Modelos de Regressão no Mapeamento Genético em Experimentos com Cruzamentos Controlados

Paula Mitiko Yamakawa

DISSERTAÇÃO APRESENTADA
AO
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
DA
UNIVERSIDADE DE SÃO PAULO
PARA
OBTENÇÃO DO TÍTULO DE MESTRE
EM CIÊNCIAS

Área de Concentração: Estatística
Orientadora: Profa. Dra. Júlia Maria Pavan Soler

São Paulo, julho de 2006

Modelos de Regressão no Mapeamento Genético em Experimentos com Cruzamentos Controlados

Este exemplar corresponde à redação final da dissertação devidamente corrigida e defendida por Paula Mitiko Yamakawa e aprovada pela Comissão Julgadora.

São Paulo, 23 de julho de 2006.

Banca Examinadora:

Profa. Dra. Júlia Maria Pavan Soler (orientadora) – IME/USP

Profa. Dra. Mônica Carneiro Sandoval – IME/USP

Profa. Dra. Roseli Aparecida Leandro – ESALQ/USP

Modelos de Regressão no Mapeamento Genético em Experimentos com Cruzamentos Controlados

PAULA MITIKO YAMAKAWA

Orientadora:

Profa. Dra. Júlia Maria Pavan Soler

Dissertação apresentada ao Instituto de Matemática e Estatística da Universidade de São Paulo, como parte dos requisitos para obtenção do título de Mestre em Ciências na área de Estatística.

USP - UNIVERSIDADE DE SÃO PAULO
São Paulo, julho de 2006.

Aos meus pais Carlos e Mitiko.

Agradecimentos

Agradeço a todos que contribuíram para a elaboração deste trabalho, em especial:
À professora **Júlia Maria Pavan Soler**, pela orientação, confiança e incentivo.

À minha querida amiga **Tatiana Terabayashi**, pelo inestimável apoio.

Ao **InCor**, por conceder o banco de dados para análise.

Resumo

O mapeamento genético pode ser definido como uma série de procedimentos experimentais e estatísticos que buscam detectar genes associados à etiologia e regulação de traços quantitativos, além de estimar os efeitos genéticos e as localizações genômicas correspondentes. Neste trabalho, apresentamos diferentes formulações de modelos de regressão úteis na identificação de QTL's (*quantitative trait loci*) considerando delineamentos experimentais que envolvem cruzamentos controlados de animais ou plantas. Com base em informações obtidas por meio dos mapas de marcadores moleculares, pode-se ajustar desde modelos mais simples; como é o caso daqueles utilizados na busca por evidências da presença de um único QTL, até modelos mais complexos; como os que permitem a busca por múltiplos QTL's e seus possíveis efeitos de interação (epistasia), e aqueles modelos com efeito de pleiotropia para os casos em que um mesmo conjunto de genes está associado a mais de um fenótipo. A aplicação destes modelos é considerada por meio da análise de um conjunto de dados genotípicos e fenotípicos de ratos provenientes de um delineamento F2, coletados no Laboratório de Genética e Cardiologia Molecular do Instituto do Coração de São Paulo (InCor), com o objetivo de identificar genes relacionados a doenças cardiovasculares.

Abstract

QTL mapping can be defined as a series of experiment and statistical procedures used to identify genes associated to quantitative traits etiology and regulation. Estimation of QTL's (quantitative trait loci) position and effects are considered through this metodology. In this work, we introduce different formulations of regression models useful to identification of QTL's considering experimental designs that envolves inbred crosses of animals and plants. Based on informations obtained from molecular markers maps, we are able to adjust since simple models, as those used on the search for evidences of only one QTL, until complex models, such as those which allows the search for multiple QTL's and their possible interaction effects (epistasis), and pleiotropic models that consider one common group of genes that underlies more than one phenotype. The application of these models is considered by the analysis of genotypes and fenotypes data sets of rats that came from a F2 progeny collected at the São Paulo Heart Institute to identify genes related with blood pressure regulation.

Sumário

Lista de Tabelas

Lista de Figuras

Capítulo 1

Introdução

1.1 A Estatística e a Genética

Métodos estatísticos são ferramentas poderosas que auxiliam de forma decisiva as ciências factuais, entre elas um ramo da Biologia que estuda as leis de transmissão dos caracteres hereditários em indivíduos: a Genética.

O interesse pela Genética vem crescendo vertiginosamente em função de que quanto maior for a compreensão e o domínio sobre estas leis de segregação dos genes, maior será o controle sobre várias doenças, bem como de várias características de importância econômica em humanos, plantas e animais.

A fim de facilitar o entendimento dos mecanismos envolvidos na transmissão de caracteres hereditários, vale relembrar algumas definições como a de cromossomos homólogos, que são aqueles que possuem a mesma morfologia e que contêm o mesmo conjunto de locos gênicos, lembrando que cada par de cromossomos homólogos é formado por um ramo de origem paterna e outro de origem materna [Farah, 1997]. Outra definição importante é a de alelos, que são formas alternativas de um mesmo gene que ocupam a mesma posição relativa (loco) em cromossomos homólogos. Confira a Figura 1.1 para ilustração destes conceitos.

Sabe-se que fenótipo é o conjunto de características que um indivíduo possui e, em geral, ele é o resultado de interações entre o genótipo (constituição genética do indivíduo) e o ambiente [Silva Jr & Sasson, 1990]. O fenótipo não corresponde necessariamente a uma característica morfológica; ele pode indicar uma característica

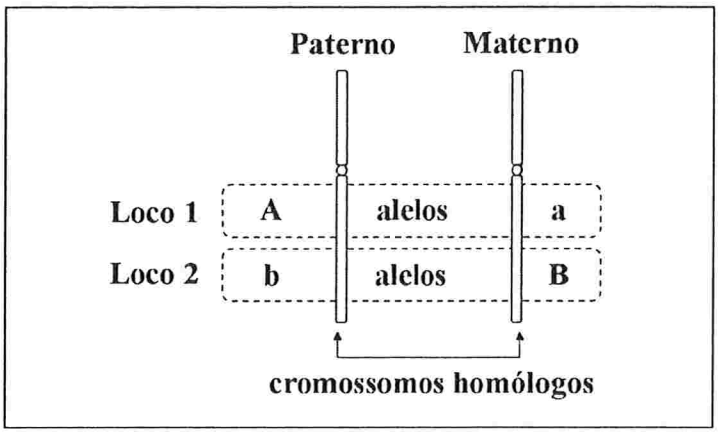


Figura 1.1: Dois cromossomos homólogos com dois pares de locos dialélicos.

fisiológica. Como exemplo de uma característica de forma, pode-se citar o tamanho de um osso ou o número de vértebras de um indivíduo, e como exemplo de uma característica de função, pode-se falar em substâncias presentes na corrente sanguínea como o nível de glicose, colesterol e hormonal.

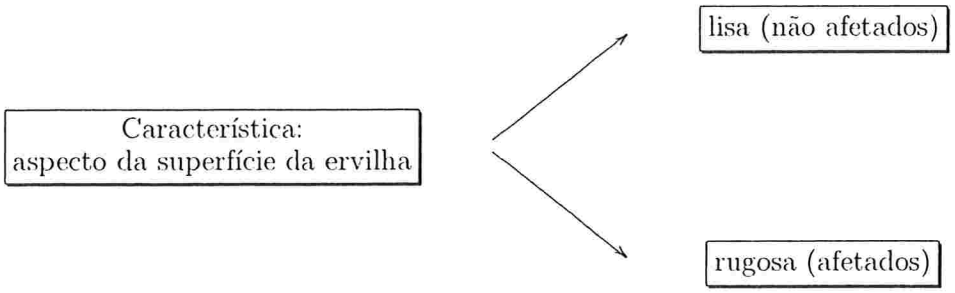
De maneira geral, quando não existe efeito de ambiente sobre uma determinada característica, o fenótipo apresentado é idêntico para todos os indivíduos que possuem o mesmo genótipo. Pode-se estabelecer o seguinte modelo [Balding *et al.*, 2003]:

$$\textbf{Fenótipo} = \textbf{Genótipo} + \textbf{Ambiente} \tag{1.1}$$

Os fenótipos, definidos como variáveis estatísticas, podem ser classificados como qualitativos ou quantitativos.

1.2 A Genética Qualitativa

A Genética Mendeliana estuda basicamente os fenótipos qualitativos. Gregor Mendel (1822-1884) conduziu algumas experiências de cruzamento entre plantas caracterizando a variável resposta de forma categorizada: afetado/não afetado e considerando apenas uma característica por vez. Para ilustrar, tomemos o seguinte exemplo esquemático:



Mendel buscou linhagens puras quanto à característica de interesse, deixando as plantas se autofecundarem durante algumas gerações até observar que estas apresentassem sempre a mesma característica, no caso, apenas o aspecto liso ou o rugoso. De posse destas linhagens puras, ele promoveu o cruzamento das plantas puras de sementes lisas com plantas puras de sementes rugosas, determinando assim a geração parental.

Além disso, Mendel postulava que "cada característica genética de um organismo é condicionada por dois genes, um proveniente do pai e outro da mãe" e que "quando o indivíduo fosse reproduzir-se, apenas um gene do par seria transmitido da célula sexual". Estas duas idéias formam a essência da chamada **Primeira Lei de Mendel** [Silva Jr & Sasson, 1990].

Retomando o exemplo anterior e sabendo que dois genes que se localizam na mesma região de um par de cromossomos homólogos (loco genético) definem um genótipo por meio da combinação dos alelos que compõem o loco, vejamos a Figura 1.2.

Considere que o gene sob estudo tem dois alelos na população: A e a. Se um determinado indivíduo tem duas cópias de A, então seu genótipo é AA e ele é homozigoto (para o alelo A). Um indivíduo com o genótipo Aa é heterozigoto e um indíviduo com genótipo aa é outro homozigoto (para o alelo a).

Se cada um dos três genótipos (AA, Aa, aa) determinarem três fenótipos distintos, então os alelos são ditos codominantes. Entretanto, se os indivíduos com genótipos AA e Aa expressarem o mesmo fenótipo, então A é definido como o alelo dominante e a é o alelo recessivo (caso da superfície da ervilha).

Quando a probabilidade de um indivíduo ser afetado dado que ele possui um determinado genótipo é 1, temos que o efeito do gene é completamente efetivo e tal

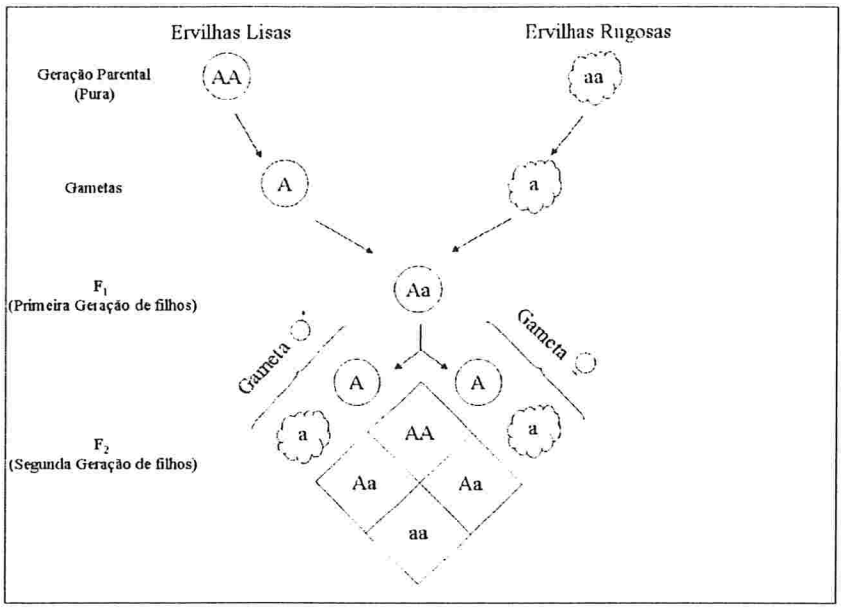


Figura 1.2: Cruzamento Mendeliano entre indivíduos F1 heterozigotos.

situação determina um modelo monogênico de herança (controlado por um único gene) e de penetrância completa. No entanto, existem casos em que o indivíduo carrega um gene para uma doença autossômica (relativa à qualquer cromossomo que não seja os sexuais) dominante e não manifesta os sinais clínicos desta doença. Nestes casos, dizemos que não houve penetrância do gene [Farah, 1997], ou que ela é incompleta, muitas vezes dependente de um outro fator, como a idade por exemplo.

De qualquer forma, vale salientar que fenótipos qualitativos como, por exemplo, ter ou não ter uma síndrome, são controlados por 1 único gene e não há efeito do ambiente.

1.3 A Genética Quantitativa

A genética quantitativa, por sua vez, foca a herdabilidade de traços (fenótipos) quantitativos. Quando o número de genes que controlam o traço aumenta e a influência de fatores ambientais no fenótipo também aumenta, o poder de modelos Mendelianos de herança (para dados categorizados) deixa de ser satisfatório.

Variações na estatura, na cor da pele, na produção de leite e carne do gado, no peso e no tamanho de frutos são exemplos de caracteres que dizemos variar quantitativamente.

Diferentemente dos traços Mendelianos que são categóricos na sua natureza, os caracteres de variabilidade contínua se manifestam de acordo com a ação conjunta de vários genes e do ambiente. Considere um modelo simples, em que cada gene teria um efeito individual sobre o fenótipo, que se somaria ao efeito de outros genes. Considere ainda que o efeito de cada gene seria aditivo e não haveria entre os alelos de um par de locos cromossômicos qualquer relação.

Como um exemplo simplificado de herança quantitativa, podemos citar a hipótese de Davenport, apresentada em 1913, para explicar a transmissão da cor da pele [Silva Jr & Sasson, 1990]. Foi suposto que a quantidade de melanina da pele fosse determinada por dois pares de genes aditivos. Os dois fenótipos extremos, branco e negro, seriam, respectivamente, aabb e AABB. Teríamos 5 classes fenotípicas, determinadas pelos seguintes genótipos apresentados na Tabela 1.1.

GENÓTIPO	FENÓTIPO
AABB	Negro
AABb AaBB	Mulato Escuro
AAbb AaBb aaBB	Mulato Médio
Aabb aaBb	Mulato Claro
aabb	Branco

Tabela 1.1: Classes Genotípicas e Fenotípicas da cor da pele.

Note que a quantidade de genes A ou B nos dois locos genéticos é que determina a "intensidade" do fenótipo, neste caso, a cor da pele. Vale ressaltar também que sob o modelo de Davenport, um fenótipo quantitativo, cor da pele, é denotado de forma simplificada como qualitativo (categórico).

É provável que a hipótese de Davenport represente uma simplificação do modo de regulação gênica, pois existem, na realidade, variações maiores (e contínuas) na

cor de pele do que as propostas pelo autor. Ainda assim, mesmo que um número maior que dois pares de genes estejam envolvidos no modelo, que a existência de efeito de fatores ambientais deva ser levada em conta ou que o modelo aditivo não seja válido, sua hipótese de herança quantitativa controlada por mais de um gene é correta.

Em linhas gerais, tanto na abordagem qualitativa quanto na quantitativa, o objetivo principal dos modelos genéticos é identificar genes que estejam relacionados com determinados fenótipos. Como ilustração, podemos citar o grande interesse em determinar quais genes estão relacionados a certas doenças humanas ou quais deles controlam traços de importância econômica em plantas e animais [Liu, 1998].

A partir do momento em que existe o interesse em determinar quanto da variação de um traço quantitativo se deve a um gene específico e quanto se deve ao ambiente, o mapeamento genético do traço se faz necessário.

1.4 Mapeamento Genético e QTL's

O mapa genético de espécies de plantas ou animais é um modelo abstrato da disposição linear de um conjunto de genes e de marcadores moleculares [Liu, 1998]. Marcadores moleculares são caracteres de herança ou locos cromossômicos facilmente identificáveis e, por esta razão, são utilizados para a caracterização de um indivíduo. Tais locos podem ser genes de função conhecida ou até mesmo fragmentos de DNA com função desconhecida.

Nos casos em que há interesse na identificação de genes que estão envolvidos na regulação de uma característica fenotípica quantitativa, o conceito de QTL's *(Quantitative Trait Loci)* é introduzido, isto é, QTL's são regiões cromossômicas candidatas a conterem genes reguladores de traços quantitativos de interesse. O mapeamento genético associado é denominado mapeamento de QTL's e é definido por um conjunto de procedimentos experimentais e estatísticos que buscam detectar genes associados à etiologia e regulação de traços quantitativos e estimar os efeitos genéticos e localizações genômicas correspondentes.

É importante destacar que o efeito genético de um gene (de um marcador, de

um QTL) pode ser decomposto em dois componentes: **efeito genético aditivo** e **efeito de dominância** [Falconer, 1964]. Retomando a equação 1.1, tem-se:

Fenótipo=Genótipo+Ambiente

Fenótipo= (Efeito Aditivo+Efeito de Dominância)+Efeito de Ambiente (1.2)

O efeito genético aditivo é o valor fenotípico que pode ser predito linearmente por meio do número de alelos de um certo tipo que definem o genótipo, enquanto que o efeito de dominância é o valor fenotípico que não pode ser explicado linearmente (resíduo genético do modelo). Veja Figura 1.3 [Lynch & Walsh, 1998].

