times \text{Amb}; \text{QMResíduo})] \\
 &= \frac{1}{r^2} \left[\frac{2(\text{QMDesv} \times \text{Amb})^2}{(g-k-1)(l-1)} + \frac{2(\text{QMResíduo})^2}{l(g-1)(r-1)} \right] \\
 &= \frac{1}{r^2} \left[\frac{2(\mathcal{M}^2 + r\mathcal{M}_{hs}^2)^2}{(g-k-1)(l-1)} + \frac{2(\mathcal{M}^2)^2}{l(g-1)(r-1)} \right] : \quad (89)
 \end{aligned}$$

$$\begin{aligned}
 \text{Cov}(\mathcal{M}_{ms}^2; \mathcal{M}_{hs}^2) &= \\
 &= \text{Cov} \left[\frac{k}{r(g-1)} (\text{QMReg} \times \text{Amb} - \text{QMDesv} \times \text{Amb}); \frac{1}{r} (\text{QMDesv} \times \text{Amb} - \text{QMResíduo}) \right] \\
 &= \frac{k}{r^2(g-1)} [\text{Cov}(\text{QMReg} \times \text{Amb}; \text{QMDesv} \times \text{Amb}) - \text{Cov}(\text{QMReg} \times \text{Amb}; \text{QMResíduo}) \\
 &\quad - \text{Cov}(\text{QMDesv} \times \text{Amb}; \text{QMDesv} \times \text{Amb}) + \text{Cov}(\text{QMDesv} \times \text{Amb}; \text{QMResíduo})] \\
 &= \frac{k}{r^2(g-1)} V(\text{QMDesv} \times \text{Amb}) = \frac{k}{r^2(g-1)} \left[\frac{2(\text{QMDesv} \times \text{Amb})^2}{(g-k-1)(l-1)} \right] \\
 &= \frac{k}{r^2(g-1)} \left[\frac{2(\mathcal{M}^2 + r\mathcal{M}_{hs}^2)^2}{(g-k-1)(l-1)} \right] : \quad (90)
 \end{aligned}$$

Variância de $\hat{\rho}_{ms}$

De 81 e 48 tem-se:

$$\text{Var}[\hat{\rho}_{ms}] = \text{Var} \left[\frac{\mathcal{M}_{ms}^2}{\mathcal{M}_{ms}^2 + \mathcal{M}_{hs}^2} \right] = \frac{\mathcal{M}_{hs}^4 V(\mathcal{M}_{ms}^2) + \mathcal{M}_{ms}^4 V(\mathcal{M}_{hs}^2) - 2\mathcal{M}_{ms}^2 \mathcal{M}_{hs}^2 \text{Cov}(\mathcal{M}_{ms}^2; \mathcal{M}_{hs}^2)}{(\mathcal{M}_{ms}^2 + \mathcal{M}_{hs}^2)^4} \quad (91)$$

com

$V(\chi^2_{ms})$ dada por 88;

$V(\chi^2_{hs})$ dada por 89 e

$Cov(\chi^2_{ms}, \chi^2_{hs})$ dada por 90;

Estas expressões das variâncias de $\hat{p}_m$ e de $\hat{p}_{ms}$ podem ser empregadas na construção de intervalos de confiança, utilizando a aproximação normal. Porém, devido ao fato de não serem expressões triviais, necessitando da utilização de um grande número de outras estimativas (χ^2 , χ^2_m , χ^2_h , χ^2_{ms} e χ^2_{hs}), as quais já estão sujeitas a erros, os valores obtidos por estas expressões podem ser um tanto quanto viesados. Adicionando-se o fato de Bearzoti & Vencovsky terem verificado, via simulação computacional, que a aproximação normal para a distribuição de $\hat{p}$ não é satisfatória, para valores elevados de p ($p = 0;9$), mesmo ao utilizar um grande número de progênies ($g=100$), sugere-se lançar mão de métodos de reamostragem, tais como o "bootstrap", para obtenção dos intervalos de confiança para $\hat{p}_m$ e de $\hat{p}_{ms}$.

Probabilidade de obtenção de estimativas de $\hat{p}_m$ e $\hat{p}_{ms}$ maiores que um

Conforme mostrado por Bearzoti & Vencovsky (1998), para o caso de experimentos conduzidos em um único ambiente, a obtenção de estimativas de p fora do intervalo paramétrico ($0 \leq p \leq 1$) não é tão incomum, principalmente para valores elevados de p ($p > 0;9$). Sendo o estimador $\hat{p}_m$ dado por: $\hat{p}_m = \frac{\chi^2_m}{\chi^2_m + \chi^2_h}$, tem-se que:

$$\begin{aligned}
P(\hat{p}_m > 1) &= P\left(\frac{\mathcal{Y}_m^2}{\mathcal{Y}_m^2 + \mathcal{Y}_h^2} > 1\right) \\
&= P(\mathcal{Y}_h^2 < 0) \\
&= P(\text{QMDesv:} < \text{QMDesv:} \times \text{Amb:}), \text{ de 83} \\
&= P\left(\frac{\text{QMDesv:}}{\text{QMDesv:} \times \text{Amb:}} < 1\right) \\
&= P\left(\frac{\mathcal{Y}_m^2 + r\mathcal{Y}_{hs}^2 + r\mathcal{Y}_h^2}{\mathcal{Y}_m^2 + r\mathcal{Y}_{hs}^2} \mathcal{F}_{(g_i - k_i - 1):(g_i - k_i - 1)(l_i - 1)} < 1\right), \text{ de 76} \\
&= P\left(\mathcal{F}_{(g_i - k_i - 1):(g_i - k_i - 1)(l_i - 1)} < \frac{\mathcal{Y}_m^2 + r\mathcal{Y}_{hs}^2}{\mathcal{Y}_m^2 + r\mathcal{Y}_{hs}^2 + r\mathcal{Y}_h^2}\right) \quad (92)
\end{aligned}$$

Verifica-se na eq. 92 que o aumento do número de repetições por local não altera os parâmetros da distribuição, mas diminui o valor f , abaixo do qual resultam estimativas de p_m maiores que um, e, consequentemente, a probabilidade de ocorrência desse evento. Aumentando-se o número de ambientes, também diminui-se o valor f , porém essa alteração provoca simultaneamente uma alteração na distribuição $\mathcal{F}$, que, mantido o mesmo ponto de corte, aumenta a probabilidade de ocorrência de valores inferiores a este ponto. Os parâmetros g (número de progênes) e k (número de marcas) alteram apenas o formato da distribuição, sendo que, para um mesmo ponto de corte, o aumento de g , mantendo k fixo, diminui a área sob a curva abaixo do ponto de corte. Por outro lado, o aumento de k , mantendo g fixo, aumenta essa área, aumentando assim, a probabilidade de obtenção de estimativas de p_m maiores que um.

Para o estimador $\hat{p}_{ms}$, tem-se:

$$\begin{aligned}
P(\hat{p}_{ms} > 1) &= P\left(\frac{\frac{3}{4}^2_{ms}}{\frac{3}{4}^2_{ms} + \frac{3}{4}^2_{hs}} > 1\right) \\
&= P(\frac{3}{4}^2_{hs} < 0) \\
&= P(QMDesv: \times Amb: < QMResiduo), \text{ de } 70 \\
&= P\left(\frac{QMDesv: \times Amb:}{QMResiduo} < 1\right) \\
&= P\left(\frac{\frac{3}{4}^2 + r\frac{3}{4}^2_{hs}}{\frac{3}{4}^2} \mathcal{F}_{(g_i - k_i - 1)(l_i - 1): gl(r_i - 1)} < 1\right), \text{ de } 76 \\
&= P\left(\mathcal{F}_{(g_i - k_i - 1)(l_i - 1): gl(r_i - 1)} < \frac{\frac{3}{4}^2}{\frac{3}{4}^2 + r\frac{3}{4}^2_{hs}}\right) \quad (93)
\end{aligned}$$

No caso da proporção da interação genótipos x ambientes, devida à interação marcadores x ambientes, tem-se que os parâmetros experimentais número de progênies avaliadas (g) e número de ambientes (l) atuam somente sobre o formato da distribuição, sendo que, para um mesmo ponto de corte, o aumento desses valores diminui a probabilidade de ocorrência de um valor inferior ao ponto de corte. Ou seja, tanto o aumento do número de progênies, quanto o de ambientes nas quais as progênies são avaliadas diminuem a probabilidade de obtenção de estimativas de p_{ms} maiores que um, sendo que a maior redução é obtida aumentando-se o número de ambientes. Assim como no caso de $\hat{p}_m$, o incremento no número de marcas utilizadas aumenta a probabilidade de obtenção de estimativas de p_{ms} maiores que um.

A título de exemplo, encontram-se apresentadas na Tabela 14 as probabilidades de obtenção de estimativas de p_m e p_{ms} maiores que um, em função do número de marcas utilizadas no modelo (k) e dos valores paramétricos de p_m e p_{ms} , considerando um experimento, no qual foram avaliadas 100 progênies em quatro ambientes, com duas repetições por ambiente. Considerou-se que o caráter apresentou uma herdabilidade ao nível de média de progênies igual a 0,40, e que $\frac{3}{4}^2_g = \frac{3}{4}^2_{gs}$. Verifica-se, nesta tabela, que, para valores elevados, tanto de p_m , quanto de p_{ms} , são grandes as probabilidades de obtenção de estimativas desses parâmetros maiores que um. Isso ocorre mesmo se trabalhando com 200 indivíduos, sendo que essas probabilidades crescem como o aumento no número de marcas utilizadas (k). Co-

mo anteriormente comentado, um modo de reduzir a probabilidade de obtenção de estimativas maiores que um consiste em aumentar o número de repetições. Nesse exemplo o aumento para 3 repetições diminuiu a probabilidade de obtenção de estimativas de p_m maiores que um de 0,37, quando se avaliaram 100 progênies com $p = 0,9$, para 0,31 e $P(p_{ms} > 1)$ de 0,43 para 0,39 na mesma situação.

Tabela 14. Probabilidade de obtenção de estimativas de p_m e p_{ms} maiores que um, em um experimento no qual foram avaliadas 100 e 200 progênies em 4 ambientes com 2 repetições por ambiente. Considerou-se $h^2 = 0,40$, e $\frac{3}{4}g^2 = \frac{3}{4}g_s^2$.

p_m	$P(p_m > 1)$					
	$g = 100$			$g = 200$		
	$k = 10$	$k = 15$	$k = 20$	$k = 10$	$k = 15$	$k = 20$
0,9	0,3641	0,3685	0,3730	0,2984	0,3010	0,3037
0,7	0,1533	0,1607	0,1686	0,0651	0,0677	0,0705
0,5	0,0533	0,0587	0,0647	0,0087	0,0095	0,0104
0,3	0,0163	0,0190	0,0222	0,0008	0,0009	0,0013
0,1	0,0045	0,0057	0,0071	0,0000	0,0000	0,0001

p_{ms}	$P(p_{ms} > 1)$					
	$g = 100$			$g = 200$		
	$k = 10$	$k = 15$	$k = 20$	$k = 10$	$k = 15$	$k = 20$
0,9	0,4327	0,4343	0,4361	0,4011	0,4020	0,4030
0,7	0,3041	0,3077	0,3116	0,2282	0,2302	0,2321
0,5	0,1999	0,2045	0,2095	0,1113	0,1133	0,1153
0,3	0,1233	0,1278	0,1328	0,0468	0,0482	0,0497
0,1	0,0716	0,0755	0,0798	0,0171	0,0179	0,0187

4.4 Conclusões

- i. Obteve-se um estimador da proporção da variância genética, explicada pelos marcadores moleculares, aplicável ao índice de seleção de Lande & Thompson (1992), no caso de experimentos conduzidos em mais de um ambiente, bem como um estimador da proporção da variância da interação genótipos x ambientes explicada pela interação QTLs x ambientes.
- ii. Os estimadores apresentados independem de desvios das frequências alélicas dos marcadores em relação às esperadas, porém possuem alta probabilidade de geração de estimativas fora do intervalo paramétrico, que pode ser reduzida com o aumento do número de ambientes e/ou repetições nos quais os genótipos são avaliados.
- iii. As expressões da esperança e da variância dos estimadores apresentados, obtidas via expansão em série de Taylor, não são triviais; em vista disto sugere-se obter tais estimativas via processos de reamostragem.