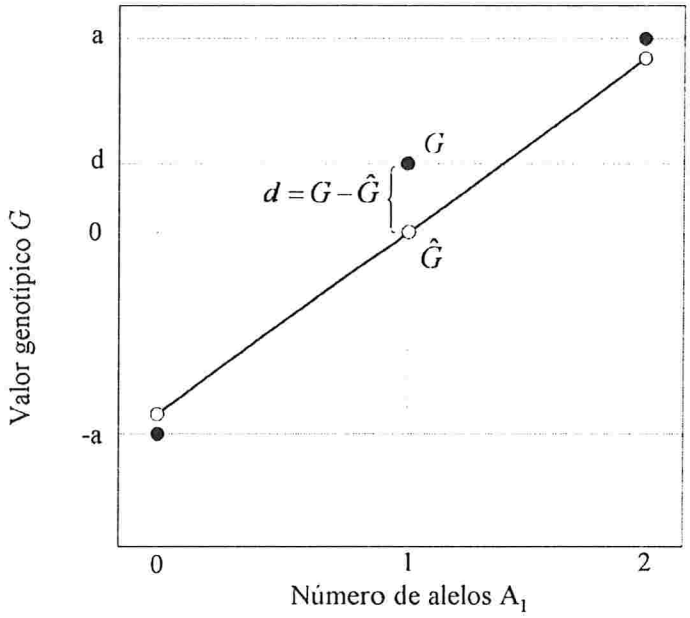


Figura 1.3: Decomposição dos efeitos genéticos. Da esquerda para a direita, os pontos da abscissa representam os genótipos A2A2, A1A2 e A1A1, respectivamente. Os círculos pretos representam os verdadeiros valores genotípicos, enquanto que os círculos brancos são os valores esperados com base no efeito aditivo $\hat{G}$. O desvio entre G e $\hat{G}$ para cada genótipo é chamado de desvio de dominância (d).

O valor genotípico de um indivíduo, em geral, não é mensurável na prática. Exceto quando estamos interessados em um loco onde os genótipos são fenotipicamente

distinguíveis.

Como apresentado na Figura 1.3, quando apenas um loco genético está sendo considerado, a diferença entre o valor genotípico, G, e o efeito aditivo, $\hat{G}$, de um genótipo em particular é conhecida como desvio de dominância d, então:

$$\mathbf{Y = G + E}$$

onde, $\mathbf{G = \hat{G} + d}$;
Y é o valor fenotípico;
G é o valor genotípico;
E é o efeito ambiental;
$\hat{G}$ é o efeito aditivo do gene;
d é o efeito de dominância do gene.

Vejamos o esquema representado pela Figura 1.4.

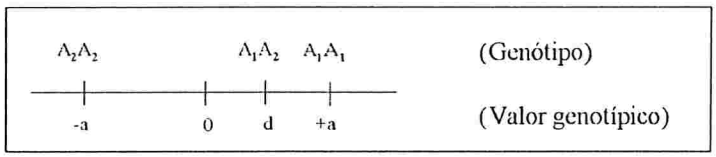


Figura 1.4: Valores genotípicos.

A Figura 1.4 ilustra valores genotípicos centrados no valor 0 e determinados arbitrariamente, pois d depende do grau de dominância do heterozigoto, tal que:

- Se $d = 0$, então não há dominância entre os alelos;
- Se $d > 0$, então A1 é dominante sobre A2;
- Se $d < 0$, então A2 é dominante sobre A1.

A Figura 1.5 apresenta os 3 modelos de herança genética mais comuns: o **codominante**, que assume um efeito linear de acordo com o número de alelos de um tipo específico (no caso, 0, 1 ou 2 alelos do tipo A1); o **dominante**, que apresenta fuga da linearidade, bastando a presença de um alelo A1 para a elevação da média do

fenótipo y e, conseqüentemente, a manifestação do traço; e o **recessivo**, que também apresenta fuga da linearidade, pois o traço se manifesta somente no homozigoto A1A1.

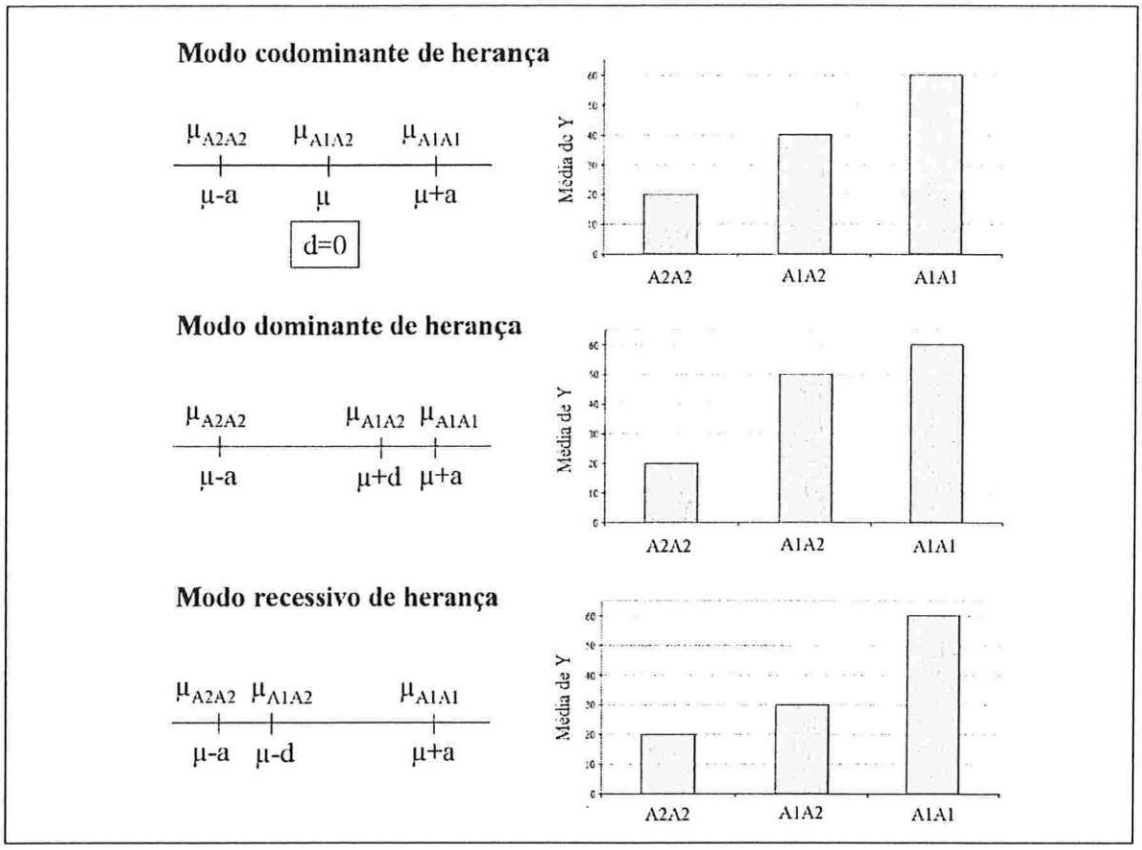


Figura 1.5: Modos de herança.

Do ponto de vista estatístico, os efeitos de dominância podem ser entendidos como efeitos de interações entre os alelos dentro de um loco [Falconer, 1964]. Ao considerarmos o gene como um fator em 3 níveis (por exemplo, A1A1, A1A2 e A2A2), teremos 2 graus de liberdade para estudar os efeitos deste fator, os quais estão ortogonalmente particionados no componente aditivo e no componente de dominância [Kempthorne, 1957].

Deste modo, se um conjunto de indivíduos é genotipado para um certo loco cromossômico (por exemplo, como A1A1, A1A2 e A2A2) pode-se estimar os efeitos aditivo e de dominância deste loco. No caso de mapeamento genético de um fenótipo

qualitativo, imagine que um conjunto de indivíduos é genotipado para um mapa de marcadores, isto é, para muitos locos cromossômicos. Neste caso, pode-se estimar diretamente o efeito (aditivo e de dominância) dos marcadores genotipados, como veremos no Capítulo 3. Ainda mais, imagine que deseja-se estimar o efeito de um gene (QTL) que não se consegue genotipar, pois nem mesmo sua localização é conhecida. Como então estima-se seu efeito? Uma alternativa é estimar os efeitos de QTL's fixados em posições intermediárias entre os marcadores genotipados, utilizando a informação do mapa de marcadores. Para tanto, é necessário definir um planejamento experimental de coleta de dados adequado, que favoreça o estudo destes efeitos. Delineamentos com cruzamentos controlados de plantas e animais têm sido muito utilizados para esta finalidade.

1.5 A Utilização de Modelos Animais

Por volta de 1865, Claude Bernard lançou os princípios do uso de animais como modelo de estudo e transposição para a fisiologia humana. Atualmente, o modelo animal é usado em todos os campos da pesquisa biológica, pois a indução dos resultados com animais para a espécie humana tem critérios claros e objetivos a serem preenchidos, especialmente na área da Saúde [Fagundes & Taha, 2004]. É importante comentar que o manejo de qualquer espécie viva, animal ou vegetal, deve seguir regras e conduta ética estabelecida, o que tem sido atualmente normalizado nas pesquisas experimentais por meio de conselhos de ética regulamentados.

Sabe-se que um modelo deve ter características suficientes para ser semelhante ao objeto imitado e ter a suficiente capacidade de ser manejado sem as limitações do objeto imitado. Levando isto em conta, temos que um modelo animal deverá atender aos seguintes pressupostos:

- que permita o estudo dos fenômenos biológicos ou de comportamento do animal;
- que um processo patológico espontâneo ou induzido possa ser investigado;
- que o fenômeno, em um ou mais aspectos, seja semelhante ao fenômeno em seres humanos.

O uso de modelos de doença animal proporciona a investigação de uma relação causal de modo mais rápido, menos trabalhoso e menos oneroso, comparado com a experimentação em seres humanos. Logo, estes aspectos tornam tais modelos bastante interessantes na pesquisa dos motivos causais de doenças em humanos.

Como exemplo, podemos citar a doença da hiperplasia benigna da próstata em cães que é utilizada com freqüência como modelo para o estudo do tratamento operatório da próstata, pela sua semelhança morfológica com a próstata humana [Fagundes & Taha, 2004].

Posteriormente, veremos com mais detalhes o caso dos ratos hipertensos que são utilizados na identificação de genes (em termos de localização e efeito) controladores da hipertensão humana.

1.6 Delineamentos com Cruzamentos Controlados

O fato dos traços quantitativos serem na maioria das vezes influenciados por muitos genes (além de fatores ambientais), faz com que seu estudo seja não trivial. Isto porque características poligênicas dificultam o isolamento e a caracterização de cada um dos fatores que controlam o traço [Schork *et al.*, 1995].

As aplicações consideradas neste trabalho são baseadas em dados reais coletados no Laboratório de Genética e Cardiologia Molecular do Instituto do Coração de São Paulo (InCor). O objetivo do estudo é a identificação de genes que estejam associados a doenças relacionadas à pressão arterial. O estudo é baseado em 221 ratos F2 resultantes do cruzamento entre uma linhagem de animais normotensos e outra de hipertensos, desenvolvidas de forma que a geração de filhos possuam características praticamente idênticas, excetuando-se apenas a pressão arterial. O planejamento experimental foi definido baseando-se em um cruzamento controlado, de acordo com a Figura 1.6.

Foram coletadas 23 variáveis fenotípicas nos ratos F2, entre elas: pressão basal, pressão após o uso de medicamentos como o Captopril, pressão sistólica antes e pós-sal, pressão diastólica antes e pós-sal e o peso. Além disso, os 221 animais foram

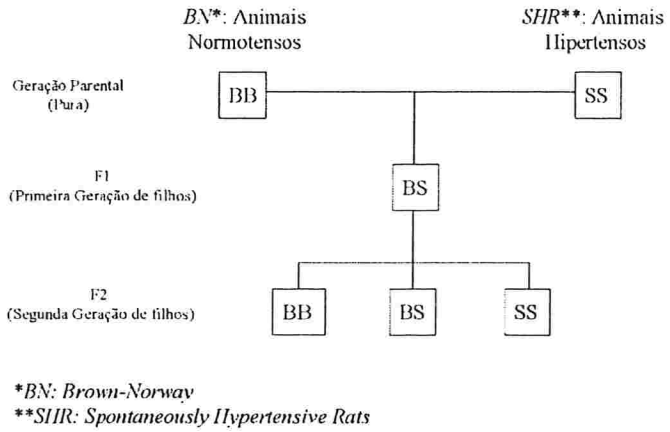


Figura 1.6: Delineamento F2 do Projeto InCor.

genotipados em 182 marcadores moleculares espalhados nos 21 cromossomos do rato. Estas informações são fundamentais na busca pelos genes que estão associados à hipertensão. Observe a Figura 1.7.

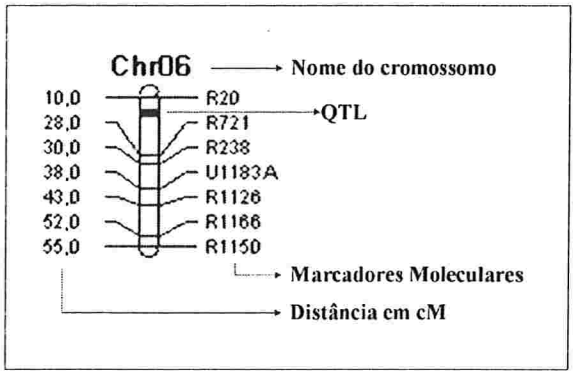


Figura 1.7: Cromossomo 6 com seus respectivos Marcadores Moleculares e um possível QTL.

A Figura 1.7 apresenta o cromossomo 6 dos ratos F2 do Projeto InCor juntamente com seus 7 marcadores moleculares e as respectivas localizações. Note que, como exemplo, há um suposto QTL entre os marcadores $R20$ e $R271$.

As populações derivadas de cruzamentos controlados são interessantes devido à flexibilidade na escolha dos pais e na determinação do esquema de cruzamento. Quando o objetivo do estudo é encontrar genes controladores de um traço em par-

ticular, as variações genéticas do traço entre os pais é crucial. Se os pais forem completamente diferentes em um nível fenotípico para o traço de interesse, existe uma chance razoável de que exista uma variação genética entre eles. Mesmo ao levarmos em consideração que efeitos do meio ambiente não-controlados podem causar uma grande variação fenotípica sem base genética alguma, ainda assim a chance de variação genética é considerável [Liu, 1998].

Assim, estas populações apresentam uma característica importante para o desenvolvimento de nossos estudos: os cruzamentos controlados podem gerar um desequilíbrio de ligação entre os genes. Tal desequilíbrio pode ser definido como uma associação na distribuição dos alelos de dois locos em uma população. Por exemplo: se o fato de haver uma base C em um determinado loco do genoma aumenta a chance de haver uma base G em outro loco do mesmo cromossomo, pode-se afirmar que existe associação alélica entre estes dois locos e eles estão em desequilíbrio de ligação [Silva & Coelho, 2005].

Portanto, fazendo uso de modelos animais, temos que a utilização de delineamentos com cruzamentos controlados são decisivamente úteis para finalidade de mapeamento genético, uma vez que pode-se controlar as linhagens a serem cruzadas de tal forma a:

- Definir uma segregação homogênea (de mesma origem) dos genes ao longo das gerações;
- Garantir que as linhagens parentais sejam praticamente idênticas, exceto nos QTL's e Marcadores de interesse;
- Efeitos dos QTL's estão fixados em cada linhagem parental com respeito a alelos alternativos, por exemplo: BB e SS;
- A herdabilidade[1] do QTL dentro das linhagens parentais é aproximadamente nula gerando uma estrutura de covariância pouco informativa (como a covariância é muito pequena, fica praticamente impossível decompô-la);

[1] A herdabilidade expressa a proporção da variância total que é atribuível aos efeitos dos genes sobre o fenótipo.

- A Geração F1 (primeira geração de descendentes) é composta exclusivamente por animais idênticos (todos heterozigotos; BS) mostrando um completo desequilíbrio de ligação nos genes que diferem entre as linhagens, isto é, os genes não segregam independentemente;
- A Geração F2 (segunda geração de descendentes) é fruto de uma distribuição aleatória do conteúdo genético das linhagens parentais e a variância fenotípica observada é devido, principalmente, aos QTL's;
- A amostra de indivíduos F2 contém todos os alelos em estudo para as populações sob cruzamento, o que pode não ocorrer em populações naturais, em que os cruzamentos são aleatórios.

Vale lembrar que os cruzamentos controlados apresentam, ainda, duas principais alternativas de delineamento experimental:

- Delineamento F2, onde um indivíduo da Geração F1 é cruzado com outro indivíduo da Geração F1. Veja Figura 1.8, que utiliza o esquema do experimento adotado no projeto com ratos hipertensos do InCor.

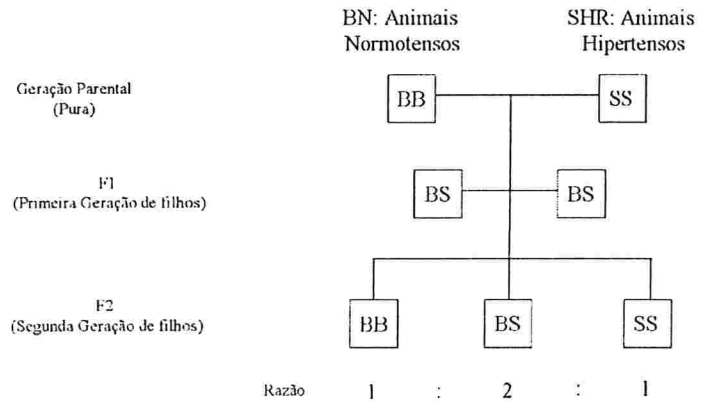


Figura 1.8: Delineamento F2.

- Delineamento *Backcross*, onde um indivíduo da Geração F1 é cruzado com uma das linhagens parentais (veja Figura 1.9).

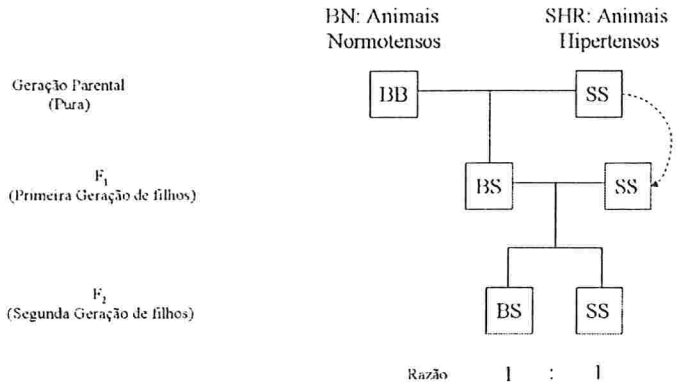


Figura 1.9: Delineamento *Backcross* (Retrocruzamento).

Como qualquer uma das linhagens parentais é homozigota, nota-se que a variabilidade genotípica obtida por meio do Delineamento *Backcross* é menor, fazendo com que o Delineamento F2 se torne mais interessante neste aspecto. Afinal, este delineamento gera 3 genótipos para cada loco de marcador, o que permite a estimação do grau de dominância associado a cada QTL detectado [Lynch & Walsh, 1998]. Maiores detalhes sobre delineamentos *Backcross* podem ser encontradas em [Liu, 1998], [Churchill & Doerge, 1994].

De qualquer forma, é importante ressaltar que a idéia-chave dos cruzamentos controlados é permitir que linhagens homozigotas com fenótipos diferentes para o traço de interesse sejam cruzadas, a fim de que originem gerações cuja variação genética esteja relacionada com o gene controlador do traço em questão.

O objetivo deste trabalho é apresentar alguns modelos estatísticos úteis na identificação de genes controladores de um traço de interesse em delineamentos experimentais envolvendo cruzamentos controlados.

No capítulo 2, apresentaremos alguns tópicos de genética que vão facilitar a compreensão do uso de determinadas técnicas estatísticas que serão aprofundadas nos capítulos posteriores. Abordaremos, inclusive, a Segunda Lei de Mendel que postula a segregação independente dos genes. Tal descoberta foi de extrema importância e discutiremos o motivo.

No capítulo 3, apresentaremos o modelo de regressão intervalar adotado no ma-

peamento de QTL's. Trataremos dos modelos Mistura de Normais e de Mínimos Quadrados, adotados no ajuste do modelo de regressão e introduziremos o gráfico de perfil da estatística *Lod score* bastante difundido na literatura da área de Genética.

No capítulo 4, apresentaremos modelos mais gerais que abordam efeitos de epistasia (interação entre genes) e efeitos de pleiotropia (incluem mais de um fenótipo).

No capítulo 5, mostraremos aplicações interessantes de alguns tópicos abordados nesta dissertação utilizando o conjunto de dados reais do InCor.

Capítulo 2

Genética

Neste capítulo vamos introduzir conceitos importantes de Genética, úteis para o entendimento dos próximos capítulos, que envolvem a relação entre alelos de diferentes locos genéticos.

2.1 Segregação Independente

Como vimos na seção 1.2 do capítulo anterior, em suas primeiras experiências, Mendel trabalhava com modelos genéticos envolvendo uma característica fenotípica por vez, observada de forma categorizada. A partir dos resultados, ele concluiu o que atualmente denomina-se de Primeira Lei de Mendel, que postula que "cada característica genética de um organismo é condicionada por dois genes, um proveniente do pai e outro da mãe" e que "quando o indivíduo for reproduzir-se, apenas um gene do par seria transmitido da célula sexual".

Após as primeiras conclusões, Mendel dedicou-se ao estudo de cruzamentos em que acompanhava a transmissão de dois caracteres fenotípicos simultaneamente. Em uma destas experiências, ele estudou a transmissão da cor da semente (amarela ou verde) e da forma (lisa ou rugosa); cruzou ervilhas amarelas e lisas puras, com ervilhas verdes e rugosas. Em seguida, deixou que a geração F1 se autofecundasse. A geração F1 foi 100% amarela lisa, uma vez que o gene para amarelo domina aquele para verde, e o liso domina o rugoso. No entanto, na geração F2 apareceram não só ervilhas com fenótipos idênticos aos parentais (amarelas lisas e verdes rugosas),

como também fenótipos novos: amarelas rugosas e verdes lisas. Veja Figura 2.1.

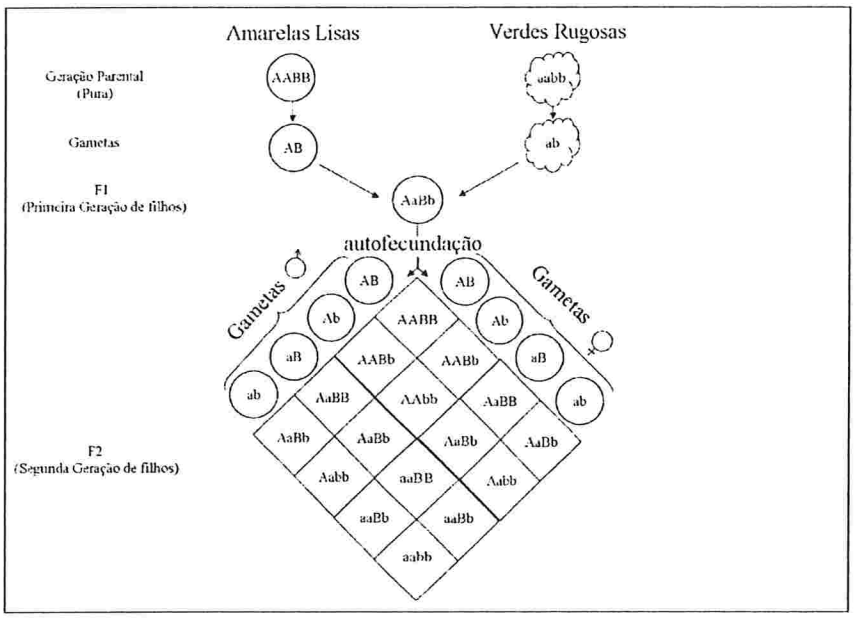


Figura 2.1: Geração F2 considerando a segregação de dois locos genéticos.

Mendel concluiu então, que o gene para cor amarela não é transmitido obrigatoriamente com o gene para forma lisa, assim como o gene para cor verde não é transmitido obrigatoriamente com o gene para forma rugosa. Sendo assim, qualquer um dos genes para cor pode ir ao gameta com qualquer um dos genes para forma, com igual probabilidade, pois não há vínculo entre eles. Trata-se da segregação de dois locos genéticos, onde 2 fenótipos estão sendo observados: a cor e a forma da semente.

A fim de confirmar sua hipótese, Mendel realizou o mesmo experimento com outros caracteres da ervilha; como altura do pé da ervilha e posição da flor, analisando-os sempre dois a dois. Todos os resultados confirmavam sua suspeita de que "os genes para dois ou mais caracteres fenotípicos passam aos gametas de forma totalmente independente um do outro, formando tantas combinações gaméticas quanto possível, com igual probabilidade", a qual ficou conhecida como **Segunda Lei de**

Mendel ou **Lei da Segregação Independente** [Silva Jr & Sasson, 1990]."

Curiosamente, esta independência na segregação dos locos genéticos, postulada por Mendel, nem sempre ocorre. Para que a Segunda Lei se verifique, é necessário que os pares de genes se localizem em cromossomos diferentes ou em locos no mesmo cromossomo suficientemente distantes entre si. Isto significa que a Segunda Lei não é tão geral como a Primeira, pois ela é válida apenas em algumas circunstâncias.

2.2 Recombinação Gênica (*Crossing-over*)

Suponha dois locos, localizados bem próximos no mesmo cromossomo, definidos pelos genótipos **Aa** e **Bb**, respectivamente, tal que os genes **A** e **B** estejam em um mesmo cromossomo, enquanto que os outros alelos **a** e **b** estejam juntos no cromossomo homólogo. No momento da divisão celular, haverá separação dos cromossomos homólogos, e os genes **A** e **B** irão para um mesmo gameta e os genes **a** e **b** irão juntos para outro gameta. Tudo se passa como se os genes **A** e **B**, por um lado, e **a** e **b**, pelo outro, estivessem amarrados, já que estão no mesmo cromossomo. Isso é exatamente o inverso da segregação independente; por estarem em locos bem próximos no mesmo cromossomo, diz-se que os genes estão **ligados** ou em *linkage*. Obteremos, neste caso, apenas dois tipos de gametas; **AB** e **ab**, ao invés de quatro; **AB**, **Ab**, **aB** e **ab**.

Às vezes, entretanto, é possível que ocorra a formação de 4 tipos de gametas mesmo que os genes estejam no mesmo cromossomo. Neste caso, os locos estão não ligados. Isso acontece quando ocorre *crossing-over* (também conhecido como recombinação gênica), que é a troca recíproca de material cromossômico entre segmentos correspondentes de cromossomos homólogos, a qual ocorre na primeira divisão da meiose [Farah, 1997]. Observe a Figura 2.2.

É importante ressaltar que quando dois locos genéticos estão muito próximos no mesmo cromossomo (ligados), a ocorrência de *crossing-over* entre eles é improvável, e os genes que compõem estes locos tendem a ser transmitidos juntos em cada meiose. Por outro lado, quando dois locos estão distantes no mesmo cromossomo (não ligados), é muito provável que ocorra *crossing-over* em algum ponto entre os genes, produzindo os genótipos recombinantes. Note que este raciocínio parte do

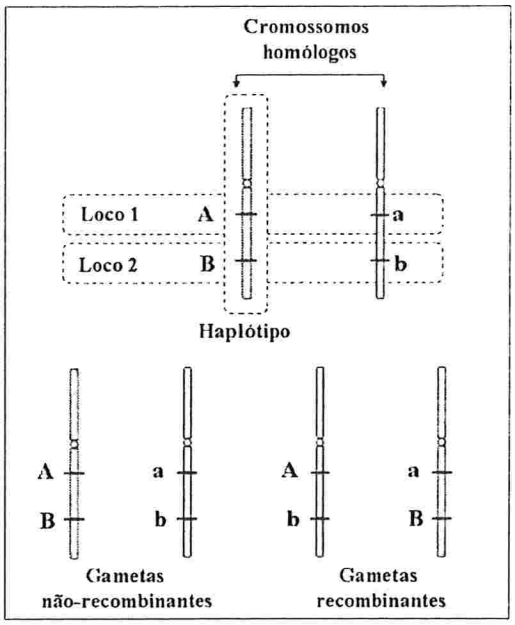


Figura 2.2: Recombinação Gênica. O loco 1 está definido pelos alelos (ou genes) A e a. O loco 2 está definido pelos alelos (ou genes) B e b. De tal forma que tem-se o haplótipo (um dos ramos cromossômicos) AB e o haplótipo ab.

pressuposto de que os genes se distribuem linearmente ao longo dos cromossomos, ocupando posições bem definidas [Amabis & Martho, 1994].

Podemos perceber então, que a freqüência de *crossing-over* depende da distância entre os locos genéticos em questão. Conseqüentemente, a posição de um gene em relação a outro pode, nesse caso, ser estimada pela freqüência dos genótipos recombinantes que eles compartilham entre si. Assim, mapas genéticos não revelam as distâncias físicas entre os genes, mas a posição citológica relativa entre eles [Farah, 1997].

Sendo assim, a herança de dois genes diferentes em um cruzamento depende da disposição e localização destes nos cromossomos parentais. A fração de recombinação pode ser definida como a proporção de todos os gametas que tiveram recombinação gênica (*crossing-over*) entre dois locos de interesse [Pereira & Krieger, 2001].

Se conhecermos a exata localização de um vasto número de locos de interesse, poderemos estudar a fração de recombinação entre estes marcadores genéticos e o

loco no qual se encontra o gene mutante causador da doença em estudo. Portanto, nota-se que um dos objetivos mais importantes da estimação de frações de recombinação é a construção de mapas genéticos. Tradicionalmente, a ligação entre locos genéticos é quantificada usando a fração de recombinação entre um par de locos. Porém, ela pode não ser eficiente para mais de dois locos, pois usualmente a relação entre eles não é aditiva.

Para uma melhor compreensão sobre este tópico, imagine que os genes e/ou marcadores genéticos estão dispostos linearmente em um mapa e suas posições relativas podem ser quantificadas de forma aditiva. Por exemplo, a relação entre 4 locos A, B, C e D, dispostos nesta ordem, poderia ser quantificada da seguinte forma:

$$m_{AD} = m_{AC} + m_{CD}$$
$$m_{AD} = m_{AB} + m_{BC} + m_{CD}$$

onde m_{ij} é definida como a medida de distância citogenética entre os locos i e j e advém do número esperado de *crossing-over* (r_{ij}) entre eles e da atribuição de uma função de distância relacionando estas duas quantidades.

Quando os genes e/ou marcadores genéticos em estudo estão localizados em um intervalo curto, a chance de ocorrência de múltiplos *crossing-over* entre eles é mínima. No entanto, sabemos que quanto maior o número de locos envolvidos, maior a complexidade das relações entre suas posições e, conseqüentemente, das frações de recombinação.

Para uma determinada fração de recombinação r_{ij}, entre dois locos i e j, se uma função f, tal que,

$$m_{ij} = f(r_{ij})$$

existe para todos os pares de genes e/ou marcadores genéticos e é uma função contínua, então $f(r_{ij})$ é definida como uma função de distância citogenética [Liu, 1998].

Para algumas funções de distância citogenética, as inversas podem ser definidas como:

$$r_{ij} = f^{-1}(m_{ij})$$

e ela é utilizada para converter distâncias no mapa genético em frações de recombinação.

A Tabela 2.1 apresenta alguns exemplos de funções de distância citogenética e suas respectivas funções inversas.

	Funções de distância ($m = F(r)$)	Inversa ($r = F^{-1}(m)$)
Morgan (1928)	r	m
Haldane (1919)	$-0,5\log(1-2r)$	$0,5(1-e^{-2\|m\|})$
Kosambi (1944)	$\frac{1}{2}tg^{-1}(2r) = \frac{1}{4}\log\frac{1+2r}{1-2r}$	$\frac{1}{2}tg(2m) = \frac{1}{2}\frac{e^{4m}-1}{e^{4m}+1}$

Tabela 2.1: Exemplo de Funções de Distância Citogenética e suas Funções Inversas.

A Figura 2.3 apresenta as curvas das principais Funções de Distância de acordo com as frações de recombinação e a distância m em Morgans (unidade de medida citogenética) [Ott, 1991].

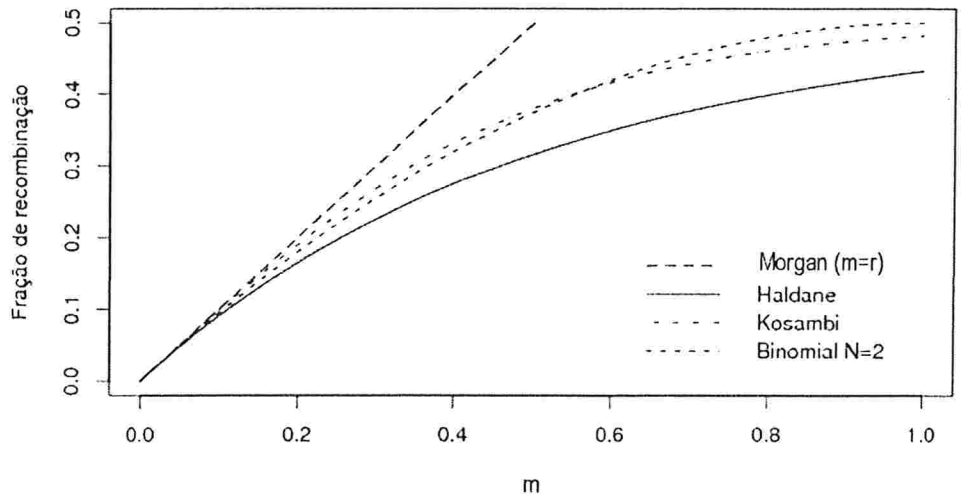


Figura 2.3: Funções de Distância Citogenética.