5 ANÁLISE DA INTERAÇÃO GENÓTIPOS X AMBIENTES EM MILHO, UTILIZANDO MARCADORES MOLECULARES CODOMINANTES

Resumo

Com base em um conjunto de dados de produtividade de grãos de 68 progênies S_2 de milho, genotipadas para 77 marcadores RFLP e avaliadas em "Top cross", com um testador do grupo Lancaster, em quatro diferentes localidades, realizou-se a identificação de QTLs associados à produtividade de grãos por meio de um modelo de regressão múltipla, utilizando o procedimento "stepwise" para a seleção das marcas associadas a QTLs. Após identificadas as marcas associadas a QTLs envolvidos no controle genético da produtividade de grãos, em cada local e na média dos quatro locais, realizou-se um estudo da interação QTLs x ambientes, por meio da quantificação da proporção da interação genótipos x ambientes devida à interação QTLs x ambientes, estimada via componentes de variância, e de análises individualizadas de seus efeitos. As metodologias utilizadas permitiram classificar as marcas identificadas como associadas a QTLs conforme o seu nível de interação, permitindo, ainda, a identificação de regiões cromossômicas, envolvidas no controle genético de caracteres quantitativos, com maior ou menor estabilidade ao longo dos ambientes.

Summary

Based on a set of grain yield data, obtained from the evaluation of 67 maize progenies genotyped for 77 codominant molecular markers, and evaluated as top crosses in four environments, the presented methodologies allowed the classification of markers associated to QTL, in relation to its level of genotype x environment interaction. The procedure also allowed the identification of chromosomic regions, involved in the genetical control of the considered trait, according to its stability, in relation to the observed environmental variation.

5.1 Introdução

Segundo Lynch & Walsh (1998), um problema que permanece não resolvido na análise de QTLs é a ocorrência da interação QTLs x ambientes. Duas metodologias têm sido utilizadas no estudo dessas interações: a consistência das associações marcador-caráter através dos ambientes e métodos de ANOVA, incorporando termos específicos para a referida interação. A primeira é muito incipiente, pois consiste simplesmente em verificar se uma associação marca-caráter é detectada em todos os ambientes avaliados. Em caso afirmativo, isto é tomado como evidência da não existência de interação marcadores x ambientes, enquanto que a detecção de associação em apenas alguns dos ambientes é sempre considerada como evidência de interação, embora um QTL possa ter efeito significativo em todos os ambientes, mesmo na presença de forte interação. De forma semelhante, baixos poderes de detecção podem resultar em que um QTL venha a ser detectado em apenas algumas repetições do experimento, mesmo que seus efeitos sejam idênticos em todos os ambientes. Fortalecendo essa explicação, Koester et al. (1993) observaram que QTLs com efeitos pequenos eram menos propensos de serem detectados através dos locais, do que QTLs de efeito pronunciado.

Lynch & Walsh (1998) afirmam que, pelo teste explícito da interação marcadores x ambientes, os métodos de ANOVA provêm uma medida mais sensível dos efeitos dessa interação. Atuando nessa linha, Cockerham & Zeng (1996) estendem a análise de variância do delineamento III de Comstock & Robinson (1952) para incluir os efeitos de ligação entre marcadores moleculares e QTLs, epistasia entre dois locos e utilização de gerações F_3 . Além disso, desenvolveram uma metodologia baseada em contrastes ortogonais, para estimação e teste de hipóteses sobre esses efeitos bem como suas interações com o ambiente.

As metodologias encontradas na literatura, para abordagem da interação QTLs x ambientes (Jansen et al., 1995; Romagosa et al., 1996; Korol et al., 1998 e Wang et al., 1999), consideram-na como um complicador da análise de QTLs e atuam no sentido de eliminar seu efeito sobre o seu mapeamento. Mas, nem sempre, a interação genótipos x ambientes é um complicador para o trabalho do melhorista, pois, quantificando o seu efeito, podem-se selecionar genótipos que sejam altamente adaptados a um determinado ambiente, captando, assim, seu efeito. Alternativamente, é possível selecionar genótipos que não apresentem essa interação, isto é, apresentem um comportamento estável nos ambientes em que são cultivados. Esse entendimento pode ser transportado para o nível de QTLs, de modo que, mediante a quantificação dos efeitos da interação QTLs x ambientes, é possível identificar aqueles mais estáveis, que podem ser muito úteis em programas de seleção assistida por marcadores moleculares, visando à produção de materiais com alta estabilidade, e os QTLs que apresentam alta interação, passíveis de serem utilizados em programas de seleção assistida visando à obtenção de genótipos específicos para um determinado tipo de ambiente. Deste modo, objetivou-se, no presente capítulo, aprofundar o conhecimento da interação QTLs x ambientes, por meio da quantificação da proporção da interação genótipos x ambientes devida à interação QTLs x ambientes, estimada via componentes de variâncias e de análises individualizadas de seus efeitos.

5.2 Material e Métodos

5.2.1 Material

Utilizou-se, no presente trabalho, uma parte dos dados apresentados por Rumin (2000), referentes à produtividade de grãos em kg/ha de 68 progênies S_2 de milho, obtidas do cruzamento de duas linhagens endogâmicas pertencentes ao grupo ISSS. Essas progênies foram genotipadas para 157 marcadores RFLP e avaliadas em "top crosses", com um testador do grupo Lancaster, em quatro diferentes ambientes no estado de Iowa, EUA, com duas repetições por ambiente, no ano agrícola de 1996. Esses dados foram gerados pelo departamento de pesquisa e desenvolvimento da empresa Garst Seeds, componente do então grupo Zeneca, com base em Slater, Iowa, EUA, no ano agrícola de 1996.

Devido ao fato de o interesse nesse conjunto de dados ser puramente didático e ilustrativo das técnicas desenvolvidas e discutidas nesta tese, trabalhou-se apenas com 77 das 157 marcas disponíveis no conjunto de dados apresentados por Rumin (2000).

Visando à seleção das 77 marcas, utilizadas no presente trabalho, para as quais as distâncias entre estas, em cM, estão ilustradas na Figura 16, eliminaram-se, inicialmente, todas as marcas que apresentaram mais de dois alelos por loco. No caso de marcas que compartilhavam a mesma posição no cromossomo, selecionou-se a que apresentou segregação mais próxima de $\frac{3}{8} : \frac{2}{8} : \frac{3}{8}$, que é a esperada em materiais com uma geração de autofecundação e frequência alélica 0,5.

5.2.2 Métodos

Detecção de QTLs

Uma vez que não se dispunha dos genótipos das linhagens parentais, a

Figura 16 - Distâncias (cM) entre as 77 marcas utilizadas, e cromossomos aos quais estão associadas. Obs: Os tamanho dos cromossomos não representam seu tamanho real.

Única técnica de mapeamento de QTLs possível de utilização é a análise de marcas simples ("Single Marker Analysis"), que foi realizada por meio de um modelo de regressão linear múltipla, utilizando o procedimento "stepwise" para seleção das marcas associadas a QTLs. O procedimento "stepwise" foi realizado considerando todo o genoma simultaneamente, isto é, utilizando todas as 77 marcas disponíveis, tendo sido adotados os níveis de significância de 0,10 para entrada de uma marca no modelo e 0,15 para sua permanência.

Consideraram-se, ainda, dois critérios para a detecção de QTLs:

1. Seleção realizada considerando todos os locais conjuntamente (Seleção conjunta);
2. Seleção realizada em cada local separadamente (Seleção por local).

Partição da variância da interação em efeitos explicados pelos marcadores e desvios

Para avaliação da interação genótipos x locais e partição da variância da interação genótipos x locais em efeitos de interação marcadores x locais e desvios x locais, adotou-se o seguinte modelo matemático:

$$y_{ij\cdot} = \mu + s_{\cdot} + r=s_{j(\cdot)} + g_i + (gs)_{i\cdot} + e_{ij\cdot} \quad (94)$$

em que

$y_{ij\cdot}$ é o valor observado da i -ésima progênie, $i = 1; 2; \dots; 68$, na j -ésima repetição, $j = 1; 2$; do $\cdot$ -ésimo local, $\cdot = 1; 2; \dots; 4$;

μ é uma constante inerente a todas as observações;

$s_{\cdot}$ é o efeito do $\cdot$ -ésimo local, $s_{\cdot} \sim N(0; \frac{3}{4})$;

$r=s_{j(\cdot)}$ é o efeito da j -ésima repetição dentro do $\cdot$ -ésimo local, $r=s_{j(\cdot)} \sim N(0; \frac{3}{4})$;

g_i é o efeito da i -ésima progênie, $g_i \sim N(0; \frac{3}{4})$;

$(gs)_{i\cdot}$ é o efeito da interação entre a i -ésima progênie e o $\cdot$ -ésimo local, $(gs)_{i\cdot} \sim N(0; \frac{3}{4})$;

$e_{ij\cdot}$ é o erro aleatório associado à observação $y_{ij\cdot}$, $e_{ij\cdot} \sim N(0; \frac{3}{4})$.

sendo g_i modelado por:

$$g_i = \sum_{t=1}^k \beta_t Z_{it} + h_i; \quad (95)$$

em que:

β_t é o coeficiente de regressão da marca t ;

Z_{it} é uma variável "dummy", codificadora do genótipo da i -ésima progênie quanto à marca t , assumindo os valores 1 para o genótipo homozigótico quanto ao alelo favorável ao caráter, 0 para o heterozigótico e -1 para o homozigótico desfavorável; ;

h_i é a porção de g_i não explicada pelas marcas (desvios).

Para o qual, o esquema da análise de variância, com as respectivas esperanças matemáticas dos quadrados médios, encontra-se apresentado na tabela 15.

As proporções da variância genética e da variância da interação genótipos x ambientes explicadas pelos marcadores moleculares (p_m e p_{ms}) foram estimadas por (Capítulo 3):

$$\hat{p}_m = \frac{\%_m^2}{\%_m^2 + \%_h^2};$$

$$\hat{p}_{ms} = \frac{\%_{ms}^2}{\%_{sm}^2 + \%_{hs}^2};$$

Tendo sido os estimadores $\%_m^2$, $\%_h^2$, $\%_{ms}^2$ e $\%_{hs}^2$ obtidos pelo método dos momentos, a partir das esperanças matemáticas dos quadrados médios apresentados na tabela 15.

Os erros padrões de $\hat{p}_m$ e de $\hat{p}_{ms}$ foram obtidos via bootstrap, tendo sido realizadas 2000 reamostragens bootstrap.

O efeito do critério de seleção sobre a proporção da variância da interação genótipos x locais, explicada pelos marcadores moleculares, foi avaliado em termos das estimativas $\hat{p}_{ms}$ e da análise AMMI, realizada com estimativas dos coeficientes de regressão linear das marcas em cada um dos quatro locais.

5.3 Resultados e Discussão

Na tabela 16, encontram-se apresentadas as marcas selecionadas pelo procedimento "stepwise" considerando os dois critérios de seleção (seleção conjunta e por local).

Tabela 15. Esquema da análise de variância conjunta com o desdobramento da soma de quadrados de tratamentos e interação, em efeitos explicados pelos marcadores moleculares e desvios, com as respectivas esperanças matemáticas dos quadrados médios.