Retornando ao exemplo citado nesta seção, em que os genes **A** e **B** estão no

mesmo cromossomo, e seus alelos **a** e **b** estão em um cromossomo homólogo, imagine que **A** e **B** sejam transmitidos juntos em 95% das vezes. Isto significa que eles se separam, durante a meiose, em 5% das vezes devido à ocorrência de *crossing-over* entre eles, dando origem aos genótipos recombinantes.

A unidade de distância utilizada em mapas genéticos é o centiMorgan (cM), nome dado em homenagem ao geneticista Thomas Hunt Morgan (1866-1945). Cada centiMorgan representa 1% de chance de dois genes ligados separarem-se durante a meiose. Portanto, em nosso exemplo, a distância entre os genes **A** e **B** é de 5cM. Além disso, pode-se dizer que a unidade centiMorgan expressa a probabilidade de *crossing-over* entre dois genes.

Observe na Figura 2.4 o mapa de marcadores moleculares usado no projeto do InCor. Neste mapa estão indicados os marcadores moleculares localizados nos 21 cromossomos dos ratos, sendo indicadas as respectivas distâncias em centiMorgans.

2.3 Epistasia

Retomando os casos de herança abordados até o momento, temos a situação de uma característica fenotípica ser controlada por apenas um loco genético, o que caracteriza os traços Mendelianos. Contudo, existem vários casos em que o traço em estudo é controlado por mais que um loco genético com efeitos não aditivos. Este fenômeno é denominado **interação gênica** [Silva Jr & Sasson, 1990].

Quando dois pares de genes afetam o mesmo fenótipo, e o gene dominante de um dos pares "domina" o gene do outro par, temos um caso de **epistasia** [Silva Jr & Sasson, 1990]. Portanto, a epistasia é um tipo de interação gênica em que um gene de um loco interfere na atuação dos genes de outro loco. Neste caso, o efeito aditivo não se aplica, dando lugar ao efeito de interação. O fenômeno é semelhante ao efeito de dominância, mas difere no seguinte aspecto: trata-se de uma relação de dominância entre pares de genes em diferentes locos, e não simplesmente entre alelos do mesmo loco genético. Veja Figura 2.5.

O gene que exerce a ação inibitória é chamado epistático, e o que sofre a inibição é chamado hipostático [Amabis & Martho, 1994]. Um alelo dominante de um par

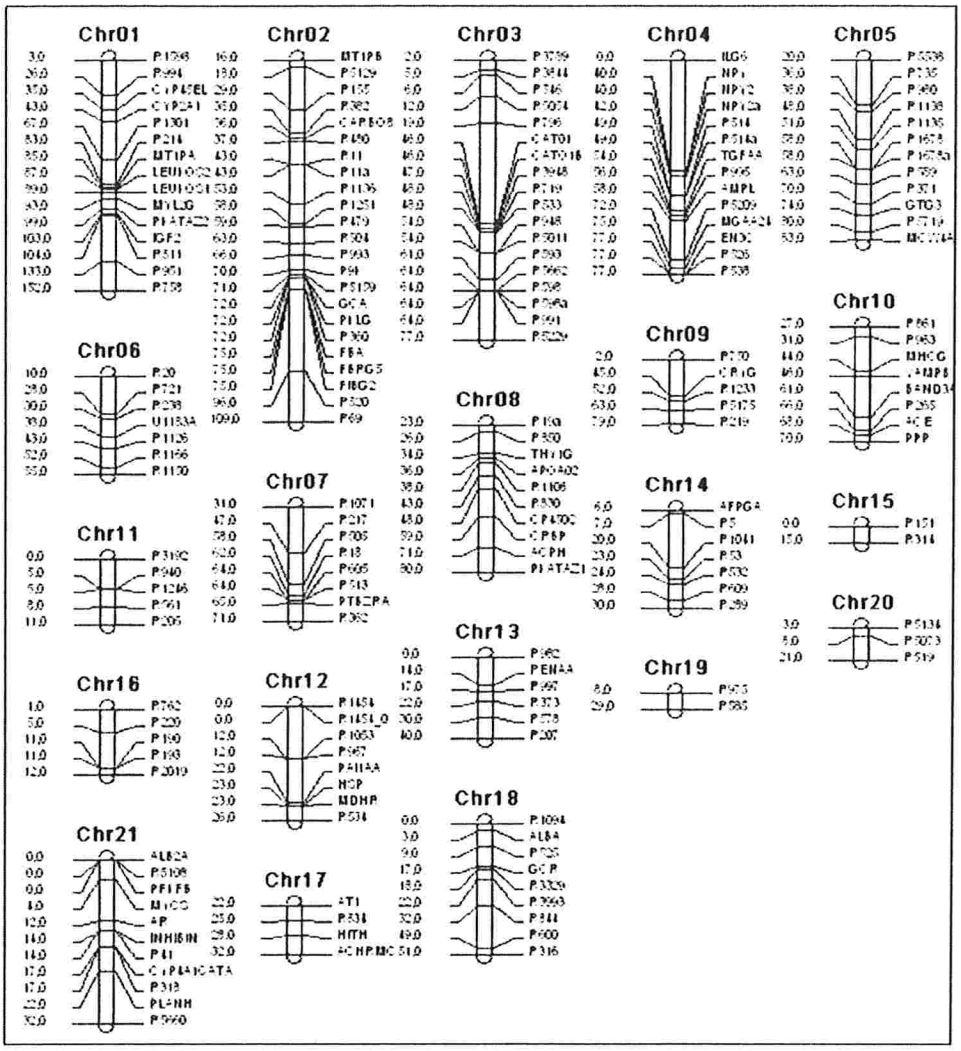


Figura 2.4: Mapa de marcadores moleculares e as respectivas distâncias (em cM) para os dados dos ratos F2 do projeto InCor.

pode ser epistático tanto sobre um alelo dominante quanto sobre um alelo recessivo de outro par.

Para exemplificar, citaremos um caso de epistasia dominante e outro de epistasia recessiva.

Nos cães há dois pares de genes (locos) com segregação independente que controlam a cor do pêlo. Em um loco, o alelo **A** determina a cor preta e o alelo **a**

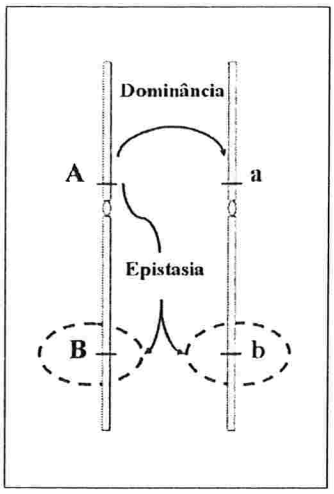


Figura 2.5: Diferença entre Dominância e Epistasia. Dominância: Interação DENTRO do mesmo loco. Epistasia: Interação ENTRE locos.

determina uma coloração marrom. No outro loco, o alelo **B** dominante sobre seu alelo **b**, inibe qualquer manifestação de cor, enquanto que **b** permite que a cor se manifeste. Portanto, qualquer genótipo com o gene **B** manifestará a cor branca, pois **B** é epistático sobre **A** e **a**, além de ser dominante sobre **b**. Este é um caso clássico de epistasia dominante [Silva Jr & Sasson, 1990].

Para os casos de epistasia recessiva, é necessário que o gene recessivo apareça em dose dupla. Nos ratos, as cores possíveis de pêlo são: aguti (pigmentos amarelos e pretos), preto e albino. O alelo **P** produz pigmento preto, e o alelo **p**, quando em dose dupla, inibe totalmente a produção de pigmento e é epistático sobre o gene **A**, que produz pigmento amarelo. O gene **a** não produz pigmento algum. Note que todos aqueles que possuírem genótipo **pp** no genótipo serão necessariamente albinos.

Observe os gráficos apresentados nas Figuras 2.6 e 2.7 e vejamos o comportamento das médias de um fenótipo Y em um exemplo em que não há interação entre dois locos (Loco 1 e Loco 2) e outros dois exemplos em que há interação entre os dois locos, caracterizando um efeito de epistasia.

Veja que na Figura 2.6 o efeito do Loco 1 independe do Loco 2, isto é, o genótipo

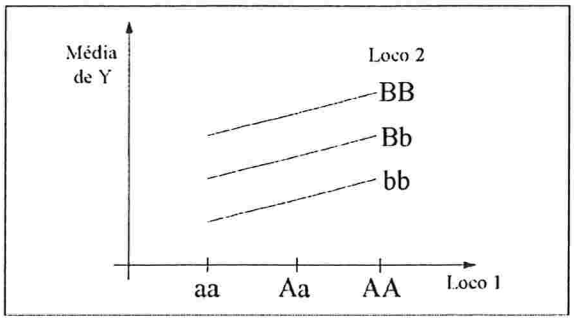


Figura 2.6: Perfis de médias de Y quando não há interação entre os 2 locos.

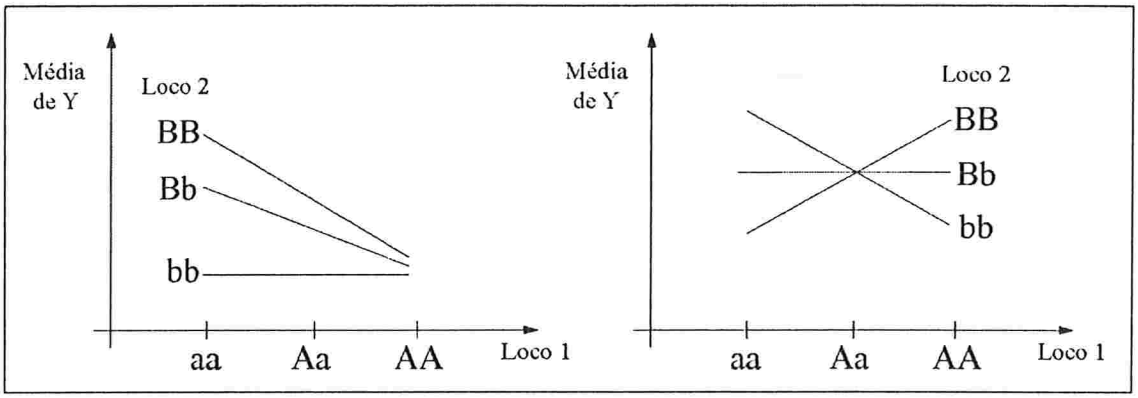


Figura 2.7: Perfis de médias de Y em que há interação entre os 2 locos.

do Loco 2 pode ser tanto BB, Bb quanto bb, que a média fenotípica de Y apresentará sempre um efeito linear crescente. Já na Figura 2.7, as médias da resposta Y não apresentam a mesma tendência como na Figura 2.6; dependendo do genótipo do Loco 1, a tendência pode ser crescente, decrescente ou constante de acordo com os genótipos apresentados pelo Loco 2.

2.4 Pleiotropia

A pleiotropia é a propriedade que certos genes apresentam de controlar mais de uma característica fenotípica, ao mesmo tempo, em determinado organismo. Na interação, dois ou mais locos genéticos condicionam uma mesma característica fenotípica, enquanto que na pleiotropia um único loco genético está associado à manifestação de várias características diferentes [Silva Jr & Sasson, 1990].

Em ervilhas, por exemplo, um único par de alelos condiciona simultaneamente três traços fenotípicos: cor das flores (branca ou vermelha), cor da semente (cinza ou parda) e presença ou ausência de manchas roxas nas axilas das folhas [Amabis & Martho, 1994].

Acredita-se que a maior parte dos genes tenha efeito pleiotrópico. O que ocorre, na prática, é que usualmente percebemos apenas seus efeitos fenotípicos mais marcantes. Além disso, deve haver inúmeros casos em que características completamente diferentes sejam reguladas pelo mesmo gene ou conjunto comum de genes, mas a relação passa despercebida. É importante comentar que o efeito de pleiotropia pode conduzir a uma covariância entre os traços (variáveis quantitativas) que são controlados pelos mesmos genes. A Figura 2.8 apresenta exemplos de epistasia e pleiotropia.

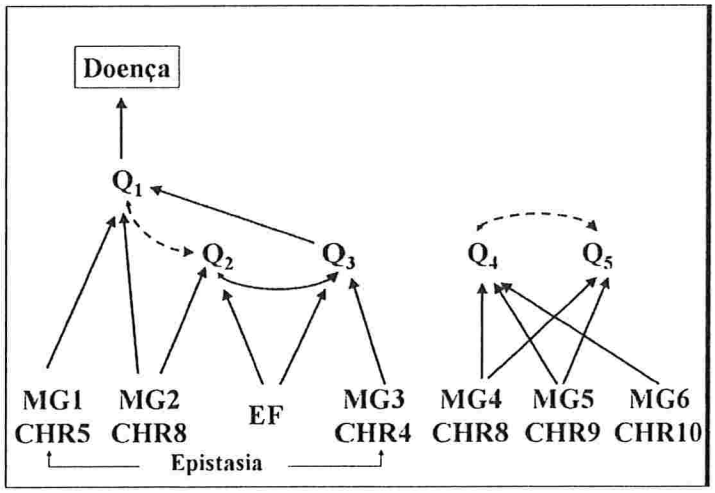


Figura 2.8: Dados Simulados - Q_1,Q_2,Q_3,Q_4 e Q_5 são fenótipos. MG1, MG2, MG3, MG4, MG5 e MG6 são locos genéticos e EF é um fator ambiental. As flechas duplas pontilhadas representam a Correlação Genética. A flecha dupla sólida representa a Correlação Ambiental.

Observe que MG1 e MG3 estão em epistasia, o primeiro loco atua diretamente sobre Q_1 e o segundo atua indiretamente por meio de Q_3, conforme indicado na Figura 2.8, o que significa que um dos locos interfere na atuação do outro loco. Já MG2 e MG5 possuem efeitos pleiotrópicos; o primeiro controla tanto o fenótipo Q_1 quanto o Q_2, enquanto que o loco MG5 controla Q_4 e Q_5.

2.5 Modelagem Genética

A Figura 2.9 apresenta diferentes modelos que explicam a variação de fenótipos em função de variáveis genéticas e ambientais, tal qual foi indicado na expressão 1.1.

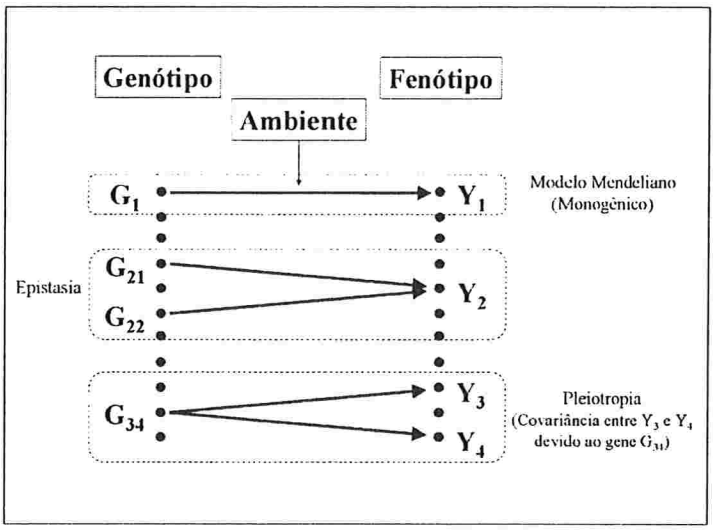


Figura 2.9: Modelagem Genética.

O modelo Mendeliano tem sido adotado para os traços qualitativos, que são aqueles controlados por um único gene e não há efeito do ambiente. O modelo de epistasia trata da situação na qual diferentes locos estão envolvidos na regulação de um mesmo traço quantitativo, em que um gene de um determinado loco interfere na atuação de genes de outro loco, caracterizando um efeito de interação entre eles que pode ser observado pelo padrão de variação da média do traço. Já o modelo de pleiotropia aborda os casos em que um único loco controla a manifestação de vários traços fenotípicos simultaneamente. Tal efeito pode conduzir à existência de covariância (dependência estatística) entre os fenótipos envolvidos. É esperado que no processo de regulação celular da maioria das variáveis quantitativas em humanos, animais e plantas esteja envolvida uma complexa rede em que todos os modelos se comunicam. Nos Capítulos 3 e 4 apresentaremos modelos estatísticos que abordam os casos comentados.

Capítulo 3

Modelo de Regressão Intervalar

Sabe-se que traços quantitativos resultam da influência de múltiplos genes (QTL's) e fatores ambientais. Entretanto, estimar o efeito e a localização de QTL's não é uma tarefa trivial.

Neste capítulo apresentamos a utilização do ajuste de modelos de regressão intervalar como ferramenta para nos auxiliar nesta busca.

3.1 Padrão de variação entre locos genéticos

O modelo básico para a genética quantitativa é aquele apresentado em 1.1, onde o fenótipo é definido pela soma dos fatores genotípicos e ambientais.

Ao utilizarmos dados de indivíduos provenientes da segunda geração de filhos em um delineamento F2, sabemos que há uma grande chance de que as duas linhagens parentais tenham alelos diferentes em todos os genes que controlam os traços de interesse. Entretanto, a diferença entre os traços de ambos os pais reflete um efeito total de todos os genes, ao invés de efeitos individuais de cada um. O mecanismo da segregação genética vai originar proles com novas combinações alélicas, geradas principalmente por segregações independentes de locos em cromossomos diferentes e por recombinações entre locos no mesmo cromossomo. Em função disso, na prole, genes em cromossomos diferentes segregam de forma independente (genes não ligados), enquanto que genes em um mesmo cromossomo podem apresentar dependência estatística ou associação alélica [Sham, 1998]. Genes não-ligados e na ausência de as-

sociação podem ser descritos como fatores ortogonais (em populações finitas) e genes ligados (muito próximos no mesmo cromossomo) e em associação alélica podem ser considerados como fatores com alto grau de colinearidade [Balding *et al.*, 2003].

Na literatura, existem duas principais formas para a análise da estrutura de relacionamento entre os locos em um Mapeamento Genético, são elas: estudos de associação e estudos de ligação.

Os estudos de associação genética utilizam delineamentos observacionais (prospectivos, retrospectivos e transversais) na coleta de dados, onde a estrutura de dados familiares (ou de cruzamentos controlados) não é necessária e o fenótipo é analisado de forma categorizada, por exemplo, ter ou não uma doença. A relevância em se identificar um marcador molecular como fator de risco para a doença está em se assumir a existência de associação entre o marcador e um gene regulador (para que seja possível extrair informações sobre o gene através do marcador). Neste trabalho o problema de mapeamento genético não será abordado via estudos de associação. Maiores detalhes sobre análises deste tipo podem ser encontradas em [Balding *et al.*, 2003].

Os estudos de ligação entre locos baseiam-se na hereditariedade, isto é, na dependência existente entre as informações presentes na geração parental e nas gerações de seus filhos. Para que seja possível realizar mapeamento genético por meio deste tipo de estudo, utilizam-se dados familiares ou de cruzamentos controlados. Este último é objeto de consideração neste trabalho.

O ideal para o mapeamento genético seria que tivéssemos locos ligados e em associação, o que não ocorre com freqüência, pois em uma população com alta taxa de miscigenação, como é a população brasileira, bem como os indivíduos F2, é possível que haja associação entre locos não ligados (distantes).

Neste trabalho vamos explorar os estudos de ligação e dados de cruzamentos controlados, particularmente, os delineamentos F2.

3.2 Análise com genótipos conhecidos

O método mais elementar para entender o efeito de marcadores moleculares, cujos genótipos são conhecidos, na variação de traços quantitativos é testar se há diferenças nos valores esperados do traço entre grupos diferentes de genótipos para um marcador em particular [Zeng, 1994].

A forma mais simples de analisar um mapa de marcadores moleculares é examinar a distribuição dos valores do traço separadamente para cada loco de marcador. Cada teste do efeito de um determinado marcador sobre o traço é realizado independentemente das informações de todos os demais marcadores, o que significa que um cromossomo com M marcadores fornecerá M testes supostamente independentes [Lynch & Walsh, 1998].

Neste tipo de análise o seguinte modelo de regressão linear simples pode ser adotado aos dados dos indivíduos F2 para cada um dos marcadores:

$$y_j = b_0 + b_1 x_j + e_j, \text{com j=1, 2,..., n} \tag{3.1}$$

onde:
y_j é o valor do traço para o j-ésimo indivíduo da população em estudo, sendo n o número de indivíduos na amostra;
b_0 é a média geral;
b_1 é o efeito linear aditivo do marcador;
x_j é a variável que indica o genótipo do marcador para o j-ésimo indivíduo, tal que em um delineamento F2, x_j=0,1,2 para genótipos aa, Aa e AA, respectivamente, para o efeito do alelo A;
e_j é o resíduo aleatório para o j-ésimo indivíduo.

Note que outros tipos de categorização para a variável x_j podem ser utilizados, por exemplo:
x_j=0,1, para genótipos (aa, Aa) e AA;
x_j=0,1, para genótipos aa e (Aa, AA).

No modelo clássico adotado aos dados de indivíduos F2, assume-se que os erros são independentes e identicamente distribuídos seguindo o modelo de probabilidade Normal tal que, $e_j \sim N(0; \sigma^2)$, $j = 1, 2, ..., n$, onde σ^2 é a variância do traço.

Como estamos interessados em saber se o marcador em questão está associado a variações na média do fenótipo, é fácil perceber que sob esta formalização devemos testar se b_1 é igual a zero. Veja que o efeito de dominância do marcador não está sendo considerado nesta formulação do modelo. Assim, da teoria clássica de modelos lineares (ver, por exemplo, Neter et al., 1996), a estatística F, bem como a estatística razão de verossimilhanças, pode ser usada para testar as sequintes hipóteses:

$$H_0 : b_1 = 0$$
$$H_1 : b_1 \neq 0.$$

Para ilustrar o procedimento de testar as hipóteses acima no contexto de mapeamento genético com genótipos conhecidos, apresentamos um gráfico de perfis da estatística razão de verossimilhanças na Figura 3.1, considerando os dados de marcadores moleculares e da pressão sistólica pós-sal (SBPS) do projeto InCor.

Note que os picos mais altos (acima do valor crítico 11,5 assumido empiricamente) indicam que devemos rejeitar H_0 e que há indícios de que o referido marcador tem um efeito significativo sobre a regulação da SBPS, isto é, que tal marcador é o próprio QTL ou que está ligado ao QTL que controla o traço em estudo.

A análise para mapeamento com genótipos conhecidos é útil para verificar o quão informativo um mapa de marcadores é para a análise de QTL's e pode ser usada para detectar se um QTL está ligado a algum marcador. Deste modo, se o objetivo for estimar a posição e o efeito do QTL, cujo genótipo é desconhecido, os sinais dos marcadores (avaliados, por exemplo, pelo valor da estatística razão de verossimilhanças do teste de ligação correspondente) podem ser usados no procedimento de estimação, como descreveremos na próxima seção.

Veja que se algum efeito significativo é detectado na análise dos marcadores, isto indica que o próprio marcador é o gene de interesse ou tal marcador está ligado ao gene de interesse e, portanto, mostra um efeito "aparente" devido à sua associação (proximidade) com o gene.

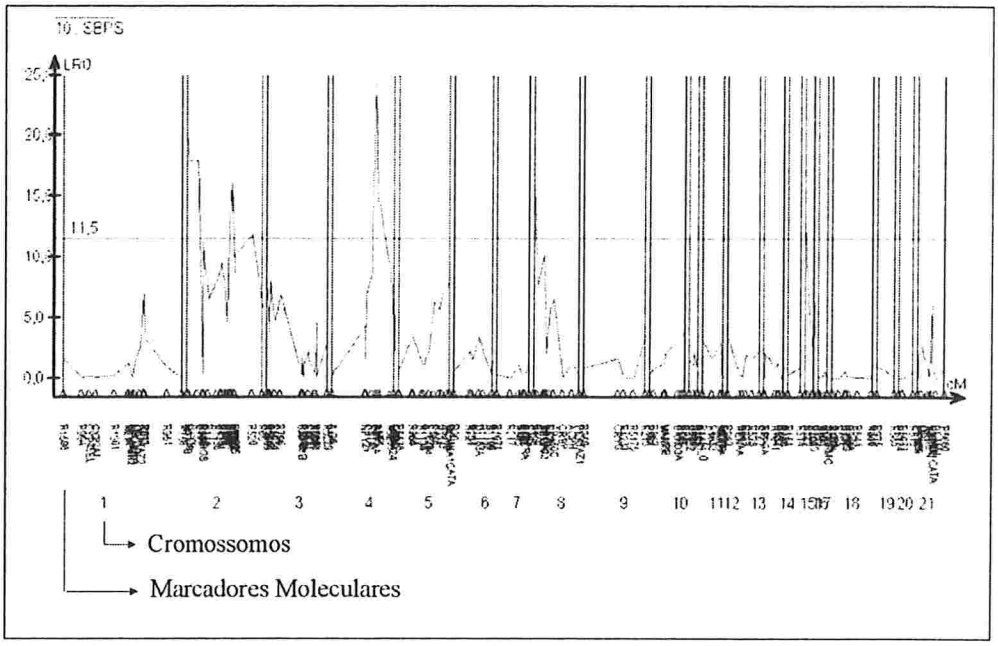


Figura 3.1: Gráfico de Perfis da estatística razão de verossimilhanças para o mapeamento com genótipos conhecidos da Pressão Sistólica Pós-Sal para os ratos F2 do projeto InCor.

Embora simples, a análise para genótipos conhecidos capta as idéias básicas do mapeamento de QTL's. No entanto, esta abordagem simples apresenta muitos problemas, entre os quais destacam-se [Zeng, 1994]:

- O método não é capaz de distingüir se os marcadores estão associados à um ou mais QTL's;
- O método não estima as posições mais prováveis dos QTL's;
- Os efeitos dos QTL's são, provavelmente, subestimados porque são confundidos com as freqüências de recombinação entre os locos;
- Devido aos efeitos de confundimento, comentados acima, o método não tem muito poder e são necessários muitos indivíduos para o teste.

3.3 Mapeamento Intervalar

Com o intuito de sanar algumas deficiências do modelo de regressão para genótipos conhecidos, Lander & Botstein (1989) propuseram um modelo de regressão intervalar.

O Mapeamento Intervalar, também conhecido como análise com marcadores flanqueadores[1], tem como objetivo percorrer todo o genoma, fixando posições, na busca por evidências da presença de um QTL. Este mapeamento utiliza dois marcadores com genótipos observáveis para definir o intervalo no qual será procurado o QTL [Haley & Knott, 1992]. Com base nos genótipos conhecidos destes marcadores flanqueadores, é possível estimar os efeitos e a localização de um possível QTL utilizando sub-intervalos de 1 ou 2cM, por exemplo. Verifique a Figura 3.2.

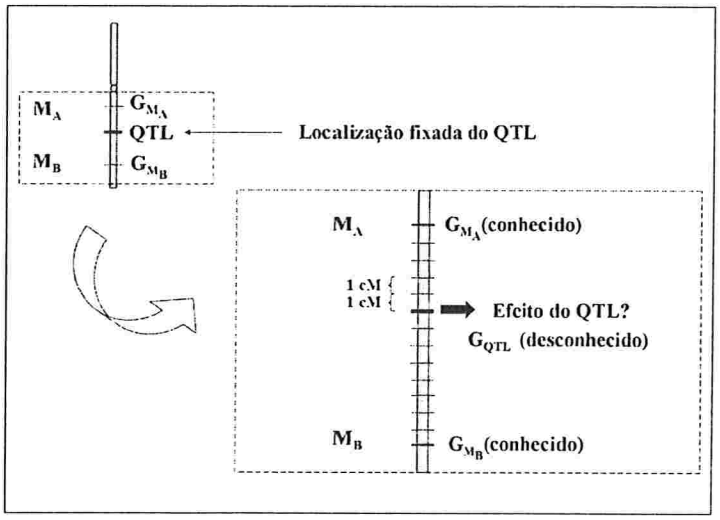


Figura 3.2: Exemplo de Mapeamento Intervalar. M_A e M_B são os marcadores flanqueadores com genótipos, G_{M_A} e G_{M_B}, conhecidos. G_{QTL} é o genótipo desconhecido do QTL, que pode ser igual a Q1Q1, Q1Q2 ou Q2Q2 para indivíduos F2.

Como discutido em Haley & Knott (1992), a utilização de métodos que envolvem marcadores flanqueadores tem provado ser uma ferramenta poderosa no mapeamento de QTL's em segregações derivadas de cruzamentos controlados. Além disso, este mapeamento apresenta um aumento no poder de detecção do QTL e também fornece estimações mais precisas da posição e do efeito do mesmo.

[1] Marcadores flanqueadores são aqueles que margeiam a localização de um suposto QTL.

No mapeamento intervalar a análise de ligação é realizada utilizando-se uma função de distância citogenética, digamos a função de distância de Haldane, que transforma distâncias em frações de recombinação entre locos e, a partir da informação do genótipo dos marcadores, calcula-se uma determinada estatística para cada incremento (sub-intervalos entre os dois marcadores flanqueadores) do intervalo. Quando o valor de tal estatística exceder um determinado valor de significância, pode-se assumir que um QTL foi encontrado naquela localização.

O modelo para um único QTL presente entre dois marcadores flanqueadores codominantes (isto é, com as três categorias genotípicas) pode ser aplicado no mapeamento da geração F2 de um cruzamento entre duas linhagens que carregam alelos diferentes para os 3 locos sob estudo; os dois referentes aos marcadores flanqueadores e um referente ao QTL.

Suponha que os 9 genótipos possíveis para os marcadores flanqueadores sejam os apresentados na Tabela 3.1.

	G1	G2	G3	G4	G5	G6	G7	G8	G9
Loco 1	A1A1	A1A1	A1A1	A1A2	A1A2	A1A2	A2A2	A2A2	A2A2
Loco 2	B1B1	B1B2	B2B2	B1B1	B1B2	B2B2	B1B1	B1B2	B2B2

Tabela 3.1: Possíveis genótipos dos marcadores flanqueadores considerando uma população F2.

Digamos que os genótipos das linhagens parentais cruzadas para um certo conjunto de 3 locos sejam $A1A1Q1Q1B1B1$ e $A2A2Q2Q2B2B2$. Logo, na geração F2 cada um dos três locos pode estar definido para uma das três constituições genéticas possíveis (dois tipos homozigotos e um heterozigoto). Por exemplo, no loco do QTL os indivíduos podem ser Q1Q1, Q1Q2 ou Q2Q2.

O valor esperado do traço para os 3 genótipos possíveis do QTL na geração F2 são: $\mu + a$, $\mu + d$ ou $\mu - a$ para $Q1Q1$, $Q1Q2$ (ou, equivalentemente $Q2Q1$) e $Q2Q2$, respectivamente, onde μ é a média do traço e a e d são os desvios devido aos efeitos aditivo e de dominância do QTL, respectivamente.

Como já foi visto no Capítulo 2, a distância m entre dois marcadores moleculares, digamos marcador A e marcador B, sempre é conhecida, pois sabemos onde os mesmos se localizam no cromossomo (como exemplo, ver Figura 1.7). Ao fixarmos

uma posição para um suposto QTL entre estes marcadores, podemos determinar sua distância até o marcador A (m_A) e até o marcador B (m_B). Com estas informações, de acordo com a seção 2.2, vimos que é possível obter também as respectivas frações de recombinação, (r_A) e (r_B), fazendo uso de funções de distância citogenética.

Então, suponha que a fração de recombinação entre os marcadores flanqueadores é conhecida e igual a r (esta fração pode ser calculada a partir dos dados do marcador antes da análise de QTL's). Já a fração de recombinação entre A e Q é r_A e entre Q e B é r_B. Para toda e qualquer análise assumimos ausência de interferência, isto é, que o número possível de recombinações entre os locos é infinito, pois os eventos são independentes, e assim esperamos que $r = r_A + r_B - 2r_A r_B$. Por fim, utilizamos, por exemplo, a função de distância citogenética de Haldane para converter distâncias (em Morgans) em frações de recombinação.

A média esperada do traço em termos do suposto QTL, para cada combinação possível do genótipo dos marcadores flanqueadores para indivíduos do cruzamento F2, pode ser obtida como veremos a seguir. Considere os gametas $A1Q1B1$ e $A1Q2B2$ que podem ser transmitidos a um indivíduo F2. O gameta $A1Q1B1$ tem probabilidade de ocorrência $(1 - r_A)(1 - r_B)/2$ e o gameta $A1Q2B2$ tem probabilidade de ocorrência $r_A(1 - r_B)/2$. A Figura 3.3 ilustra o esquema de cálculo destas probabilidades.

Vamos calcular a probabilidade marginal do genótipo dos marcadores flanqueadores. Observe a Figura 3.4 para um caso em particular.

A seguir, a Figura 3.5 apresenta os 4 genótipos possíveis de QTL para o genótipo de marcador $A1A1B1B2$ e o cálculo das respectivas probabilidades conjunta de ocorrência.