FV	GL	SQ	E(QM)
Locais	$I-1$	SQ_{Loc}	$\frac{3}{4}^2 + r\frac{3}{4}_{gs}^2 + g\frac{3}{4}_{r=1}^2 + rg\frac{3}{4}^2$
Rep/Locais	$I(r-1)$	$SQ_{Rep/Loc}$	$\frac{3}{4}^2 + g\frac{3}{4}_{r=1}^2$
Progenies	$g-1$	SQ_{Prog}	$\frac{3}{4}^2 + r\frac{3}{4}_{gs}^2 + rl\frac{3}{4}_g^2$
Reg	k	SQ_{Reg}	$\frac{3}{4}^2 + r\frac{3}{4}_{hs}^2 + rl\frac{3}{4}_h^2 + \frac{(g_i-1)}{k}(r\frac{3}{4}_{ms}^2 + rl\frac{3}{4}_m^2)$
Desv	$g-k-1$	SQ_{Desv}	$\frac{3}{4}^2 + r\frac{3}{4}_{hs}^2 + rl\frac{3}{4}_h^2$
Prog x Loc	$(g-1)(I-1)$	$SQ_{Prog \times Loc}$	$\frac{3}{4}^2 + r\frac{3}{4}_{gs}^2$
Reg x Loc	$k(I-1)$	$SQ_{Reg \times Loc}$	$\frac{3}{4}^2 + r\frac{3}{4}_{hs}^2 + \frac{(g_i-1)}{k}r\frac{3}{4}_{ms}^2$
Desv x Loc	$(g-k-1)(I-1)$	$SQ_{Desv \times Loc}$	$\frac{3}{4}^2 + r\frac{3}{4}_{hs}^2$
Resíduo	$I(g-1)(r-1)$	$SQ_{Res\acute{d}uos}$	$\frac{3}{4}^2$
Total	$Irg-1$	SQ_{Total}	

Verifica-se, nesta tabela, que, ao realizar a seleção das marcas na análise conjunta, isto é, tendo por base a produtividade média dos genótipos nos quatro locais, 14 marcas foram identificadas como associadas a QTLs. Esse resultado concorda em parte com os apresentados por Melchinger et al. (1998), que, ao realizarem o mapeamento por intervalo composto, da produtividade média de grãos, de 107 progênies F_3 de milho, avaliadas em "Top cross" com dois testadores nos anos de 1990 e 1991 em duas localidades da Alemanha, encontraram nove QTL's associados a este caráter. Sendo que, destes nove, três pertenciam ao cromossomo 3, dois ao cromossomo 8, e os demais estavam localizados nos cromossomos 1, 2, 7 e 10.

Para uma melhor visualização das distâncias entre as marcas identificadas como associadas a QTLs, estas encontram-se ilustradas na figura 17, na qual se observa a pequena distância entre algumas delas, como, por exemplo, entre as marcas M_5 e M_6 , de apenas 4,6 cM. Estas duas marcas podem, então, ser consideradas como identificadoras de uma única região cromossômica associada a QTL's. Seguindo a

Tabela 16. Marcas selecionadas como associadas a QTL's, envolvidos no controle da produtividade de grãos de milho (kg/ha), pelo procedimento "stepwise" em cada local e na análise conjunta

Local 1	Local 2	Local 3	Local 4	Análise Conjunta
5c1	1c1	5c1	3c1	1c1
8c1	10c1	6c1	5c1	12c2
14c2	11c2	19c3	11c2	21c3
17c2	12c2	21c3	15c2	24c3
24c3	13c2	22c3	19c3	27c3
30c4	14c2	24c3	21c3	30c4
36c5	15c2	34c4	34c4	40c5
43c5	17c2	40c5	35c5	49c6
51c6	21c3	49c6	45c5	51c6
57c7	24c3	57c7	52c6	57c7
63c8	28c3	59c7	57c7	59c7
67c9	30c4	60c8	72c7	60c8
	35c5	62c8		61c8
	41c5	63c8		72c9
	45c5			
	49c6			
	51c6			
	52c6			
	65c9			
	72c9			
	74c9			
	76c10			

ci é o cromossomo no qual a marca está localizada.

metodologia de Melchinger et al. (1998), que consideraram QTLs identificados por diferentes testadores, localizados a menos de 20 cM um do outro, como sendo um único QTL e fundamentado em Darvasi et al. (1993), que verificaram, via simulação computacional de dados, que os intervalos de confiança para as posições dos QTL, estimadas via mapeamento por intervalo, não são menores que 10 cM e Dupuis & Siegmund (1999), que, ao compararem, também, via simulação computacional de dados, três metodologias para obtenção de intervalos de confiança para as posições dos QTL's, considerando vários espaçamentos entre marcas, posições do QTL no

intervalo entre duas marcas e efeitos dos mesmos, obtiveram estimativas variando de 2,2 cM a 21,3 cM, para esses intervalos, pode-se então considerar que marcas separadas por distâncias inferiores a 20 cM sejam identificadoras de uma única região cromossômica associada a QTLs. Deste modo, as 40 marcas selecionadas identificam 26 regiões cromossômicas. O número dessas regiões por cromossomos, bem como o número de QTL's envolvidos no controle da produtividade de grãos, já identificados e cadastrados no "Maize Database" (<http://www.agron.missouri.edu>), estão apresentadas na tabela 17, na qual se nota uma grande concordância entre o número de regiões aqui identificadas e o número de QTL's já cadastrados.

Figura 17 - Distâncias (cM) entre as 40 marcas identificadas como associadas a QTL's utilizadas, e cromossomos aos quais estão associadas. Obs: Os tamanho dos cromossomos não representam seu tamanho real.

Na tabela 16, observa-se uma grande discrepância entre os dois critérios com relação às marcas selecionadas. Uma vez que 14 marcas foram identificadas

Tabela 17. Número de regiões identificadas como associadas a QTL's e de QTL's envolvidos no controle da produtividade de grãos, cadastrados no "Maize Database" por cromossomo

Cromossomo	1	2	3	4	5	6	7	8	9	10
Regiões cromossômicas identificadas	4	2	4	2	4	2	2	3	2	1
QTL's cadastrados	4	1	0	2	3	5	1	2	2	1

como estando associadas a QTL's no local 1, 22 no local 2, 14 no local 3 e 12 no local 4, perfazendo um total de 38 diferentes marcas. Ao considerar todos os locais, simultaneamente, 14 marcas foram selecionadas, sendo que destas, nenhuma havia sido selecionada em todos os quatro locais, três o foram em três locais, quatro em dois, cinco em apenas um local e duas não haviam sido selecionadas em nenhum dos quatro locais.

Essa discrepância, quanto às marcas selecionadas em um e outro local, já era esperada, consistindo, inclusive, em um critério de identificação de interação QTLs x ambientes, conforme relatado por Lynch & Walsh (1998).

Vislumbrando uma melhor compreensão da variância da interação genótipos x ambientes, Venously & Barriga (1992) consideraram uma situação, como a esquematizada na tabela 18, em que se está admitindo que os genótipos de um determinado QTL foram avaliados em I ambientes. Na tabela 18, a nomenclatura, adotada por Vencovsky & Barriga (1992), foi adaptada para se adequar ao enfoque de interação QTLs x ambientes, em detrimento de genótipos x ambientes, como adotado pelos autores. Nessa tabela, está-se admitindo, ainda, que os efeitos genotípicos do QTL modificam-se conforme os ambientes, ou seja, os parâmetros a , d e 1 são diferentes de ambiente para ambiente, indicando que os genótipos reagem diferentemente de ambiente para ambiente.

A variância da interação QTLs x ambientes $V(GS)$ é dada por:

$$V(GS) = \sum_{i=1}^I V(G(S)_i) - IV(G(\hat{S}))$$

Tabela 18. Valores genotípicos de um QTL, avaliado em I ambientes

Genótipos	Frequência	Valores genotípicos					Médias
		Ambientes					
		1	2	...	I		
QQ	$\frac{1}{4}$	$1_1 + a_1$	$1_2 + a_2$	...	$1_I + a_I$	$1 + a$	
Qq	$\frac{1}{2}$	$1_1 + d_1$	$1_2 + d_2$	...	$1_I + d_I$	$1 + d$	
qq	$\frac{1}{4}$	$1_1 - a_1$	$1_2 - a_2$	...	$1_I - a_I$	$1 - a$	
Médias	1	$\frac{1}{2}(1_1 + d_1)$	$\frac{1}{2}(1_2 + d_2)$	...	$\frac{1}{2}(1_I + d_I)$	$\frac{1}{2}(1 + d)$	

Adaptado de Vencovsky & Barriga (1992)

em que:

 $V(G(S)_i)$ é a variância genética dentro do i-ésimo local, e $V(G(\bar{S}))$ é a variância genética média.

Deste modo, Vencovsky & Barriga (1992) demonstram que:

$$V(GS) = \frac{1}{2}V(a) + \frac{1}{4}V(d);$$

e concluíram que a variância da interação genótipos x ambientes ($\frac{1}{4}\sigma_{Qs}^2$) é estritamente de natureza genética, decorrente da instabilidade das manifestações genotípicas ao longo dos ambientes. Um fato interessante de lembrar é que os modelos biométricos empregados no presente trabalho, para explicar o fenótipo de um caráter, são ajustados utilizando os valores genotípicos de um indivíduo quanto a uma série de marcadores moleculares, que estão associados aos QTL's controladores do caráter. Desse modo, os componentes de variâncias estimados referem-se aos efeitos de marcadores moleculares associados a QTLs, e não aos QTLs propriamente ditos.

Em vista disso, estudou-se a natureza da interação entre efeitos associados a marcadores moleculares x ambientes ($\frac{1}{4}\sigma_{ms}^2$), que é objeto principal do presente trabalho. Para tanto, considerou-se um marcador molecular M, ligado em fase de associação a um QTL Q, apresentando uma frequência de recombinação c com o mesmo, avaliado em I ambientes (locais) como esquematizado na tabela 19.

Tabela 19. Valores genotípicos associados aos genótipos de um marcador molecular, avaliado em I ambientes

Genótipo	Gen.	Valores genotípicos						
		Marcas	Freq	QTL	Freq/Marcas	Ambientes		
						1	...	I
MM	1/4	QQ	(1 - c) ²	1 + a ₁	...	1 + a _I	1 + ā	
		Qq	2c(1 - c)	1 + d ₁	...	1 + d _I	1 + d̄	
		qq	c ²	1 - a ₁	...	1 - a _I	1 - ā	
Média/MM				1 + (1 - 2c)a ₁ + 2c(1 - c)d ₁	...	1 + (1 - 2c)a _I + 2c(1 - c)d _I	1 + (1 - 2c)ā + 2c(1 - c)d̄	
Mm	1/2	QQ	c(1 - c)	1 + a ₁	...	1 + a _I	1 + ā	
		Qq	(1 - 2c + 2c ²)	1 + d ₁	...	1 + d _I	1 + d̄	
		qq	c(1 - c)	1 - a ₁	...	1 - a _I	1 - ā	
Média/Mm				1 + (1 - 2c + 2c ²)d ₁	...	1 + (1 - 2c + 2c ²)d _I	1 + (1 - 2c + 2c ²)d̄	
mm	1/4	QQ	c ²	1 + a ₁	...	1 + a _I	1 + ā	
		Qq	2c(1 - c)	1 + d ₁	...	1 + d _I	1 + d̄	
		qq	(1 - c) ²	1 - a ₁	...	1 - a _I	1 - ā	
Média/mm				1 - (1 - 2c)a ₁ + 2c(1 - c)d ₁	...	1 - (1 - 2c)a _I + 2c(1 - c)d _I	1 - (1 - 2c)ā + 2c(1 - c)d̄	
Média				1 + 1/2 d ₁	...	1 + 1/2 d _I	1 + 1/2 d̄	