O valor esperado do traço y considerando o efeito de um QTL flanqueado pelos marcadores com genótipos G_A e G_B, denotado por $E(y|G_A, G_B)$, é dado por:

$$\begin{aligned} E(y|G_A, G_B) &= \mu_{Q_1Q_1} P(G_{QTL} = Q_1Q_1|G_A, G_B) + \mu_{Q_1Q_2} P(G_{QTL} = Q_1Q_2|G_A, G_B) \\ &\quad + \mu_{Q_2Q_2} P(G_{QTL} = Q_2Q_2|G_A, G_B) \\ &= (\mu + a)P(Q1Q1|G_A, G_B) + (\mu - a)P(Q2Q2|G_A, G_B) \end{aligned}$$

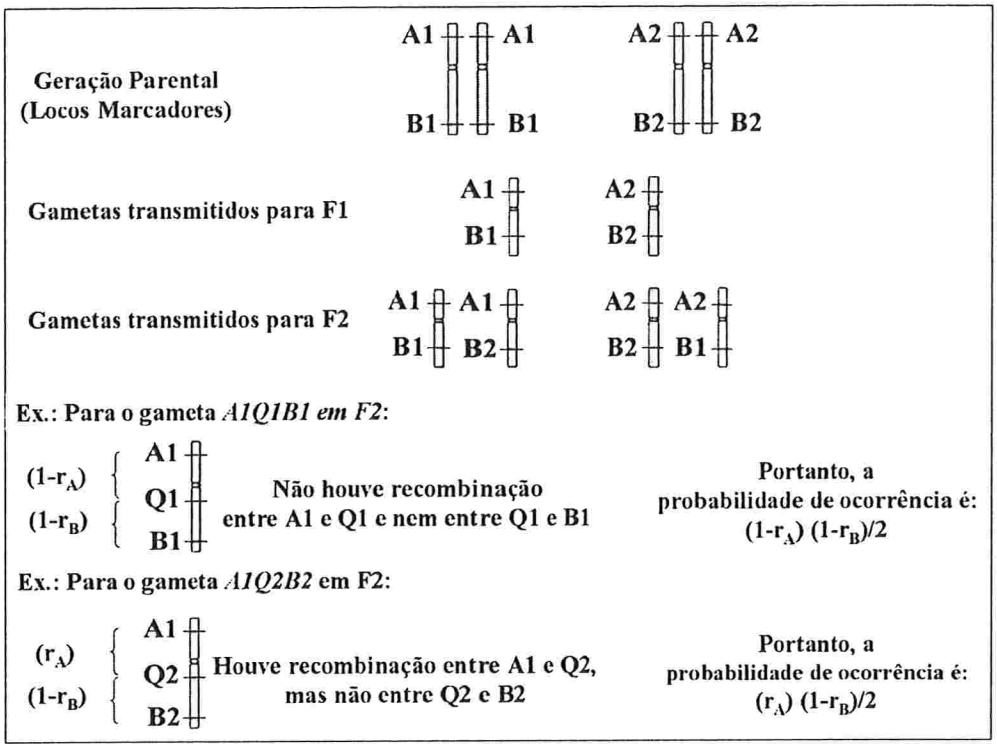


Figura 3.3: Probabilidades de ocorrência dos gametas $A1Q1B1$ e $A1Q2B2$.

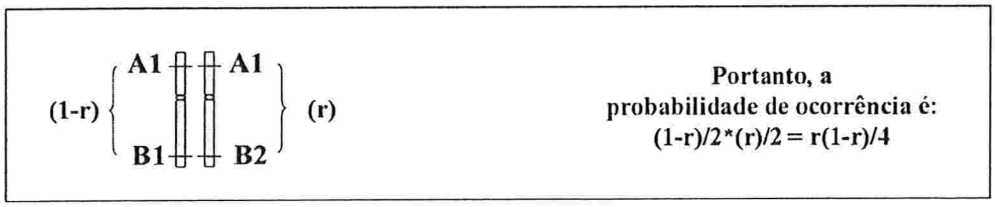


Figura 3.4: Probabilidade de ocorrência do genótipo $A1A1B1B2$ dos marcadores flanqueadores.

$$
\begin{aligned}
& +(\mu + d)P(Q1Q2|G_A, G_B) \\
= \; & \mu + a[P(Q1Q1|G_A, G_B) - P(Q2Q2|G_A, G_B)] + dP(Q1Q2|G_A, G_B) \\
= \; & \mu + aX_{ag} + dX_{dg}. \qquad (3.2)
\end{aligned}
$$

onde,

$\mu_{Q_1Q_1} = \mu + a$, $\mu_{Q_1Q_2} = \mu + d$, $\mu_{Q_2Q_2} = \mu - a$ são os valores esperados do traço y para o indivíduo F2 com os respectivos genótipos no loco referente ao QTL;

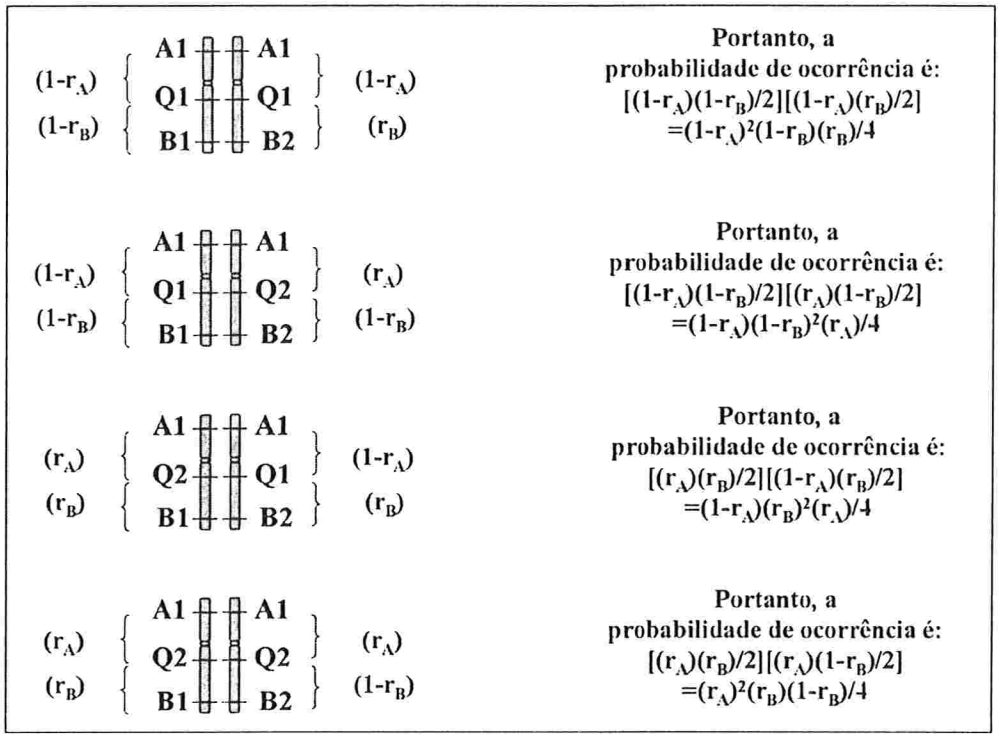


Figura 3.5: Probabilidades de ocorrência dos 4 genótipos possíveis de QTL para o genótipo de marcador A1A1B1B2.

G_A, G_B são os genótipos dos marcadores flanqueadores M_A e M_B;
μ é a média geral;
a é o efeito aditivo do QTL;
d é o efeito de dominância do QTL;
X_{ag} e X_{dg} são os valores preditos do QTL associados aos efeitos aditivo e de dominância, respectivamente, calculados por meio das informações genotípicas dos marcadores flanqueadores do suposto QTL e por meio das frações de recombinação associadas.

Retomando o exemplo dos marcadores A e B com genótipos G_M=$A1A1B1B2$, dos cálculos apresentados nas Figuras 3.4 e 3.5 temos que:

$$P(G_M) = P(G_A, G_B) = P(A1A1B1B2) = \frac{r(1-r)}{4} \quad (3.3)$$

$$P(Q1Q1;A1A1B1B2) = \frac{(1-r_A)^2 r_B(1-r_B)}{4} \tag{3.4}$$

$$P(Q1Q2;A1A1B1B2) = \frac{(1-r_A)(1-r_B)^2(r_A)}{4} + \frac{(1-r_A)^2(r_B)(1-r_B)}{4} \tag{3.5}$$

$$P(Q2Q2;A1A1B1B2) = \frac{(r_A)^2(r_B)(1-r_B)}{4}. \tag{3.6}$$

Utilizando as equações 3.2, 3.3, 3.4, 3.5 e 3.6, temos que:

$$\begin{aligned} X_{ag} &= P(Q1Q1|G_M) - P(Q2Q2|G_M) \\ &= \frac{P(Q1Q1;G_M))}{P(G_M)} - \frac{P(Q2Q2;G_M)}{P(G_M)} \\ &= \frac{\frac{(1-r_A)^2 r_B(1-r_B)}{4} - \frac{(r_A)^2(r_B)(1-r_B)}{4}}{\frac{r(1-r)}{4}} \\ &= \frac{(1-r_A)^2 r_B(1-r_B) - (r_A)^2 r_B(1-r_B)}{r(1-r)}. \end{aligned} \tag{3.7}$$

$$\begin{aligned} X_{dg} &= P(Q1Q2|G_M) \\ &= \frac{P(Q1Q2;G_M)}{P(G_M)} \\ &= \frac{\frac{(1-r_A)(1-r_B)^2(r_A)}{4} + \frac{(1-r_A)^2(r_B)(1-r_B)}{4}}{\frac{r(1-r)}{4}} \\ &= \frac{r_A(1-r_A)(1-r_B)^2 + r_A(1-r_A)r_B^2}{r(1-r)}. \end{aligned} \tag{3.8}$$

Portanto, a média esperada do traço y para um indivíduo F2 de genótipo de marcador G_M=$A1A1B1B2$ é:

$$\begin{aligned} E(y|\mathrm{G}_M) &= \mu + a\frac{(1-r_A)^2 r_B(1-r_B) - r_A^2 r_B(1-r_B)}{r(1-r)} \\ &\quad + d\frac{r_A(1-r_A)(1-r_B)^2 + r_A(1-r_A)r_B^2}{r(1-r)}. \end{aligned} \tag{3.9}$$

As variáveis preditoras associadas aos coeficientes a e d em termos da fração de recombinação para cada um dos 9 genótipos possíveis para os marcadores flanqueadores em uma população F2 são dados na Tabela 3.2 [Haley & Knott, 1992], e podem ser calculados de maneira análoga à apresentada para o genótipo $A1A1B1B2$.

Genótipo do marcador	X_{ag}	X_{dg}
$A1A1B1B1$	$\frac{[(1-r_A)^2(1-r_B)^2-r_A^2r_B^2]}{(1-r)^2}$	$\frac{[2r_A(1-r_A)r_B(1-r_B)]}{(1-r)^2}$
$A1A1B1B2$	$\frac{[(1-r_A)^2r_B(1-r_B)-r_A^2r_B(1-r_B)]}{r(1-r)}$	$\frac{[r_A(1-r_A)(1-r_B)^2+r_A(1-r_A)r_B^2]}{r(1-r)}$
$A1A1B2B2$	$\frac{[(1-r_A)^2r_B^2-r_A^2(1-r_B)^2]}{r^2}$	$\frac{[2r_A(1-r_A)r_B(1-r_B]}{r^2}$
$A1A2B1B1$	$\frac{[r_A(1-r_A)(1-r_B)^2-r_A(1-r_A)r_B^2)]}{r(1-r)}$	$\frac{[(1-r_A)^2r_B(1-r_B)+r_A^2r_B(1-r_B)]}{r(1-r)}$
$A1A2B1B2$	0	$\frac{[r_A^2r_B^2+r_A^2(1-r_B)^2(1-r_A)^2r_B^2+(1-r_A)^2(1-r_B)^2]}{[r^2+(1-r)^2]}$
$A1A2B2B2$	$\frac{[r_A(1-r_A)r_B^2-r_A(1-r_A)(1-r_B)^2)]}{r(1-r)}$	$\frac{[(1-r_A)^2r_B(1-r_B)+r_A^2r_B(1-r_B)]}{r(1-r)}$
$A2A2B1B1$	$\frac{[r_A^2(1-r_B)^2-(1-r_A)^2r_B^2]}{r^2}$	$\frac{[2r_A(1-r_A)(r_B)(1-r_B)]}{r^2}$
$A2A2B1B2$	$\frac{[r_A^2r_B(1-r_B)-(1-r_A)^2r_B(1-r_B)]}{r(1-r)}$	$\frac{[r_A(1-r_A)(1-r_B)^2+r_A(1-r_A)r_B^2]}{r(1-r)}$
$A2A2B2B2$	$\frac{[r_A^2r_B^2-(1-r_A)^2(1-r_B)^2]}{(1-r)^2}$	$\frac{[2r_A(1-r_A)r_B(1-r_B)]}{(1-r)^2}$

Tabela 3.2: Variáveis preditoras do efeito genotípico aditivo e de dominância de um QTL para todos os genótipos de marcadores flanqueadores possíveis em uma população F2.

Conhecendo as expressões para as variáveis preditoras X_a e X_d apresentadas na Tabela 3.2, o modelo de regressão linear múltipla dado em 3.2 pode ser ajustado via a teoria de modelos lineares clássicos da seguinte maneira: dado um intervalo entre 2 marcadores e um suposto QTL fixado em posições intermediárias (intervalos de 1cM, por exemplo) entre eles as variáveis X_{ag} e X_{dg} são calculadas para todos os indivíduos da amostra e inferências podem ser feitas para os coeficientes de regressão a e d, por meio de procedimentos de estimação e teste destes parâmetros genéticos.

A teoria clássica de regressão múltipla pode ser utilizada para ajustar μ, a e d para cada posição fixada de QTL separadamente. Deste modo, por meio da análise clássica pode-se obter a soma de quadrados dos resíduos e também o quadrado médio dos resíduos, permitindo o cálculo do coeficiente de determinação, isto é, a proporção da variância total de y explicada pelo efeito do QTL. Esta medida é conhecida em Genética como a herdabilidade do traço y. Além disso, pode-se construir testes para os efeitos do QTL via estatística F ou mesmo por meio da estatística razão de verossimilhanças.

3.3.1 Testes sobre o efeito do QTL

Acompanhe a aplicação do método de Mapeamento Intervalar utilizando os marcadores **R589** e **R371** como marcadores flanqueadores no cromossomo 5 dos dados dos ratos F2 do projeto do InCor. Observe a Figura 3.6.

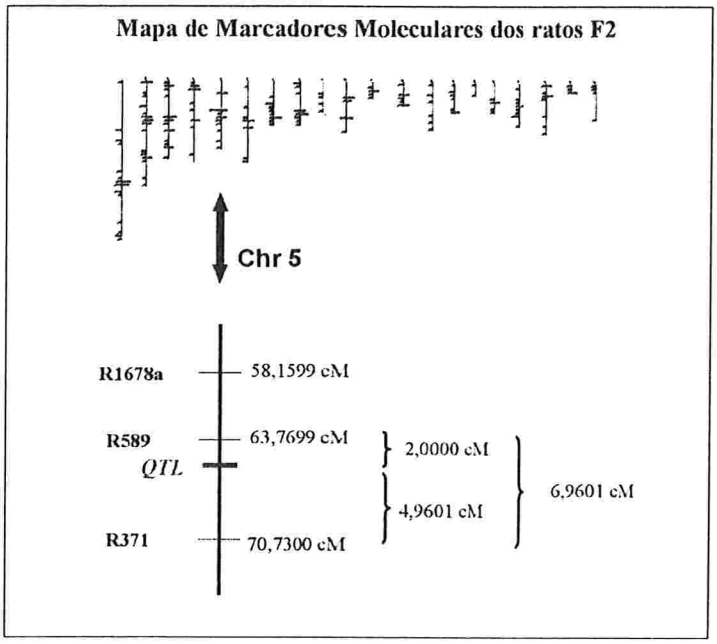


Figura 3.6: Distância (em cM) entre os marcadores flanqueadores e a suposta localização do QTL.

De acordo com a Tabela 2.1, assumindo a função de distância de Haldane, temos que a fração de recombinação (r) deve ser calculada de acordo com a seguinte expressão:

$$r = 0,5(1 - e^{-2|m|}), \text{ onde m é a distância em Morgans.}$$

Portanto, as distâncias m (entre **R589** e **R371**), m_1 (entre **R589** e o QTL) e m_2 (entre o QTL e **R371**) podem ser convertidas em frações de recombinação r, r_1 e r_2, respectivamente, de acordo com a Tabela 5.3

	Distância em cM	Fração de Recombinação
entre **R589** e **R371**	m=6,9601	r=0,06497387
entre **R589** e o QTL	m_1=2,0000	r_1=0,01960528
entre o QTL e **R371**	m_2=4,9601	r_2=0,04722012

Tabela 3.3: Distância citogenética e respectivas frações de recombinação.

Utilizando o mesmo raciocínio apresentado nas Figuras 3.3, 3.4 e 3.5, e utilizando as equações para X_{ag} e X_{dg} podemos construir a seguinte Tabela 3.4.

	$P(Q_1Q_1)$	$P(Q_1Q_2)$	$P(Q_2Q_2)$	$P(G_M)$	X_{ai}	X_{di}
A_1A_1/B_1B_1	0,2181	0,0004	0,0000	0,2185	0,99604	0,00198
A_1A_1/B_1B_2	0,0216	0,0087	0,0000	0,0303	0,42389	0,28791
A_1A_1/B_2B_2	0,0005	0,0004	0,0001	0,0010	0,09799	0,40968
A_1A_2/B_1B_1	0,0087	0,0216	0,0000	0,0303	-0,42488	0,71209
A_1A_2/B_1B_2	0,0009	0,4375	0,0009	0,4393	-0,99409	0,99606
A_1A_2/B_2B_2	0,0000	0,0216	0,0087	0,0303	-0,71138	0,71209
A_2A_2/B_1B_1	0,0001	0,0004	0,0005	0,0010	-0,32703	0,40968
A_2A_2/B_1B_2	0,0000	0,0087	0,0216	0,0303	-0,28763	0,28791
A_2A_2/B_2B_2	0,0000	0,0004	0,2181	0,2185	-0,00198	0,00198
	0,2500	0,5000	0,2500	1,0000		

Tabela 3.4: Valores para X_a e X_d para cada genótipo de marcador.

Para a atribuição dos valores de X_a e X_d para os indivíduos F2, deve-se observar os genótipos conhecidos dos marcadores flanqueadores para cada indivíduo e associá-los aos valores pertinentes de acordo com a Tabela 3.4. A partir daí, as seguintes hipóteses estatísticas podem ser formuladas sobre os efeitos aditivo a e o de dominância d do QTL:

$$\begin{aligned} \mathrm{H}_0 &: a = 0, d = 0 \\ \mathrm{H}_1 &: a \neq 0, d = 0 \\ \mathrm{H}_2 &: a = 0, d \neq 0 \\ \mathrm{H}_3 &: a \neq 0, d \neq 0. \end{aligned}$$

Zeng (1994), propõe os seguintes testes do efeito do QTL:

(1) H_0 x H_3;
(2) H_1 x H_3;
(3) H_2 x H_3;
(4) H_0 x H_1 ou
(5) H_0 x H_2.

Veja que o teste (1) H_0 x H_3 equivale a testar simultaneamente se há efeito aditivo e de dominância do QTL, assim como o teste (4) H_0 x H_1 equivale a testar se há efeito aditivo do QTL na ausência de efeito de dominância, e assim por diante.

3.3.2 Inferências clássicas sobre o Modelo de Regressão Intervalar

Fazendo uso da teoria clássica de Modelos Lineares , Haley & Knott (1992) adotam o seguinte modelo:

$$y_j = \mu + aX_{aj} + dX_{dj} + e_j,$$
$$\text{onde } e_j \sim N(0;\sigma^2), j = 1, 2, ..., n.$$

onde n é o número de indivíduos da geração F2 na amostra.

Sob esta formulação, inferências sobre os coeficientes a e d podem ser obtidas de duas formas:
(i) via Máxima-Verossimilhança;
(ii) via Mínimos Quadrados.
Sob a suposição clássica de normalidade e homocedasticidade dos resíduos, é conhecido que tais procedimentos conduzem aos mesmos estimadores [Draper & Smith, 1981].

Convencionalmente, na literatura de Genética adotam-se os métodos baseados em máxima-verossimilhança, embora tais métodos sejam relativamente complexos e possam ser computacionalmente lentos quando se considera modelos mais gerais para a distribuição dos erros. De qualquer forma, certamente estes métodos fornecem boas estimativas dos efeitos do QTL.

Em análises baseadas na teoria de máxima-verossimilhança, o teste dos parâmetros do modelo é obtido por meio da estatística razão de verossimilhanças, definida pela razão entre a verossimilhança maximizada no espaço geral, isto é, sob H_1 (L_1) e a verossimilhança maximizada sob o espaço restrito definido pela hipótese nula H_0 (L_0). Assim, a estatística razão de verossimilhanças é:

$$\text{LRT} = 2\ln\left(\frac{L_1}{L_0}\right) \tag{3.10}$$

a qual, sob condições de regularidade, é assintoticamente distribuída como uma Qui-quadrado com p graus de liberdade, onde p é definido como a diferença entre as dimensões dos dois espaços dos parâmetros sob H_1 e H_0.

Quando os resíduos são independentes e identicamente distribuídos como uma $N(0, \sigma^2)$, a estatística da razão de verossimilhanças pode ser escrita em termos da Soma de Quadrados Residual do modelo completo ($SQR_{completo}$), do modelo reduzido ($SQR_{reduzido}$), e do total de observações (n):

$$\text{LRT} = n\ln\left(\frac{SQR_{reduzido}}{SQR_{completo}}\right) \tag{3.11}$$

Muitas vezes, as suposições de que os resíduos são identicamente distribuídos segundo uma Normal entre as classes de genótipos de marcadores não estão satisfeitas. Mesmo assim, estudos empíricos mostram que o valor de $n\ln\left(\frac{SQR_{reduzido}}{SQR_{completo}}\right)$ fornece uma boa aproximação para a estatística da razão de verossimilhanças, sendo uma estatística robusta [Haley & Knott, 1992].

3.3.3 Estatística *Lod score*

A estatística *Lod Score* tem sido muito utilizada na área da Genética, especialmente em análise de ligação, e é definida como o logaritmo na base 10 da razão de verossimilhanças:

$$Lod = \log_{10}\left(\frac{L_1}{L_0}\right). \tag{3.12}$$

A razão de sua popularidade advém essencialmente da facilidade de sua interpretação, pois *Lod scores* iguais a, por exemplo, 1, 2 e 3 representam razões de verossimilhanças iguais a 10, 100 e 1000 [Sham, 1998].

Sob condições de regularidade e premissas clássicas, temos que a estatística *Lod Score* (*Lod*), a estatística Razão de Verossimilhanças (LRT) e a estatística F (usual) são alternativas equivalentes para se testar o efeito do QTL.

Sejam:

$$SQR = \sum_{j=1}^{n}(y_j - \hat{y}_j)^2$$

$$SQT = \sum_{j=1}^{n}(y_j - \bar{y}_j)^2$$

$$\lambda = \left(\frac{SQT}{SQR}\right)^{n/2}$$

onde,
y_j é o valor do traço para o j-ésimo indivíduo da população em estudo;
$\hat{y}_j$ é o valor predito do traço para o j-ésimo indivíduo da população em estudo;
$\bar{y}_j$ é a média amostral do traço para a população em estudo.

Pode-se mostrar que:

- Lod=$\log_{10}(\lambda)$;
- LRT=$2\ln(\lambda) = 4,61\ Lod$;

- $F = \frac{n-k}{k-1}\left(\frac{SQT - SQR}{SQR}\right) = \frac{n-k}{k-1}(\lambda^{2/n} - 1)$.

No Mapeamento Intervalar, deve-se testar a significância dos efeitos aditivo (a) e de dominância (d) para cada localização possível do QTL. Para tanto, podemos calcular o valor da estatística *Lod score* para cada incremento e, assim, construir um Gráfico de Perfis da estatística de teste.

O próximo passo é fixar um valor crítico da estatística de teste que nos indique a significância do efeito. Em geral, tem sido apontada na literatura [Lander & Botstein, 1989] a necessidade do ajuste do valor crítico para os múltiplos testes realizados, de acordo com: o número e o tamanho dos intervalos, níveis diferentes de herdabilidade, números diferentes de múltiplos QTL's ligados e não-ligados, etc. Churchill & Doerge (1994) sugeriram um teste de permutação para estimar empiricamente um valor crítico apropriado para um determinado banco de dados.

Em geral, adota-se o valor crítico 2,5, que é sugerido por Kao et al. (1999) de forma exploratória. Note que $Lod = 2,5$ em um teste com 1 grau de liberdade equivale a um nível descritivo $p < 0,001$. Isto significa que, quando a estatística *Lod score* apresentar um valor maior que 2,5 para uma determinada posição, temos uma indicação de que o QTL deve estar naquela região cromossômica com uma chance de um resultado falso positivo menor que 0,001. Veja a Figura 5.2. A estatística LOD0, presente na ordenada do gráfico, corresponde ao teste H_0 x H_3, que equivale a testar simultaneamente se existe efeito aditivo e de dominância do QTL. Neste caso, conclui-se que há evidência para a presença de QTL's em posições nos cromossomos 2, 4, 8 e 16, sendo que no cromossomo 2 há dois locos de QTL's.

É importante lembrar que, neste método, múltiplos testes estão sendo feitos simultaneamente. O fato de cada teste ter um nível de significância α associado, não significa que o nível de significância global, para a família de comparações, também seja igual a α. Sendo assim, é importante corrigir o valor do α para evitar falsos positivos. A abordagem mais simples e também a mais conservativa é a **correção de Bonferroni**, que estabelece um nível de significância α para todo o conjunto de testes e α/n para cada um deles [Neter *et al.*, 1996], [Ott, 1991]. Além deste procedimento, podemos utilizar também como nível de significância conjunto do teste a "Razão de Falsas Descobertas"(FDR, *false discovery rate*), definida como

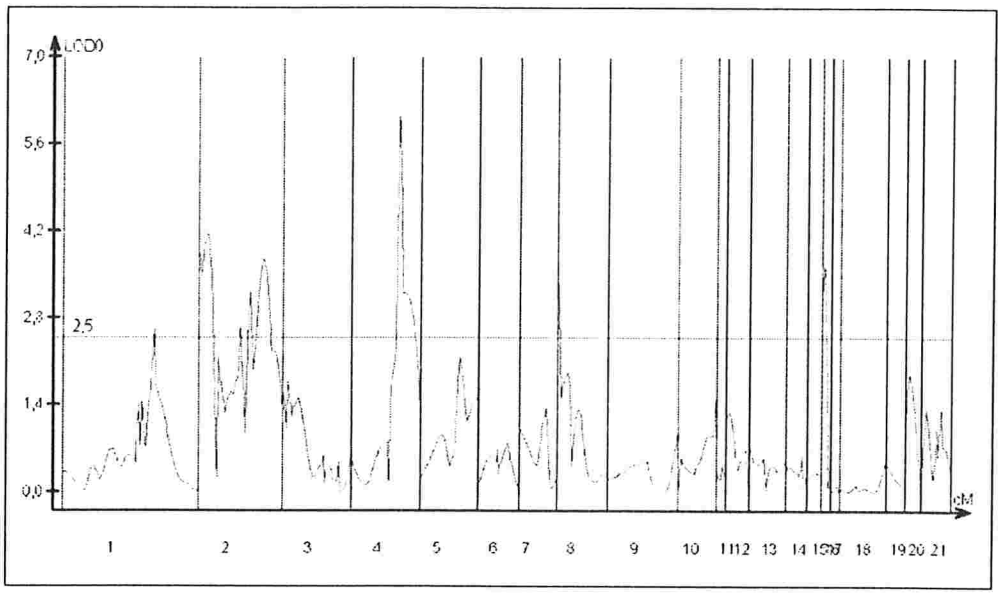


Figura 3.7: Gráfico de Perfis da estatística *Lod score* para o mapeamento intervalar da Pressão Sistólica Pós-Sal para os ratos F2 do projeto InCor.

sendo a proporção de hipóteses nulas H_0 verdadeiras, entre as hipóteses nulas rejeitadas, também chamada proporção de falsos positivos [Benjamini & Hocheberg, 1995], [Reiner *et al*, 2003].

3.3.4 Inferências via o Modelo Mistura de Normais

Sabe-se que um traço quantitativo definido em uma população em segregação pode seguir uma distribuição Mistura de Normais, onde as proporções de mistura são baseadas em leis de segregação Mendeliana [Jansen, 1992]. Observe os histogramas apresentados na Figura 3.8 [Balding *et al.*, 2003].

Note que o primeiro gráfico da Figura 3.8 refere-se a uma distribuição bimodal apropriada para uma população DH (duplo haplótipo; Q_1Q_1 e Q_2Q_2) em segregação. O segundo gráfico mostra uma distribuição unimodal apropriada, por exemplo, para uma população F2 em segregação, onde o gene é codominante. Já o terceiro gráfico apresenta uma distribuição unimodal com cauda pesada à direita, apropriada a uma população F2 em segregação, onde o gene principal é recessivo, isto é, o valor esperado do traço é o mesmo para os genótipos Q_1Q_1 e Q_1Q_2 e aumenta na presença

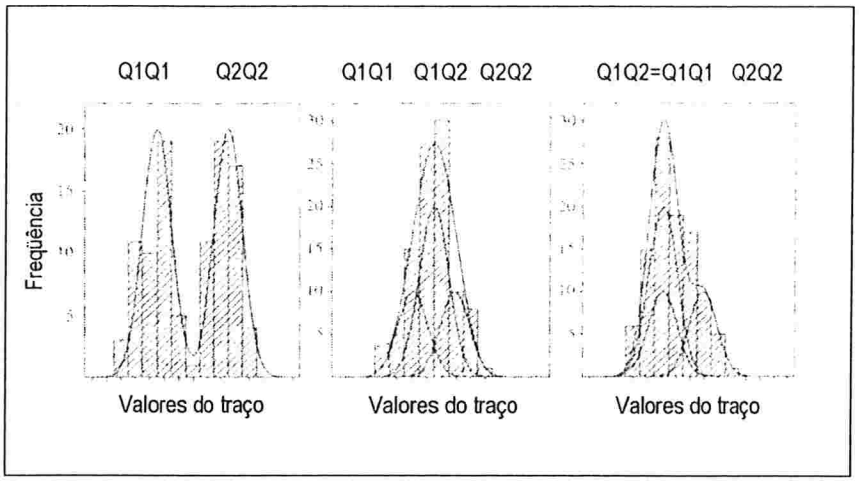


Figura 3.8: Exemplos de Histogramas que apresentam dados que seguem um modelo de Mistura de Normais.

de duas cópias do alelo Q_2Q_2 por exemplo.

O modelo de mistura proposto por Jansen (1992) para um traço y é dado por:

$$\begin{aligned} f(y_j) &= f_{Q_1Q_1}(y_j; G_{Q_1Q_1}) + f_{Q_1Q_2}(y_j; G_{Q_1Q_2}) + f_{Q_2Q_2}(y_j; G_{Q_2Q_2}) \\ &= P(Q_1Q_1) f_{Q_1Q_1}(y_j/G_{Q_1Q_1}) + P(Q_1Q_2) f_{Q_1Q_2}(y_j/G_{Q_1Q_2}) \\ &\quad + P(Q_2Q_2) f_{Q_2Q_2}(y_j/G_{Q_2Q_2}) \\ &= P(Q_1Q_1)\phi_{Q_1Q_1}(y_j; \mu_{Q_1Q_1}, \sigma^2) + P(Q_1Q_2)\phi_{Q_1Q_2}(y_j; \mu_{Q_1Q_2}, \sigma^2) \\ &\quad + P(Q_2Q_2)\phi_{Q_2Q_2}(y_j; \mu_{Q_2Q_2}, \sigma^2) \end{aligned} \tag{3.13}$$

onde ϕ é a função de densidade da Normal expressa por:
$\phi(y; \mu; \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-(y-\mu)^2/2\sigma^2)$.

Considerando uma população F2 e um loco codominante, os pesos em cada mistura são esperados obedecer 1/4, 1/2 e 1/4 para as categorias genotípicas Q_1Q_1, Q_1Q_2 e Q_2Q_2, respectivamente.

A função de verossimilhança do modelo mistura de Normais para a amostra

aleatória de indivíduos F2 sob estudo é o produto, sob independência, das contribuições individuais de verossimilhanças dada por:

$$L = \prod_{j=1}^{n} f(y_j),$$

onde $f(y_j)$ é dada em 3.13.

Para o caso da Mistura de Normais, os estimadores de máxima-verossimilhança (EMV) para μ, a e d não têm solução explícita e devem ser obtidos por métodos numéricos, como por exemplo o algoritmo EM (Esperança-Maximização). Este procedimento tem sido amplamente utilizado em análises de dados genômicos porque é poderoso especialmente quando a amostra possui dados incompletos.

Cada iteração do algoritmo EM envolve dois passos: o passo de Esperança (E) e o passo de Maximização (M). O passo da Esperança estima as estatísticas para os dados completos baseando-se nos dados observados incompletos. O passo de Maximização utiliza os dados completos estimados para obter a estimação da máxima-verossimilhança. Depois de inúmeras iterações dos passos E e M, o estimador de máxima-verossimilhança é encontrado quando a iteração converge. Ver, por exemplo, [Liu, 1998], [Jansen, 1992], [Zeng, 1994].

Os estimadores dos parâmetros sob o modelo de mistura também podem ser calculados por meio do algoritmo ECM (Esperança-Maximização Condicional) [Meng & Rubin, 1993], [Jiang & Zeng, 1995], que consiste em uma versão do algoritmo EM que converge mais rapidamente.

Vale ressaltar que as estatísticas LRT e *Lod score*, introduzidas anteriormente, também podem ser definidas para o teste dos efeitos do QTL, no contexto da verossimilhança definida em termos de Modelos de Mistura de Normais.