Assim sendo, tem-se que:

$$\begin{aligned}
 V(G(S)_i) &= \frac{1}{4} \left[1 + (1 - 2c)a_i + 2c(1 - c)d_i - \left(1 + \frac{1}{2}d_i \right) \right]^2 + \\
 &\frac{1}{2} \left[1 + (1 - 2c + 2c^2)d_i - \left(1 + \frac{1}{2}d_i \right) \right]^2 + \frac{1}{4} \left[1 - (1 - 2c)a_i + 2c(1 - c)d_i - \left(1 + \frac{1}{2}d_i \right) \right]^2 \\
 &= \frac{1}{2}a_i^2(1 - 2c)^2 + \frac{1}{4}d_i^2(1 - 8c + 24c^2 - 32c^3 + 16c^4)
 \end{aligned}$$

$$\begin{aligned}
 V(G(\bar{S})_i) &= \frac{1}{4} \left[1 + (1 - 2c)\bar{a} + 2c(1 - c)\bar{d} - \left(1 + \frac{1}{2}\bar{d} \right) \right]^2 + \\
 &\frac{1}{2} \left[1 + (1 - 2c + 2c^2)\bar{d} - \left(1 + \frac{1}{2}\bar{d} \right) \right]^2 + \frac{1}{4} \left[1 - (1 - 2c)\bar{a} + 2c(1 - c)\bar{d} - \left(1 + \frac{1}{2}\bar{d} \right) \right]^2 \\
 &= \frac{1}{2}\bar{a}^2(1 - 2c)^2 + \frac{1}{4}\bar{d}^2(1 - 8c + 24c^2 - 32c^3 + 16c^4)
 \end{aligned}$$

$$\begin{aligned}
 V(GS) &= \sum_{i=1}^I V(G(S)_i) - IV(G(\bar{S})) \\
 &= \sum_{i=1}^I \left[\frac{1}{2}a_i^2(1 - 2c)^2 + \frac{1}{4}d_i^2(1 - 8c + 24c^2 - 32c^3 + 16c^4) \right] - \\
 &I \left[\frac{1}{2}\bar{a}^2(1 - 2c)^2 + \frac{1}{4}\bar{d}^2(1 - 8c + 24c^2 - 32c^3 + 16c^4) \right] \\
 &= \frac{1}{2}(1 - 2c) \left[\sum_{i=1}^I a_i^2 - I\bar{a}^2 \right] + \frac{1}{4}(1 - 8c + 24c^2 - 32c^3 + 16c^4) \left[\sum_{i=1}^I d_i^2 - I\bar{d}^2 \right]
 \end{aligned}$$

$$\begin{aligned}
&= \frac{1}{2}(1 - 2c) \left[\sum_{i=1}^l a_i^2 - \frac{(\sum_{i=1}^l a_i)^2}{l} \right] + \frac{1}{4}(1 - 8c + 24c^2 - 32c^3 + \\
&16c^4) \left[\sum_{i=1}^l d_i^2 - \frac{(\sum_{i=1}^l d_i)^2}{l} \right] \\
&= \frac{1}{2}(1 - 2c)V(a) + \frac{1}{4}(1 - 8c + 24c^2 - 32c^3 + 16c^4)V(d)
\end{aligned}$$

$$V(GS) = \frac{1}{2}(1 - 2c)V(a) + \frac{1}{4}(1 - 2c)^4V(d) \quad (96)$$

Verifica-se, na expressão 96, que a variância da interação dos efeitos associados aos marcadores x ambientes é decorrente da instabilidade das manifestações genotípicas ao longo dos ambientes, expressa pelas variâncias dos efeitos aditivos e de dominância dos QTLs ($V(a)$ e $V(d)$), associados à distância do QTL à marca (c). Nota-se ainda, na expressão 96, que a principal fonte dessa interação é a variância dos efeitos aditivos, pois, como a constante que associa o efeito da distância do QTL à marca, que multiplica a variância dos efeitos de desvios de dominância, é elevada à quarta potência, esse valor será bem menor do que o multiplicador da variância dos efeitos aditivos, diminuindo, assim, sua contribuição à variância da interação. Por exemplo, se $c = 0,10$ ter-se-á de 96 que $V(GS) = (0,4)V(a) + (0,1024)V(d)$. Esse resultado concorda com os apresentados por Cocherham & Zeng (1996), que, ao analisarem os dados relativos à produtividade de grãos, altura de espiga, altura de plantas, área foliar, dias até o pendoamento, umidade de grãos e número de espigas de 264 progênies F_3 de milho, genotipadas para 67 marcadores RFLP e nove Izoenzimas, avaliadas em seis ambientes, observaram que os efeitos aditivos interagiram mais com os ambientes do que os efeitos de dominância em todos os caracteres avaliados.

Como pode ser verificado no Apêndice A, os coeficientes de regressão $\bar{\tau}_t$ são estimadores dos efeitos aditivos associados a cada marca (m_t). Assim, considerando que a localização do QTL em relação à marca (c) é a mesma, o fato de um coeficiente $\bar{\tau}_t$ ser estatisticamente diferente de zero em um local e não o ser em outro ou é decorrência de um erro tipo I, ou então de uma variação no efeito aditivo (a) do QTL associado à marca em questão, que, como demonstrado pela equação 96, é

a principal causa da interação QTLs x Locais.

Na Figura 18, estão apresentadas as produtividades médias dos genótipos homozigóticos para a marca m_{61} , que foi considerada associada a QTL's apenas na análise conjunta, a marca m_{25} , que foi na análise conjunta e nos locais 1, 2 e 3 e a marca m_{10} , que o foi apenas no local 2. Sabendo-se que o estimador $\hat{\mu}_t$, a menos de uma constante, igual ao contraste entre os dois genótipos homozigóticos, depreende-se que o fato de a marca m_{61} ter sido identificada como associada a QTLs apenas na análise conjunta decorre unicamente de fatores estatísticos, pois, como pode ser verificado nessa figura, as magnitudes dos contrastes entre as médias dos genótipos homozigóticos para esta marca não são muito diferentes. Deste modo, a obtenção de um resultado significativo no teste de associação entre esta marca e um QTL pode ter ocorrido em função de um erro tipo I, ou devido ao fato de, na análise conjunta, cada média ser obtida mediante de oito observações (2 por local), o que aumenta o poder do teste, em comparação aos realizados em cada local. Deste modo, justifica-se o fato de esta marca ter sido identificada como estando associada a um QTL apenas na análise conjunta, não existindo nenhuma explicação genética plausível para este resultado.

No caso da marca m_{24} , que não foi selecionada em apenas um local, e da m_{10} , que o foi em apenas um, observa-se que os contrastes que não foram estatisticamente significativos eram de magnitudes desprezíveis, em comparação aos que foram significativos. Neste caso, essas diferenças podem ser atribuídas a uma expressão variável do efeito dos QTL's associados a essas marcas, em outras palavras, à interação QTLs x Locais.

Esta expressão variável do efeito do QTL pode estar associada ao modo de ação do mesmo sobre o caráter. No caso da produtividade de grãos, que foi objeto do presente estudo, sabe-se que este é um caráter extremamente complexo, sendo influenciado por uma série de fatores, tais como a eficiência fotossintética, resistência à toxidez de alumínio, resistência à seca, etc. Seguindo essa linha de raciocínio, pode-se imaginar que uma planta, possuidora de um sistema radicular mais profundo, será

Figura 18 - Produtividade média (kg/ha) dos genótipos homozigotos (1 favorável e 1 desfavorável) considerando as marcas m_{61} , m_{24} e m_{10} , nos quatro locais avaliados e na média geral.

mais resistente à seca e, conseqüentemente, apresentará uma maior produtividade de grãos que outra com sistema radicular menos desenvolvido, quando cultivadas num ambiente com escassez d'água. Por outro lado, se essas duas plantas forem cultivadas em um local em que a água não é um fator limitante, e se considerar que a única diferença entre elas é o sistema radicular, elas apresentarão a mesma produtividade de grãos. Nesta situação hipotética, o gene que controla o tamanho do sistema radicular seria identificado como sendo um QTL para a produtividade de grãos no local com escassez d'água e não o seria no ambiente com abundância de água.

As partições da variância da interação genótipos x locais em interação QTL's x Locais e desvios por locais, considerando os modelos com as 14 marcas selecionadas na análise conjunta, as 38 na análise por local e as 40 considerando os dois critérios estão apresentadas na tabela 20. Nesta última tabela, pode-se verificar

que houve efeito significativo de progênies, avaliado em termos da variância genética entre progênies bem como da interação progênies x locais.

Tabela 20. Resultado da análise de variância conjunta, com desdobramento da soma de quadrados de progênies e da interação progênies x locais em efeitos explicados pelas marcas e desvios, em um experimento, no qual foram avaliadas 68 progênies de milho, genotipadas para 77 marcadores codominantes, quanto à produtividade de grãos (kg/ha), em 4 locais, com duas repetições por local.

seleção	FV	GL	SQ	QM	f	P(F > f)
	Locais	3	141244593,99	47081531,33		
	Rep/Locais	4	3053551,28	763387,82		
	Progênies	67	172166037,38	2569642,35	2,1685	<0,0000
	Prog. x Locais	201	238178092,54	1184965,63	1,4312	0,0043
	Resíduo	229	189595760,89	827929,09		
Análise conjunta	Regressao	14	115277873,99	8234133,86	3,8771	<0,0000
	Desvios	53	56888163,39	1073361,57	0,9058	0,6753
	Reg. x Locais	42	56497047,45	1345167,79	1,1772	0,2354
	Desv. x Locais	159	181681045,09	1142648,08	1,3801	0,0129
Análise por local	Regressao	38	135044088,33	3553791,80	1,4402	0,0732
	Desvios	29	37121949,05	1280067,21	2,0686	0,0026
	Reg. x Locais	114	184343899,56	1617051,75	2,6132	<0,0000
	Desv. x Locais	87	53834192,98	618783,83	0,7473	0,9416
Geral	Regressao	40	137176594,73	3429414,87	1,42737	0,0655
	Desvios	27	34989442,65	1295905,28	2,03006	0,0079
	Reg. x Locais	120	186471075,62	1553925,63	2,43425	<0,0000
	Desv. x Locais	81	51707016,92	638358,23	0,77103	0,9131
	Total	504	744238036,08			
	Média (kg/ha)		8978,05			
	CV(%)		10,13			

As 14 marcas selecionadas na análise conjunta explicaram muito bem a variância genética entre progênies, haja vista o efeito de marcas ter sido altamente significativo e o de desvios não, indicando a não ausência de marcas importantes na descrição do caráter. Fortalecendo esta observação, verifica-se, na tabela 21, que essas marcas explicaram 100% da variância genética ($p_m = 1;05$). Porém elas não foram tão eficientes na descrição da interação progênies x locais, pois o efeito da

interação marcas x locais foi não significativo e explicou apenas 12% da variância da interação progênies x locais ($\hat{\rho}_{ms} = 0,12$). Do ponto de vista de melhoramento, o fato de o efeito da interação desvios x locais ter sido significativo é muito interessante, pois indica que a maior parte da interação progênies x locais deve-se à interação desvios x locais, levando à conclusão de que os QTL's associados a essas 14 marcas são bastante estáveis, nos ambientes em que foram avaliados.

Tabela 21. Estimativas das proporções da variância genética e da variância da interação genótipos x locais explicadas pelos marcadores moleculares, com seus respectivos erros padrões

Critério de seleção	$\hat{\rho}_m$	$\hat{\rho}_{ms}$
Análise conjunta (14 marcas)	1,05 (0,24)	0,12 (3,94)
Análises individuais (38 marcas)	0,52 (0,22)	1,58 (4,79)
Ambos (40 marcas)	0,52 (0,17)	1,53 (7,94)

Considerando as 38 marcas selecionadas nas análises por local, nota-se que a proporção da variância genética explicada pelas marcas cai para 52% ($\hat{\rho}_m = 0,52$), sendo que, neste caso, o efeito de marcas foi não significativo e o desvios sim, indicando a ausência de marcas importantes para a descrição do caráter. Como verificado na tabela 16, das 38 marcas consideradas por esse critério, apenas as marcas m_{27} e m_{61} figuram entre as consideradas no critério anterior e não neste. Poder-se-ia, então, pensar que estas duas marcas têm um valor considerável na descrição da produtividade de grãos, mas, ao incluí-las no modelo (modelo geral), os resultados não se alteraram, isto é, a proporção da variância genética explicada pelo modelo não foi alterada, nem tampouco o resultado dos testes de hipóteses. Este fato reforça a hipótese de que a seleção dessas duas marcas apenas na análise conjunta tenha ocorrido devido a erros tipo I, em detrimento à de que se deva a um maior poder da análise conjunta.

Apesar de ter apresentado pior desempenho quanto à explicação da variância genética, o modelo com as 38 marcas selecionadas na análise por local explicou completamente a interação progênies x locais ($\rho_{ms} = 1;58$), sendo que, neste caso, o efeito da interação desvios x locais foi não significativo. Este fato não foi alterado ao se considerar o modelo com as 40 marcas, indicando que as marcas m_{27} e m_{61} muito pouco contribuíram para explicação da interação progênies x locais.

Do exposto, infere-se que a seleção de marcas realizada na análise conjunta é uma boa estratégia para identificar as marcas associadas aos QTLs, que pouco contribuem para interação genótipos x ambientes, podendo ser utilizadas em programas de seleção assistida que visem a obtenção de materiais com uma alta estabilidade. Porém este critério não é capaz de identificar as associadas aos QTLs, que apresentam alta interação com os ambientes, notadamente, quando esta interação for do tipo complexa.

Os valores $\hat{\lambda}_t$, utilizados para realização da análise AMMI estão apresentados na tabela 22. Dado que o experimento possui apenas quatro locais, o efeito da interação genótipos x ambientes ($\hat{\lambda}$ x locais) pode ser desdobrado em no máximo três eixos principais. Na figura 19, encontra-se apresentado o biplot para os quarenta marcadores moleculares utilizados, considerando o primeiro e o segundo eixos principais, que explicaram 87% da variância da interação $\hat{\lambda}$ x locais. Vale ressaltar que, neste biplot, se omitiram os efeitos dos quatro locais somente para facilitar a observação dos efeitos das marcas.

Tabela 22. Coeficientes de regressão ($\hat{\beta}_t$) associados às marcas em cada um dos quatro locais e na análise conjunta

Marca	Local 1	Local 2	Local 3	Local 4	Conjunta
m1	246,155	-668,32	-174,502	2,581	-180,525
m3	-392,493	186,61	-64,871	10,510	-71,497
m5	458,742	27,99	-756,850	469,046	173,475
m6	300,835	-66,73	-638,130	308,065	63,586
m8	357,570	-3,98	381,646	-263,747	83,170
m10	-126,648	-500,57	5,464	-118,667	-157,179
m11	-192,635	-404,62	-5,342	324,434	-10,360
m12	126,867	716,95	3,978	-34,042	182,024
m13	51,517	-421,38	44,720	22,172	-58,833
m14	-308,453	-289,91	190,597	-7,916	-122,398
m15	-3,762	900,43	-178,107	-235,839	153,865
m17	486,044	-718,44	-199,237	204,678	-21,249
m19	215,854	93,36	416,422	-426,623	33,600
m21	109,740	543,51	631,499	46,962	262,513
m22	-135,638	86,96	501,037	1,487	67,096
m24	375,242	1151,93	315,763	-41,780	440,350
m27	199,969	-88,52	-214,944	-146,712	-96,215
m28	116,073	-358,91	-244,316	53,479	-113,524
m30	408,176	-69,39	-104,776	81,218	97,856
m34	253,906	64,27	-232,988	-128,940	15,917
m35	-157,601	-252,65	5,183	-90,588	-172,467
m36	246,477	79,61	-62,085	-73,121	6,728
m40	-245,644	-30,58	468,858	60,517	7,180
m41	-143,645	484,70	-136,949	21,023	29,947
m43	359,462	22,38	40,680	-9,451	91,410
m45	205,448	-183,71	-14,021	231,305	89,601
m49	-107,991	448,41	481,208	-213,014	109,792
m51	596,420	334,15	-205,335	368,206	308,797
m52	-257,287	409,56	-7,243	215,718	79,647
m57	363,366	-30,39	52,915	201,857	143,600
m59	101,101	-123,51	384,555	-28,217	30,013
m60	127,418	219,28	405,466	45,965	195,720
m61	210,080	133,21	307,629	123,804	171,758
m62	31,164	-68,71	532,047	-306,337	68,585
m63	133,914	-17,32	-686,030	84,378	-129,960
m65	-47,900	393,42	158,421	87,385	162,274
m67	204,094	-180,98	-44,984	52,493	-0,160
m72	-177,946	210,93	32,828	409,545	144,294
m74	-165,474	-580,85	-66,838	70,458	-187,518
m76	78,119	672,51	-20,827	-217,795	176,003

Nesta figura, as marcas situadas mais próximas à origem dos eixos (coordenadas 0,0) são as que menos contribuem para interação, sendo

Figura 19 - Biplot das 40 marcas seleccionadas, considerando o primeiro e segundo eixos principais, os quais explicaram, conjuntamente, 87% da variância da interação Δ x locais.

que esta contribuição aumenta à medida que a marca se afasta da origem. Observa-se, então, que, das 14 marcas seleccionadas na análise conjunta ($M_1; M_{12}; M_{21}; M_{24}; M_{27}; M_{30}; M_{40}; M_{49}; M_{51}; M_{57}; M_{59}; M_{60}; M_{61}; M_{72}$), seis estão localizadas na região mais central do gráfico, ou seja, entre as mais estáveis. Sete estão na região intermediária e apenas uma está entre as seis, mais instáveis. As marcas M_{27} e M_{61} , fonte de discordância entre os dois critérios, estão localizadas na região de baixa contribuição para a interação, confirmando as conclusões obtidas a partir da tabela 20.

Confrontando as figuras 17 e 19, identifica-se uma associação entre a classificação das marcas quanto à estabilidade dos efeitos dos QTLs, aos quais estão associadas, e suas localizações nos cromossomos, pois, das seis marcas classificadas como instáveis, duas localizam-se no cromossomo 1, estando a apenas 4,3 cM uma da outra (marcas M_5 e M_6), duas outras (M_{15} e M_{16}) estão no cromossomo 2, e

as demais encontram-se uma no cromossomo 2 e a outra no cromossomo 8. Fato semelhante ocorre para aquelas consideradas estáveis, haja vista as duas únicas marcas identificadas como associadas a QTLs no cromossomo 4 terem sido classificadas como estáveis, das quatro pertencentes ao cromossomo 9, três serem estáveis, e nos cromossomos 5 e 8 existirem, em cada um, duas marcas adjacentes com esta classificação.

No biplot completo, isto é, incluindo os locais, apresentado na figura 20, verifica-se que o efeito de locais foi o que mais contribuiu para interação QTLs x locais, sendo que destes as maiores contribuições foram dadas pelos locais dois e três (mais distantes em relação à origem do gráfico). Nota-se ainda que os locais um e quatro apresentaram comportamento bastante semelhante. Os polígonos ilustrados procuram mostrar a influência dos locais dois e três sobre as marcas selecionadas, evidenciando a tendência das marcas selecionadas nesses locais concentrarem-se próximas a eles. Comparando esta figura com a tabela 16, verifica-se que, das seis marcas selecionadas apenas no local três (M_6 ; M_{22} ; M_{40} ; M_{59} ; M_{60} e M_{62}), quatro encontram-se na área delimitada para o referido local (M_{22} ; M_{40} ; M_{60} e M_{62}). Das nove selecionadas apenas no local 2 (M_1 ; M_{10} ; M_{12} ; M_{13} ; M_{28} ; M_{41} ; M_{65} ; M_{74} ; e M_{76}), cinco estão na sua respectiva área (M_{12} ; M_{41} ; M_{65} ; M_{74} ; e M_{76}), sendo que as outras quatro (M_1 ; M_{10} ; M_{13} ; e M_{28}) estão agrupadas no segundo quadrante, próximo à origem. As marcas M_{21} e M_{49} , que são comuns aos locais dois e três estão localizadas numa posição intermediária aos dois, região marcada como sendo de influência de ambos locais. Desse modo, o biplot das estimativas dos coeficientes de regressão $\hat{a}$ pode ser utilizado para identificação das marcas que são mais importantes em um determinado local e que devem, portanto, ser priorizadas em um programa de seleção assistida específico para aquele local.

5.4 Conclusões

- i. As metodologias apresentadas possibilitaram classificar as marcas associadas a QTLs quanto ao seu grau de interação genótipos x ambientes, permitindo,

Figura 20 - Biplot das 40 marcas selecionadas e os quatro locais, considerando o primeiro e segundo eixos principais, os quais explicaram conjuntamente, 87% da variância da interação $\hat{\alpha} \times$ locais.

ainda, a identificação de regiões cromossômicas envolvidas no controle genético de caracteres quantitativos, com maior ou menor estabilidade quanto à expressão destes caracteres.

- ii. Quando se realizar a avaliação fenotípica dos genótipos em mais de um ambiente, é desaconselhável que se proceda o mapeamento de QTL's tendo por base apenas o desempenho médio dos genótipos, quando a interação genótipos x ambientes for significativa.

6 CONCLUSÕES

- i. O procedimento baseado em regressão múltipla, utilizando o procedimento "step-wise" para seleção de variáveis, foi mais poderoso, em identificar as marcas associadas a QTL's, do que os procedimentos baseados em testes individuais, utilizando tanto o critério FDR, quanto o de Bonferroni para o controle do nível de significância conjunto.
- ii. Dentre os critérios FDR e de Bonferroni, que são aplicáveis às metodologias de mapeamento por intervalo, o primeiro apresentou-se em geral mais poderoso, devendo, portanto, ser preferido para o controle do nível de significância conjunto.
- iii. O poder de detecção de QTL's é função direta da herdabilidade do caráter, podendo ser aumentado, por meio do uso de um maior número de repetições nos experimentos para avaliação fenotípica do caráter sob análise.
- iv. Obteve-se um estimador da proporção da variância genética, explicada pelos marcadores moleculares, aplicável ao índice de seleção de Lande & Thompson (1992), no caso de experimentos conduzidos em mais de um ambiente, bem como um estimador da proporção da variância da interação genótipos x ambientes explicada pela interação QTLs x ambientes.
- v. Os estimadores apresentados independem de desvios das frequências alélicas dos marcadores em relação às esperadas, porém possuem alta probabilidade de geração de estimativas fora do intervalo paramétrico, que pode ser reduzida com o aumento do número de ambientes e/ou repetições nos quais os genótipos são avaliados.

- vi. As expressões da esperança e da variância dos estimadores apresentados, obtidas via expansão em série de Taylor, não são triviais; em vista disto sugere-se obter tais estimativas via processos de reamostragem.
- vii. As metodologias apresentadas possibilitaram classificar as marcas associadas a QTLs quanto ao seu grau de interação genótipo x ambiente, permitindo, ainda, a identificação de regiões cromossômicas envolvidas no controle genético de caracteres quantitativos, com maior ou menor estabilidade quanto à expressão destes caracteres.
- viii. Quando se realizar a avaliação fenotípica dos genótipos em mais de um ambiente, é desaconselhável que se proceda o mapeamento de QTL's tendo por base apenas o desempenho médio dos genótipos, quando a interação genótipos x ambientes for significativa.

REFERÊNCIAS BIBLIOGRÁFICAS

- BEARZOTI, E. Simulação de seleção recorrente assistida por marcadores moleculares em espécies autógamas. Piracicaba, 1997. 230p. (Doutorado) - Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo.
- BEARZOTI, E. Mapeamento de QTL. In: JORNADA EM GENÉTICA E MELHORAMENTO DE PLANTAS, 2., Goiânia, 2000. Análise de QTL no melhoramento de plantas. Goiânia: FUNAPE, 2000. p. 93-223.
- BEARZOTI, E.; VENCOSKY, R. Estimation of the proportion of genetic variance explained by molecular markers. *Genetics and Molecular Biology*, v.21, n.4, p.557-566, 1998.
- BENJAMINI, Y.; HOCHBERG, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B*, v.57, n.1, p.289-300, 1995.
- BERKE, T.G.; ROCHEFORD, T.R. Quantitative trait loci for flowering, plant and ear height, and kernel traits in maize. *Crop Science*, v. 35, p. 1542-1549, 1995.
- CHURCHIL, G.A.; DOERGE, R.W. Empirical threshold values for quantitative trait mapping. *Genetics*, v.138, p.693-971, Nov. 1994.
- COCKERHAN, C.C. An extension of the concept of partitioning hereditary variance for analysis of covariance among relatives when epistasis is present. *Genetics*, v.39, p.859-882, 1954.

- COCKERHAN, C.C.; ZENG, Z.B. Design III with marker loci. *Genetics*, v.143, p. 1437-1456, July 1996.
- COMSTOCK, R.E.; ROBINSON, H.F. Estimation of average dominance of genes. In: GOWEN, J. W. (Ed.) *Heterosis*. Ames: Iowa State College, 1952. p. 495-516.
- CROW, J.F.; KIMURA, M. *An introduction to population genetics theory*. New York: Harper & Row Publ., 1970. 591p.
- DARVASI, A.; WEINREB, V.; MINKE, V.; WELLER, J.I.; SOLLER, M. Detecting marker-QTL linkage and estimating QTL gene effect and map location using a saturated genetic map. *Genetics*, v.134, p.943-951, July 1993.
- DEMPSTER, A.P.; LAIRD, N.M.; RUBIN, D.B. Maximum likelihood from incomplete data via EM algorithm. *Journal of the Royal Statistical Society. Series B*, v.39, p.1-38, 1977.
- DUPUIS, J.; SIEGMUND, D. Statistical methods for mapping quantitative trait loci from a dense set of markers. *Genetics*, v 151, p.373-386, 1999.
- FALCONER, D.S.; MACKAY, T.F.C. *Introduction to quantitative genetics*. 4. ed. New York: Longman, 1996. 464p.
- FERREIRA, M.E; GRATTAPAGLIA, D. *Introducao ao uso de marcadores moleculares em análise genética*. 3.ed. Brasília: EMBRAPA, CENARGEN, 1998. 220p.
- FISHER, R.A. The correlations between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburg*, v.52, p.399-433, 1918.
- GOLDMAN, I.L.; ROCHEFORD, T.R.; DUDLEY, J.W. Molecular markers associate with maize kernel oil concentration in Illinois hight protein x Illinois low protein cross. *Crop Science*, v.34, p.908-915, 1994.

- HALDANE, J.B.S. The combination of linkage values, and the calculation of distance between the loci of linked factors. *Journal of Genetics*, v.8, p.299-309, 1919.
- HALEY, C.S.; KNOTT, S.A. A simple regression method for mapping quantitative trait loci in line crosses using anchoring markers. *Heredity*, v.69, p.315-324, 1992.
- HOFFMANN, R.; VIEIRA, S. *Análise de regressão: uma introdução à econometria*. São Paulo: Editora Hucitec, 1998. 378p.
- JANSEN, R.C. A general mixture model for mapping quantitative trait loci by using molecular markers. *Theoretical and Applied Genetics*, v.85, p.252-260, 1992.
- JANSEN, R.C. Interval mapping of multiple quantitative trait loci. *Genetics*, v.135, p.205-211, 1993.
- JANSEN, R.C.; VAN OOIJEN, J.W.; STAM, P.; LISTER, C.; DEAN, C. Genotype-by-environment interaction in genetic mapping of multiple quantitative trait loci. *Theoretical and Applied Genetics*, v.91, p.33-37, 1995.
- JIANG, C.; ZENG, Z.B. Multiple trait analysis of genetic mapping for quantitative trait loci. *Genetics*, v.140, p.1111-1127, July 1995.
- KAO, C.H.; ZENG, Z.B. General formulas for the obtaining the MLEs and the asymptotic variance-covariance matrix in mapping quantitative trait loci when using EM algorithm. *Biometrics*, v.53, p.653-665, 1997.
- KAO, C.H.; ZENG, Z.B.; TEASDALE, R.D. Multiple interval mapping for quantitative trait loci. *Genetics*, v.152, p.1203-1216, July 1999.
- KOESTER, R.P.; SISCO, P.H.; STUBER, C.W. Identification of quantitative trait loci controlling days to flowering and plant height in two near isogenic lines of maize. *Crop Science*, v.33, p.1209-1216, 1993.

- KOROL, A.B.; RONIN, Y.I.; NEVO, E. Approximate analysis of QTL-Environment interaction with no limits on the number of environments. *Genetics*, v.148, p.2015-2028, Apr. 1998.
- KOSAMBI, D.D. The estimation of map distances from recombination values. *Annals of Eugenics*, v.12, p.172-175, 1944.
- LANDE, R.; THOMPSON, R. Efficiency of marker-assisted selection in the improvement of quantitative traits. *Genetics*, v.126, p.743-756, 1990.
- LANDER, E.S.; BOTSTEIN, D. Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*, v.121, p.185-199, 1989.
- LIU, B.H. Statistical genomics, linkage, mapping and QTL analysis. Boca Raton: CRC Press, 1998. 611p.
- LYNCH, M.; WALSH, B. Genetics and analysis of quantitative traits. Sunderland: Sinauer Associates, 1998. 980p.
- MATHER, K.; JINKS, J.L. Introdução à genética biométrica. Trad. de F.A.M. DUARTE et al. Ribeirão Preto: SBG, 1984. 242p.
- MELCHINGER, A.E.; FRIEDRICH UTZ, H.; SCHÖN, C.C. Quantitative trait locus (QTL) mapping using different testers and independent population samples in maize reveals low power of QTL detection and large bias in estimates of QTL effects. *Genetics*, v.149, p.383-403, 1998.
- MENG, X.L.; RUBIN, D.B. Maximum Likelihood estimation via ECM algorithm: a general framework. *Biometrika*, v.80, p.267-278, 1993.
- MILACH, S.C.K. Marcadores moleculares em plantas Porto Alegre: Gráfica da UFRGS, 1998. 141p.
- MONTGOMERY, D.C.; PECK, E.A. Introduction to linear regression analysis 2.ed. New York: John Wiley, 1992. 527p.

- MOOD, A.L; GRAYBILL, F.A.; BOES, D.C. Introduction to the theory of statistics. 3.ed. Tokyo: McGraw-Hill Kogakusha, 1974. 564p.
- MORGAN, T.H. Chromosomes and Heredity. Amererican Naturalist, v.44, p.194, 1919.
- MORTON, N.E. Sequential tests for the detection of linkage American Journal of Human Genetics, v.7, p.277-318,1955.
- NETER, J.; WASSERMAN, W.; KUTNER, M.H. Applied linear statistical models: regression, analysis of variance and experimental designs. Home-wood: Richard D. Irwin, 1985. 1127p.
- PROVINCE, M.A. Sequential methods of analysis for genome scan. In RAO, D.C.; PROVINCE, M.A. Dissection of complex traits. San Diego: Academic Press, 1999. 583p. (Advances in Genetics, 42)
- ROMAGOSA, I.; ULLRICH, S.E.; HAN, F.; HAYES, P.M. Use of additive main e®ects and multiplicative interaction model in QTL mapping for adaptation in barley. Theoretical and Applied Genetics, v.93, p.30-37, 1996.
- RUMIN, G.C.R. Índice de seleção de linhagens endogâmicas de milho (Zea mays L.) visando populações sintéticas baseadas em marcadores genéticos. Piracicaba, 2000. 200p. Dissertação (Mestrado) - Escola Superior de Agricultura "Luiz de Queiroz", Universidade de Sao Paulo.
- SATTERTHWAITE, F.E. An approximate distribution of estimates of variance components. Biometrics Bulletin, v.2, p.110-114, 1946.
- SAX, K. The association of size differences with seed-coat pattern and pigmentation in Phaseolus vulgaris. Genetics, v.8, p.552-560, 1923.
- SEARLE, S.R. Linear models. New York: John-Wiley, 1971. 501p.

- SEARLE, S.R.; CASELA, J.; McCULLOCH, C.E. Variance components. New York: John-Wiley, 1992. 532p.
- SOUTTHEY, B.R.; FERNANDO, R.L. Controlling the proportion of False Positive among Significant Results in QTL Detection. In:WORLD CONGRESS ON GENETICS APPLIED TO LIVESTOCK PRODUCTION, 6., Armindale, 1996. Proceedings. Armindale: University of New England, 1996. v.26, p.221-224.
- VENCOVSKY, R.; BARRIGA, P. Genética biométrica no melhoramento. Ribeirao Preto: Sociedade Brasileira de Genética, 1992. 486p.
- WANG, D.L.; ZHU, J.; LI, Z.K.; PATERSON, A.H. Mapping QTLs with epistatic effects and QTL x environment interaction by mixed linear model approaches. Theoretical and Applied Genetics, v.99. p.1255-1264, 1999.
- WEIR, B.S. Genetic data analysis II. Sunderland: Sinauer Associates, 1996. 445p.
- WELLER, J.L. Maximum likelihood techniques for the mapping and analysis of quantitative trait loci with the aid of genetic markers. Biometrics, v.42, p.627-640, 1986.
- WELLER, J.L.; SONG, J.Z.; HEYEN, D.W.; LEWIN, H.A.; RON, M. A new Approach to the Problem of Multiple Comparisons for Detection of Quantitative Trait Loci. In:WORLD CONGRESS ON GENETICS APPLIED TO LIVESTOCK PRODUCTION, 6., Armindale, 1996. Proceedings. Armindale: University of New England, 1996. v.26, p.229-231.
- ZENG, Z.B. Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. Proceedings of the National Academic of Science of the USA, v.90, p.10972-10976, 1993.

ZENG, Z.B. Precision mapping of quantitative trait loci. *Genetics* , v.136, p.1457-1466, 1994.

APÊNDICE

Alguns aspectos sobre a interpretação do coeficiente de regressão linear em análises de QTLs, utilizando famílias S_n e Top-crosses

Introdução

O procedimento de procura e localização de QTLs é denominado mapeamento de QTLs. Os delineamentos utilizados para tal, em grande parte das espécies cultivadas, baseiam-se no uso de populações segregantes, F_2 , ou retrocruzamentos, obtidas pelo cruzamento de duas linhagens homozigóticas fenotipicamente extremas quanto ao caráter quantitativo em questão. Um determinado número de indivíduos dessas populações (geralmente algumas centenas) são avaliados para os caracteres quantitativos de interesse, e genotipados para uma série de marcadores moleculares, distribuídos a intervalos regulares pelo genoma. Se houver uma diferença na média do fenótipo entre as classes genotípicas do marcador, pode-se inferir que existe um QTL ligado a aquele marcador.

Um modelo de regressão linear simples pode ser utilizado para identificar QTLs ligados a marcadores moleculares do seguinte modo:

Considere-se o modelo de regressão linear simples:

$$y_i = \mu_0 + \mu_1 x_i + e_i; \quad (97)$$

em que:

y_i é o valor fenotípico observado do i -ésimo indivíduo;

x_i é uma variável "dummy" codificadora dos genótipos do marcador;

μ_0 é o intercepto do modelo de regressão;

μ_1 é o coeficiente angular do modelo;

e_i é o erro aleatório associado a observação y_{ij} ; $e_{ij} \sim N(0; \sigma^2)$;

Considerando um único loco QTL, com dois alelos Q e q, existirão três possíveis genótipos na população QQ, Qq e qq. Atribuindo-se, arbitrariamente os valores genotípicos $+a$ a um dos homozigotos $-a$ ao outro e d para o heterozigoto (aqui será adotado Q como o alelo favorável ao caráter, portanto o valor $+a$ será atribuído ao genótipo QQ). Tem-se a escala de valores genotípicos apresentadas na Figura 21, na qual o valor a é o efeito aditivo do QTL, d é o desvio de dominância, de tal forma que, se Q for dominante em relação a q, d será positivo, se q for dominante, d será negativo, se não houver dominância d será nulo e se houver sobredominância, d será maior que a ou menor que $-a$.

Figura 21 - Representação de um QTL, com seus valores genotípicos

Considerando este modelo, serão adotados os seguintes valores para a variável x_i , em 97, $x_i = -1$ se o genótipo do i -ésimo indivíduo for qq, $x_i = 0$ se for Qq e $x_i = 1$ se for QQ.

De modo geral, um resultado significativo para o teste da hipótese

$$\begin{cases} H_0 : \beta_1 = 0 & \text{vs} \\ H_a : \beta_1 \neq 0 \end{cases}$$

é tomado como indicativo da presença de um QTL associado à marca em questão. Porém, dependendo do delineamento genético utilizado para o mapeamento, o coeficiente β_1 do modelo pode ter diferentes interpretações. Portanto, pretende-se no presente trabalho discutir alguns aspectos relativos à interpretação desse coeficiente em diferentes delineamentos.

Retrocruzamentos

Esquema do delineamento:

Ao utilizar-se este delineamento para o mapeamento de QTLs pressupõe-se que os indivíduos das populações RC_1 ou RC_2 serão genotipados, quanto a uma série de marcadores moleculares, e estes mesmos indivíduos serão avaliados quanto ao caráter quantitativo em questão. Procedendo-se deste modo, a distribuição das frequências genotípicas esperadas, para o RC_1 , encontram-se apresentadas na tabela 23

Tabela 23. Frequências genotípicas esperadas no RC_1 , considerando dois locos ligados em fase de associação, com uma frequência de recombinação c entre os mesmos

Genótipo	Frequência	Valor Genotípico	Média da classe
MQ/MQ	$\frac{1}{2}(1 - c)$	a	$\frac{1}{2}a(1 - c) + \frac{1}{2}cd$
MQ/Mq	$\frac{1}{2}c$	d	
MQ/mQ	$\frac{1}{2}c$	a	$\frac{1}{2}ac + \frac{1}{2}(1 - c)d$
MQ/mq	$\frac{1}{2}(1 - c)$	d	

c é a frequência de recombinação entre o marcador e o QTL.

Ajustando um modelo de regressão com as médias dos indivíduos para cada classe genotípica do marcador, virá:

$$Y = X\beta + \epsilon$$

$$\begin{bmatrix} y_{MM} \\ y_{Mm} \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + "$$

Para o qual o estimador de mínimos quadrados é: $\hat{\beta} = (X^0 X)^{-1} X^0 Y$

então,

$$\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} (y_{MM} + y_{Mm}) \\ (y_{MM}) \end{bmatrix}$$

e a esperança matemática de $\hat{\beta}$ será:

$$E \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = E \begin{bmatrix} y_{MM} \\ (y_{MM} - y_{Mm}) \end{bmatrix}$$

$$E \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = \begin{bmatrix} \frac{1}{2}ac + \frac{1}{2}(1-c)d \\ \frac{1}{2}a(1-c) + \frac{1}{2}cd - [\frac{1}{2}ac + \frac{1}{2}(1-c)d] \end{bmatrix}$$

$$E \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = \begin{bmatrix} \frac{1}{2}ac + \frac{1}{2}(1-c)d \\ \frac{1}{2}(a-d)(1-2c) \end{bmatrix}$$

Deste modo, testar a hipótese:

$$\begin{cases} H_0 : \beta_1 = 0 & \text{vs} \\ H_a : \beta_1 \neq 0 \end{cases}$$

equivale a testar a hipótese de que $(a - d)$ seja nulo, ou $c = 0,5$.

Considerando-se o retrocruzamento com o outro pai (RC₂) ter-se-ia a distribuição de frequências apresentadas na tabela 24.

Neste caso, procedendo-se de modo análogo ao RC₁ obtém-se que:

$$E \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = \begin{bmatrix} -\frac{1}{2}ac + \frac{1}{2}d(1-c) \\ \frac{1}{2}(a+d)(1-2c) \end{bmatrix};$$

Portanto, testar a hipótese:

$$\begin{cases} H_0 : \beta_1 = 0 & \text{vs} \\ H_a : \beta_1 \neq 0 \end{cases}$$

utilizando um RC₂, equivale a testar se $(a + d) = 0$ ou $c = 0,5$.

Tabela 24. Frequências genotípicas esperadas no RC_2 , considerando dois locos ligados em fase de associação, com uma frequência de recombinação c entre os mesmos

Genótipo	Frequência	Valor Genotípico	Média da classe
mQ/MQ	$\frac{1}{2}(1 - c)$	d	$-\frac{1}{2}ac + \frac{1}{2}d(1 - c)$
mQ/Mq	$\frac{1}{2}c$	$-a$	
mQ/mQ	$\frac{1}{2}c$	d	$-\frac{1}{2}a(1 - c) + \frac{1}{2}dc$
mQ/mq	$\frac{1}{2}(1 - c)$	$-a$	

c é a frequência de recombinação entre o marcador e o QTL.

Geração F_2

Esquema do delineamento:

Assim como no caso dos retrocruzamentos, neste delineamento pressupõe-se a genotipagem e avaliação dos indivíduos F_2 . Neste tipo de delineamento existem 10 possíveis classes genotípicas, para as quais as frequências esperadas, considerando o marcador e o QTL ligados em fase de associação encontram-se na tabela 25

Ajustando o modelo de regressão com as médias dos indivíduos para cada classe genotípica do marcador, o modelo será:

$$Y = X\beta + \epsilon$$

Tabela 25. Frequências genotípicas esperadas para uma geração F_2 , considerando-se dois locos ligados em fase de associação, com uma frequência de recombinação c

Genótipo	Frequência	Valor Genotípico	Média da classe
MQ/MQ	$\frac{1}{4}(1 - c)^2$	+a	
MQ/Mq	$\frac{c}{2}(1 - c)$	d	$\frac{1}{4}a(1 - 2c) + \frac{d}{2}c(1 - c)$
Mq/Mq	$\frac{c^2}{4}$	-a	
MQ/mQ	$\frac{c}{2}(1 - c)$	+a	
$\left\{ \begin{array}{l} \text{MQ/mq} \\ \text{Mq/mQ} \end{array} \right\} \text{MmQq}$	$\frac{1}{2}(1 - 2c + 2c^2)$	d	$\frac{d}{2}(1 - 2c + 2c^2)$
Mq/mq	$\frac{c}{2}(1 - c)$	-a	
mQ/mQ	$\frac{c^2}{4}$	+a	
mQ/mq	$\frac{c}{2}(1 - c)$	d	$-\frac{1}{4}a(1 - 2c) + \frac{d}{2}c(1 - c)$
mq/mq	$\frac{1}{4}(1 - c)^2$	-a	

c é a frequência de recombinação entre o marcador e o QTL.

$$\begin{bmatrix} y_{MM} \\ y_{Mm} \\ y_{Mm} \\ y_{mm} \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 1 & 0 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha \end{bmatrix} + \epsilon$$

Para o qual, o estimador de mínimos quadrados $\hat{\mu} = (X'X)^{-1}X'Y$

então,

$$\begin{bmatrix} \hat{\mu}_0 \\ \hat{\mu}_1 \end{bmatrix} = \begin{bmatrix} \frac{1}{4} & 0 \\ 0 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} (y_{MM} + 2y_{Mm} + y_{mm}) \\ (y_{MM} - y_{mm}) \end{bmatrix}$$

e a esperança matemática de $\hat{\mu}$ será:

$$E \begin{bmatrix} \hat{\mu}_0 \\ \hat{\mu}_1 \end{bmatrix} = E \begin{bmatrix} \frac{1}{4}(y_{MM} + 2y_{Mm} + y_{mm}) \\ \frac{1}{2}(y_{MM} - y_{mm}) \end{bmatrix}$$

$$E \begin{bmatrix} \hat{\alpha}_0 \\ \hat{\alpha}_1 \end{bmatrix} = \begin{bmatrix} \frac{1}{4}[d(1 - c + c^2)] \\ \frac{1}{4}a(1 - 2c) \end{bmatrix}$$

assim testar a hipótese:

$$\begin{cases} H_0 : \alpha_1 = 0 & \text{vs} \\ H_a : \alpha_1 \neq 0 \end{cases}$$

equivale a testar a hipótese de que α seja nulo, ou $c = 0,5$.

Genotipagem de F_2 e avaliação fenotípica de S_n

Esquema do delineamento:

Neste delineamento, os indivíduos F_2 são genotipados e, então, autofecundados para obtenção da geração S_1 ou $F_{3:2}$, que será avaliada quanto ao caráter que se deseja mapear, ou autofecundada novamente para obtenção da geração S_2 ($F_{4:2}$ que poderá ser avaliada ou autofecundada novamente, e assim sucessivamente até a geração S_n . Este delineamento é empregado para aumentar a disponibilidade de sementes do material a ser avaliado, o que permite uma avaliação com um maior número de repetições e em mais de um local. Na tabela 26 estão apresentadas as

freqüências genotípicas esperadas, quanto ao QTL, na geração S_1 em relação ao genótipo, quanto ao marcador molecular, na geração F_2 .

Tabela 26. Frequências genotípicas esperadas quanto ao QTL, em uma geração S_1 , dado o genótipo quanto ao marcador molecular na geração F_2 considerando-se marcador e QTL ligados em fase de associação, com uma frequência de recombinação c

F_2		S_1		média da classe
Genótipo	Frequência	Genótipo	Frequência	
MQ/MQ	$\frac{1}{4}(1 - c)^2$	- _QQ	$\frac{1}{4}(1 - c)^2$	$\frac{1}{4}a(1 - 2c) + \frac{d}{4}c(1 - c)$
MQ/Mq	$\frac{c}{2}(1 - c)$	- _QQ	$\frac{c}{8}(1 - c)$	
		- _Qq	$\frac{c}{4}(1 - c)$	
		- _qq	$\frac{c}{8}(1 - c)$	
Mq/Mq	$\frac{c^2}{4}$	- _qq	$\frac{c^2}{4}$	
MQ/mQ	$\frac{c}{2}(1 - c)$	- _QQ	$\frac{c}{2}(1 - c)$	$\frac{d}{4}(1 - 2c + 2c^2)$
		- _QQ	$\frac{1}{8}(1 - 2c + 2c^2)$	
MQ/mq } Mq/mQ }	$\frac{1}{2}(1 - 2c + 2c^2)$	- _Qq	$\frac{1}{4}(1 - 2c + 2c^2)$	
		- _qq	$\frac{1}{8}(1 - 2c + 2c^2)$	
Mq/mq	$\frac{c}{2}(1 - c)$	- _qq	$\frac{c}{2}(1 - c)$	
mQ/mQ	$\frac{c^2}{4}$	- _QQ	$\frac{c^2}{4}$	$-\frac{1}{4}a(1 - 2c) + \frac{d}{4}c(1 - c)$
mQ/mq	$\frac{c}{2}(1 - c)$	- _QQ	$\frac{c}{8}(1 - c)$	
		- _Qq	$\frac{c}{4}(1 - c)$	
		- _qq	$\frac{c}{8}(1 - c)$	
mq/mq	$\frac{1}{4}(1 - c)^2$	- _qq	$\frac{1}{4}(1 - c)^2$	

c é a frequência de recombinação entre o marcador e o QTL.

Como neste tipo de delineamento apenas os indivíduos F_2 são genotipados, tem-se que, assim como para uma geração F_2 o estimador de mínimos quadrados

Tabela 27. Valores genotípicos médios esperados para cada classe de marcador da geração F_2 , nas gerações S_1 , S_2 e S_n

Gen.	valor genotípico médio			
F_2	S_1	S_2	$\dots$	S_n
MM	$\frac{1}{4}a(1 - 2c) + \frac{d}{4}c(1 - c)$	$\frac{1}{4}a(1 - 2c) + \frac{d}{8}c(1 - c)$	$\dots$	$\frac{1}{4}a(1 - 2c) + \frac{d}{2^{n+1}}c(1 - c)$
Mm	$\frac{d}{4}(1 - 2c + 2c^2)$	$\frac{d}{8}(1 - 2c + 2c^2)$	$\dots$	$\frac{d}{2^{n+1}}(1 - 2c + 2c^2)$
mm	$-\frac{1}{4}a(1 - 2c) + \frac{d}{4}c(1 - c)$	$-\frac{1}{4}a(1 - 2c) + \frac{d}{8}c(1 - c)$	$\dots$	$-\frac{1}{4}a(1 - 2c) + \frac{d}{2^{n+1}}c(1 - c)$

c é a frequência de recombinação entre o marcador e o QTL.

de $\bar{y}_1$ será:

$$\hat{a}_1 = \frac{1}{2}(\bar{y}_{MM} - \bar{y}_{mm})$$

, conseqüentemente,

$$E[\hat{a}_1] = \frac{1}{4}a(1 - 2c)$$

Então, testar-se a hipótese:

$$\begin{cases} H_0 : \bar{y}_1 = 0 & \text{vs} \\ H_a : \bar{y}_1 \neq 0 \end{cases}$$

, utilizando os dados fenotípicos de uma geração S_1 equivale a testar a hipótese de que $a = 0$ ou $c = 0,5$, assim como utilizando os dados de uma geração F_2 , o mesmo ocorrendo ao se utilizar gerações mais avançadas, como pode ser verificado na tabela 27 na qual estão apresentadas os valores médios esperados para cada classe de marcadores da geração F_2 , quando o caráter quantitativo for avaliado nas gerações S_1 , S_2 e S_n .

Genotipagem de F_2 e avaliação em Top-cross

Nesse caso as plantas F_2 serão cruzadas com um testador, que pode ser uma linhagem, diferente das parentais, um híbrido simples ou uma

população qualquer. Considerando-se que apenas as plantas F_2 serão genotipadas, desconhece-se o genótipo do testador, tanto quanto aos marcadores utilizados, quanto ao QTL. Sem perda de generalidade consideramos que o testador possui o mesmo QTL que as linhagens parentais, com frequências alélicas p para o alelo favorável (Q), e $1-p$ para o alelo desfavorável (q). Deste modo, obtêm-se as frequências esperadas para a população resultante do Top cross, apresentadas na tabela 28

Esquema do delineamento:

Dado que a informação genotípica disponível refere-se apenas aos marcadores moleculares, nas plantas f_2 , então o estimador de τ_1 será:

$$\hat{\tau}_1 = \frac{1}{2}(\bar{y}_{MM} - \bar{y}_{mm})$$

então,

$$\begin{aligned} E[\hat{\tau}_1] &= E\left[\frac{1}{2}\bar{y}_{MM} - \bar{y}_{mm}\right] \\ &= \frac{1}{8}a(1 - 2c) + \frac{1}{8}d(1 - 2c) - \frac{1}{4}d(p - 2pc) \end{aligned}$$

Verifica-se deste modo que a interpretação genética do parâmetro τ_1 estimado, utilizando os dados de uma população top cross depende do tipo do testador utilizado, ou seja da frequência alélica do QTL no testador. Assim, se o testador for, por exemplo, a linhagem L_2 para a qual $p = 0$ tem-se: $E\tau_1 = \frac{1}{8}(a + d)(1 - 2c)$, neste caso o teste da hipótese:

$$\begin{cases} H_0 : \tau_1 = 0 & \text{vs} \\ H_a : \tau_1 \neq 0 \end{cases}$$

Tabela 28. Frequências genotípicas esperadas para uma geração Top-cross, considerando-se dois locos ligados em fase de associação, com uma frequência de recombinação c

F ₂		Top Cross		média da classe
Genótipo	Frequência	Genótipo	Frequência	
MQ/MQ	$\frac{1}{4}(1 - c)^2$	- _QQ	$\frac{1}{4}(1 - c)^2p$	$\frac{1}{4}a(p - c) + \frac{d}{4}[1 - c - p(1 + 2c)]$
		- _Qq	$\frac{1}{4}(1 - c)^2(1 - p)$	
MQ/Mq	$\frac{c}{2}(1 - c)$	- _QQ	$\frac{c}{4}(1 - c)p$	
		- _Qq	$\frac{c}{4}(1 - c)$	
Mq/Mq	$\frac{c^2}{4}$	- _qq	$\frac{c}{4}(1 - c)(1 - p)$	
		- _Qq	$\frac{c^2}{4}p$	
		- _qq	$\frac{c^2}{4}(1 - p)$	
MQ/mQ	$\frac{c}{2}(1 - c)$	- _QQ	$\frac{c}{2}(1 - c)p$	
MmQq	$\frac{1}{2}(1 - 2c + 2c^2)$	- _Qq	$\frac{c}{2}(1 - c)(1 - p)$	
		- _QQ	$\frac{1}{4}(1 - 2c + 2c^2)p$	
		- _Qq	$\frac{1}{4}(1 - 2c + 2c^2)$	
Mq/mq	$\frac{c}{2}(1 - c)$	- _qq	$\frac{1}{4}(1 - 2c + 2c^2)(1 - p)$	
		- _Qq	$\frac{c}{2}(1 - c)p$	
		- _qq	$\frac{c}{2}(1 - c)(1 - p)$	
mQ/mQ	$\frac{c^2}{4}$	- _QQ	$\frac{c^2}{4}p$	$\frac{1}{4}a(p - 1 + c) + \frac{d}{4}[c + p(1 - 2c)]$
mQ/mq	$\frac{c}{2}(1 - c)$	- _Qq	$\frac{c^2}{4}(1 - p)$	
		- _QQ	$\frac{c}{4}(1 - c)p$	
		- _Qq	$\frac{c}{4}(1 - c)$	
mq/mq	$\frac{1}{4}(1 - c)^2$	- _qq	$\frac{c}{4}(1 - c)(1 - p)$	
		- _Qq	$\frac{1}{4}(1 - c)^2p$	
		- _qq	$\frac{1}{4}(1 - c)^2(1 - p)$	

c é a frequência de recombinação entre o marcador e o QTL.

, equivale a testar de $(a+d) = 0$ ou $c = 0$; 5, assim como no caso de se utilizar uma população RC₂. Se o testador for a linhagem L₁, para qual $p = 1$; 0 virá:
 $E^{-1} = \frac{1}{8}(a-d)(1-2c)$, e o teste da hipótese:

$$\begin{cases} H_0 : \tau_1 = 0 & \text{vs} \\ H_a : \tau_1 \neq 0 \end{cases}$$

, equivale-se a testar $(a - d) = 0$ ou $c = 0$; 5, como obtido para uma população RC₂. Já no caso de se utilizar como testador uma população F₁ ou híbrido simples entre L₁ e L₂ tem-se: $E[\bar{r}_1] = \frac{1}{8}(a)(1 - 2c)$, e então, testar a hipótese:

$$\begin{cases} H_0 : \bar{r}_1 = 0 \\ H_a : \bar{r}_1 \neq 0 \end{cases} \quad \text{vs}$$

, equivale a testar de $a = 0$ ou $c = 0$; 5, assim como no caso de se utilizar uma população F₂.

No caso de realizar-se o Top-cross utilizando famílias S₁, no lugar de plantas F₂, as frequências esperadas para os possíveis genótipos desta população, bem como os valores genotípicos médios esperados para cada classe de marcadores da geração F₂ estão apresentadas na tabela 29. Na qual verifica-se que os valores esperados para as classes de marcadores na geração F₂ e portanto os de $\bar{r}_1$ são iguais aos obtidos ao se realizar o top cross com as plantas da geração F₂.

Considerando-se o efeito de dominância (d) sempre maior ou igual a zero ($d \geq 0$) isto é, quando existir dominância ela seja do alelo favorável ao caráter, verifica-se na expressão $E[\bar{r}_1] = \frac{1}{8}a(1 - 2c) + \frac{1}{8}d(1 - 2c) - \frac{1}{4}d(p - 2pc)$ que o valor de $E[\bar{r}_1]$ será, para uma dada relação $a=d$ tanto maior quanto menor for p. Do ponto de vista prático, este fato indica que quanto pior for a linhagem testadora, maior será sua contribuição para o mapeamento de QTLs, pois neste caso, haverá um maior poder de rejeição da hipótese $H_0 : \bar{r}_1 = 0$.

Um resumo das esperanças matemáticas do coeficiente $\bar{r}_1$, em função do tipo de geração segregante utilizada, é apresentado na tabela 30.

Tabela 29. Frequências esperadas para uma geração top cross de S_1

F_2	S_1		Top Cross		média da classe
	genótipo	frequência	genótipo	frequência	
MM	-- QQ	$\frac{1}{8}(2 - 3c + c^2)$	-- QQ	$\frac{1}{8}(2 - 3c + c^2)p$	$\frac{a}{4}(p - c) + \frac{d}{4}[1 - c - p(1 + 2c)]$
			-- Qq	$\frac{1}{8}(2 - 3c + c^2)(1 - p)$	
	-- Qq	$\frac{1}{4}c(1 - c)$	-- QQ	$\frac{1}{8}c(1 - c)p$	
			-- Qq	$\frac{1}{8}c(1 - c)$	
			-- qq	$\frac{1}{8}c(1 - c)(1 - p)$	
	-- qq	$\frac{1}{8}(c + c^2)$	-- Qq	$\frac{1}{8}(c + c^2)p$	
			-- qq	$\frac{1}{8}(c + c^2)(1 - p)$	
Mm	-- QQ	$\frac{1}{8}(1 + 2c - 2c^2)$	-- QQ	$\frac{1}{8}(1 + 2c - 2c^2)p$	$\frac{d}{4} + \frac{a}{4}(2p - 1)$
			-- Qq	$\frac{1}{8}(1 + 2c - 2c^2)(1 - p)$	
	-- Qq	$\frac{1}{4}(1 - 2c + 2c^2)$	-- QQ	$\frac{1}{8}(1 - 2c + 2c^2)p$	
			-- Qq	$\frac{1}{8}(1 - 2c + 2c^2)$	
			-- qq	$\frac{1}{8}(1 - 2c + 2c^2)(1 - p)$	
	-- qq	$\frac{1}{8}(1 + 2c - 2c^2)$	-- Qq	$\frac{1}{8}(1 + 2c - 2c^2)p$	
			-- qq	$\frac{1}{8}(1 + 2c - 2c^2)(1 - p)$	
mm	-- QQ	$\frac{1}{8}(c + c^2)$	-- QQ	$\frac{1}{8}(c + c^2)p$	$\frac{a}{4}(p - 1 + c) + \frac{d}{4}[c + p(1 - 2c)]$
			-- Qq	$\frac{1}{8}(c + c^2)(1 - p)$	
	-- Qq	$\frac{1}{4}c(1 - c)$	-- QQ	$\frac{1}{8}c(1 - c)p$	
			-- Qq	$\frac{1}{8}c(1 - c)$	
			-- qq	$\frac{1}{8}c(1 - c)(1 - p)$	
	-- qq	$\frac{1}{8}(2 - 3c + c^2)$	-- Qq	$\frac{1}{8}(2 - 3c + c^2)p$	
			-- qq	$\frac{1}{8}(2 - 3c + c^2)(1 - p)$	

c é a frequência de recombinação entre o marcador e o QTL.

Tabela 30. Esperanças matemáticas do coeficiente β_1 num modelo de regressão linear simples para análise de QTLs utilizando a metodologia de análise de marca simples, em função do tipo de população segregante utilizada.

População	$E[\beta_1]$
RC_1	$\frac{1}{2}(a - d)(1 - 2c)$
RC_2	$\frac{1}{2}(a + d)(1 - 2c)$
F_2	$\frac{1}{4}a(1 - 2c)$
S_n	$\frac{1}{4}a(1 - 2c)$
Top cross	$\frac{1}{8}a(1 - 2c) + \frac{1}{8}d(1 - 2c) - \frac{1}{4}d(p - 2pc)$

a= efeito aditivo do QTL; d= desvio de dominância;
c=frequência de recombinação entre o QTL e o marcador utilizado e
p= frequência alélica do alelo favorável (Q) na população testadora.