3.3.5 Vantagens e desvantagens do Mapeamento Intervalar

Ao comparamos com o modelo de regressão para genótipos conhecidos, o método de mapeamento intervalar tem inúmeras vantagens, incluindo [Zeng, 1994]:

- É possível inferir a posição provável do QTL;

- As localizações e os efeitos estimados dos QTL's tendem a ser assintoticamente não-viesados se apenas 1 QTL estiver segregando no cromossomo;
- O método requer menos indivíduos que o modelo de regressão para genótipos conhecidos.

Contudo, este método também apresenta alguns problemas como:

- Nos casos em que não há um QTL presente no intervalo, mas há um QTL em alguma região próxima no cromossomo, é possível que o valor da estatística no intervalo exceda o valor crítico definido. Neste caso, não se trata de um teste intervalar, que poderia distingüir quando há ou não um QTL dentro do intervalo, pois o teste pode não ser independente dos efeitos de outros QTL's que estiverem fora da região definida;
- Se houver mais que um QTL em um cromossomo, o teste estatístico da posição que está sendo testada pode ser afetado por todos os outros QTL's e as posições e os efeitos estimados dos QTL's identificados por este método provavelmente serão viesados;
- Não é eficiente utilizar somente 2 marcadores por vez para fazer o teste, já que a informação de outros marcadores disponíveis não é utilizada.

Os problemas comentados incentivaram o desenvolvimento de modelos mais gerais, que fossem capazes de contorná-los.

Capítulo 4

Modelos Gerais

Modelos como o de regressão linear simples para avaliar o efeito dos marcadores, ou de Mapeamento Intervalar para avaliar o efeito de QTL's são valiosos e, indiscutivelmente, contribuem muito na busca pela identificação de genes controladores de um determinado traço. No entanto, tais modelos apresentam algumas limitações que prejudicam as estimativas de efeito e localização dos QTL's.

Motivados pela busca de maior precisão tanto na estimação dos efeitos quanto na localização dos QTL's, alguns autores propuseram modelos mais gerais que serão tratados neste capítulo.

4.1 Mapeamento Intervalar Composto

Considere que uma determinada análise de Mapeamento Intervalar tenha apresentado um gráfico de perfis da estatística *Lod score* semelhante ao da Figura 4.1. Considere também que o traço em estudo não seja controlado por apenas 1 gene, mas por 2, e as posições reais do QTL 1 e do QTL 2 sejam próximas como indicado na mesma Figura.

Curiosamente, o gráfico de perfis da Figura 4.1 apresenta 3 picos: os dois mais baixos, que estariam indicando corretamente a existência de QTL's relacionados ao traço naquela região, e um pico mais alto entre os dois mais baixos que, aparentemente, não faria sentido, uma vez que na realidade os dois genes que controlam o traço não estão localizados onde o pico mais alto está indicando.

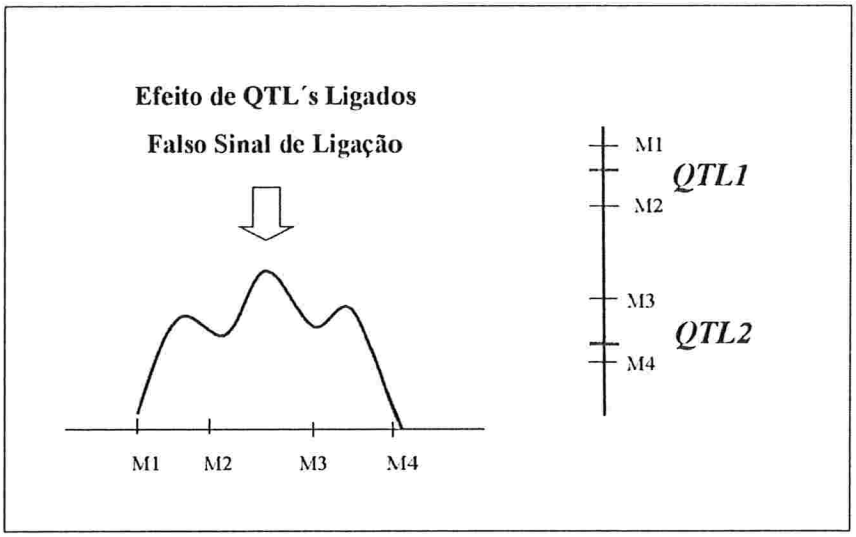


Figura 4.1: Exemplo de Gráfico de Perfis referente a um estudo de Mapeamento Intervalar cujo traço é controlado por dois genes em um mesmo cromossomo.

Sabe-se que o Mapeamento Intervalar não prevê a modelagem de mais que 1 QTL, e nos casos em que existem QTL's ligados (próximos), como no exemplo apresentado, o resultado indicado por este método deixa de ser satisfatório.

Pode-se concluir então, que tanto o método de análise com um único marcador quanto o de Mapeamento Intervalar são viesados quando múltiplos QTL's estão ligados ao marcador/intervalo considerado. Para solucionar este problema, Zeng (1993) propôs o *Mapeamento Intervalar Composto*, que consiste em considerar não apenas um intervalo entre 2 marcadores, mas também alguns outros marcadores para controlar o que chamamos de *background* genético[1] através de análises de regressão múltipla.

Quando testamos um intervalo para um QTL, nós gostaríamos que este teste fosse realizado independentemente dos efeitos de outros possíveis QTL's presentes em outras regiões do cromossomo. Caso isto fosse possível, o interessante é que poderíamos simplificar o processo de mapeamento de múltiplos QTL's de uma dimensão múltipla para um processo de mapeamento unidimensional, já que o teste para cada intervalo seria independente. Assim surgiu o modelo para Mapeamento Intervalar Composto, definido segundo a equação 4.1.

[1] *Background* genético: é o conjunto de muitos genes, cada um com pequeno efeito.

$$y_j = \mu + aX_{aj} + dX_{dj} + \sum_{i \neq j,j+1}^{M} (b_i x_{ij}) + e_j, \quad (4.1)$$

para j=1,2,...,n, onde,
y_j é o valor do traço do j-ésimo indivíduo F2;
μ, a e d são a média geral do traço, efeitos aditivo e de dominância do QTL, respectivamente;
X_{aj} e X_{dj} são os valores preditos do QTL calculados por meio das informações genotípicas dos marcadores flanqueadores do suposto QTL e por meio das frações de recombinação associadas (como descritos anteriormente);
b_i é o coeficiente de regressão associado ao i-ésimo marcador;
x_{ij} é a variável indicadora do efeito do i-ésimo marcador no j-ésimo indivíduo. Se o marcador é codominante, a variável recebe as categorias 0, 1 e 2;
M é o número total de marcadores moleculares incluídos no ajuste;
e_j é o resíduo para o j-ésimo indivíduo.

Assim como nas análises anteriores, obtemos os valores X_{aj} e X_{dj}, testamos os efeitos do QTL e obtemos estimativas destes efeitos. Medidas de herdabilidade do traço também podem ser calculadas.

Os 3 primeiros termos da equação 4.1 estimam o efeito do QTL no intervalo em questão, já o 4o. termo (somatória) refere-se ao ajuste para remoção dos efeitos de outros QTL's ligados.

Nota-se que a idéia principal do modelo 4.1 é ganhar precisão, isto é, selecionar um conjunto de marcadores (covariáveis) para formar parte do *background* genético e assim diminuir o resíduo do modelo. É essencial comentar que a principal dificuldade deste modelo reside justamente em encontrar o subconjunto de marcadores genéticos ideal para o ajuste e, para tanto, é comum a utilização de métodos de seleção de modelos por *Stepwise*.

O método *Stepwise* consiste na seleção de variáveis para inclusão em um modelo de regressão que inicia selecionando o melhor preditor da variável dependente. Outras variáveis independentes são adicionadas ao modelo de acordo com o poder ex-

plicativo incremental que podem acrescentar, assim como estas podem ser eliminadas se seu poder preditivo cair para um nível não significante quando outra variável independente for acrescentada no modelo. Para maiores detalhes, ver [Hair *et al.*, 1998], [Draper & Smith, 1981], [Neter *et al.*, 1996].

Zeng sugere a inclusão dos marcadores adjacentes aos marcadores flanqueadores no *background* genético como forma de validar que o efeito que estamos testando é realmente do QTL localizado na específica posição entre os marcadores flanqueadores e não de possíveis QTL's vizinhos [Zeng, 1993].

Contudo, o Método Intervalar Composto ainda apresenta algumas limitações como [Zeng *et al.*, 1999]:

- A análise pode ser afetada por uma distribuição desigual de marcadores no genoma, o que significa que um teste estatístico em uma região rica em marcadores moleculares pode não ser comparável àquela pobre em marcadores;
- Dificuldade em estimar a contribuição conjunta de múltiplos QTL's ligados para a variância genética;
- Não é possível extender diretamente este método para modelos mais complexos;
- A utilização de marcadores fortemente ligados entre si definidos como covariáveis no *background* genético pode reduzir o poder do teste de detecção do QTL.

Veremos a seguir que tais limitações serviram como incentivo para o desenvolvimento de modelos mais sofisticados.

4.2 Modelo para múltiplos QTL's (Epistasia)

Modelos para múltiplos QTL's apresentam uma melhora significativa em relação aos modelos para apenas um único QTL porque são capazes de isolar efeitos de QTL's ligados no mesmo cromossomo, bem como detectar QTL's em interação, que usualmente não são detectados por modelos mais simples [Sen & Churchill, 2001].

Os métodos apresentados até o momento não são adequados para estudar toda a "arquitetura genética" de traços quantitativos, pois muitos parâmetros (como aqueles necessários quando há epistasia), ainda não foram levados em conta nestes modelos [Zeng *et al.*, 1999].

Para mapear múltiplos QTL's é necessário utilizar a informação de todo o genoma e, além disso, inferir possíveis padrões de epistasia. Isto significa que a busca e o mapeamento de múltiplos QTL's epistáticos devem ser feitos em múltiplos intervalos simultaneamente, e tal fato motivou o desenvolvimento do **Mapeamento Intervalar Múltiplo**. Vale ressaltar que, quando comparado com os outros métodos como o Mapeamento Intervalar e o Mapamento Intervalar Composto, o Mapeamento Intervalar Múltiplo tende a ser mais poderoso e preciso na detecção de QTL's [Kao *et al.*, 1999].

Portanto, a idéia do Mapeamento Intervalar Múltiplo é ajustar diretamente em um único modelo tanto os efeitos dos supostos QTL's quanto os efeitos epistáticos associados, com o intuito de facilitar a busca, o teste e as estimativas de posições, efeitos e interações entre os QTL's.

Inicialmente, vamos apresentar o modelo aditivo para 2 QTL's:

$$y_j = \mu + \underbrace{a_1 X_{a_{1j}} + d_1 X_{d_{1j}}}_{QTL1} + \underbrace{a_2 X_{a_{2j}} + d_2 X_{d_{2j}}}_{QTL2} + e_j, j = 1, 2, ..., n$$

onde,

a_1 é o efeito aditivo do QTL1;

d_1 é o efeito de dominância do QTL1;

a_2 é o efeito aditivo do QTL2;

d_2 é o efeito de dominância do QTL2;

$X_{a_{1j}}$, $X_{a_{2j}}$, $X_{d_{1j}}$ e $X_{d_{2j}}$ são os valores preditos do QTL1 e do QTL2 calculados, independentemente, por meio das informações genotípicas dos correspondentes marcadores flanqueadores dos supostos QTL's e por meio das frações de recombinação associadas;

e_j é o resíduo aleatório do modelo.

Nota-se que o modelo acima não considera que possa existir epistasia entre o QTL1 e o QTL2. O problema é que se a epistasia entre os QTL's existe e não está sendo considerada, a estimação dos efeitos marginais e as respectivas posições dos QTL's podem estar viesadas caso os QTL's epistáticos estejam ligados [Kao & Zeng, 2002].

A seguir, apresentamos o modelo com epistasia para 2 QTL's, incluindo o ajuste para marcadores do *background* genético:

$$\begin{aligned} y_j = \mu + \underbrace{a_1 X_{a_{1j}} + d_1 X_{d_{1j}}}_{QTL1} + \underbrace{a_2 X_{a_{2j}} + d_2 X_{d_{2j}}}_{QTL2} \\ + \underbrace{i_{aa} X_{a_{1j}} * X_{a_{2j}}}_{\text{Epistasia efeito aditivo}} + \underbrace{i_{dd} X_{d_{1j}} * X_{d_{2j}}}_{\text{Epistasia efeito de dominância}} \\ + \underbrace{i_{ad} X_{a_{1j}} * X_{d_{2j}} + i_{da} X_{d_{1j}} * X_{a_{2j}}}_{\text{Epistasia ef. aditivo com ef. de dominância}} + \sum_{i}^{M} b_i x_{ij} + e_j, \end{aligned} \tag{4.2}$$

onde,
a_1 e d_1 são os efeitos aditivo e de dominância do QTL1;
a_2 e d_2 são os efeitos aditivo e de dominância do QTL2;
$X_{a_{1j}}$, $X_{a_{2j}}$, $X_{d_{1j}}$ e $X_{d_{2j}}$ são os valores preditos do QTL1 e QTL2 calculados, independentemente, por meio das informações genotípicas dos correspondentes marcadores flanqueadores dos supostos QTL's e por meio das frações de recombinação associadas;
i_{aa}, i_{dd}, i_{ad}, i_{da} são os efeitos epistáticos entre os QTL's 1 e 2 definidos em termos dos componentes aditivos e de dominância dos QTL's 1 e 2 e do aditivo de um e de dominância do outro, respectivamente;
M é o número total de marcadores moleculares incluídos no *background*;
b_i é o coeficiente de regressão associado ao i-ésimo marcador ;
x_{ij} é a variável indicadora do efeito do i-ésimo marcador;
e_j é o resíduo aleatório do modelo.

Para os procedimentos inferenciais, assume-se que os erros deste modelo podem seguir as premissas clássicas, isto é, são independentes e identicamente distribuídos

seguindo o modelo de probabilidade Normal tal que, $e_j \sim N(0;\sigma^2)$, ou o modelo Mistura de Normais.

Para o caso de Mistura de Normais, utiliza-se o algoritmo ECM (Esperança-Maximização Condicional) para a obtenção das estimativas de máxima-verossimilhança.

Posteriormente, podemos avaliar o gráfico de perfis da estatística *Lod score* para testarmos a significância dos efeitos principais e de interação dos QTL's envolvidos no modelo.

Veja que nos deparamos com o seguinte problema: como realizar a busca por múltiplos QTL's e efeitos de interação? A estatística *Lod score* pode ser definida de várias maneiras:

- $Lod = log_{10}\left(\frac{L(QTL1, QTL2)}{L(\mu)}\right)$;
- $Lod = log_{10}\left(\frac{L(QTL1, QTL2)}{L(QTL1)}\right)$;
- $Lod = log_{10}\left(\frac{L(QTL1, QTL2)}{L(QTL2)}\right)$;
- $Lod = log_{10}\left(\frac{L(QTL1, QTL2, QTL1 * QTL2)}{L(QTL1)}\right)$;
- $Lod = log_{10}\left(\frac{L(QTL1, QTL2, QTL1 * QTL2)}{L(QTL1), L(QTL2)}\right)$.

No primeiro caso, calculamos a estatística razão de verossimilhanças comparando os modelos de regressão que incluem os dois QTL's simultaneamente com o modelo esporádico (sem efeitos genéticos). Este é o procedimento de busca simultâneo, que apresenta um problema sério de envolver um número muito grande de combinações 2 a 2 para se testar. No segundo caso, trata-se do procedimento condicional de busca por múltiplos QTL's, calculamos a estatística razão de verossimilhanças considerando o modelo que inclui ambos os QTL's e aquele na presença apenas do QTL1, isto é, o efeito do segundo QTL é avaliado condicionalmente, na presença do QTL1. Tal problema de seleção de modelos se estende para o teste de efeitos de interação, que pode ser definido, por exemplo, via as duas últimas estatísticas *Lod score* indicadas.

Infelizmente não se tem uma resposta a priori de qual estatística *Lod* é mais conveniente. De qualquer modo, é importante frisar que o método condicional de busca é o mais utilizado. Trata-se de um procedimento de regressão com seleção de QTL's pelo método *Forward*, que consiste na inclusão de novas variáveis no modelo na presença de outras. Neste caso, o número de possíveis modelos competitivos a serem investigados é menor que o procedimento de busca simultânea por subconjuntos de QTL's.

4.3 Modelo com efeito de Pleiotropia

Como vimos na Figura 2.8 do capítulo 2, pleiotropia é a propriedade que certos genes apresentam de controlar mais de uma característica fenotípica simultaneamente. Não é difícil perceber que o fato de um mesmo gene estar associado à mais de uma característica, digamos duas, gera uma estrutura de covariância entre estas duas variáveis devido ao efeito do gene comum em questão.

Poucos estudos têm focado na busca e em uma maior compreensão dos QTL's responsáveis pelas interações entre genes e ambiente. As análises de múltiplos traços possuem diversas propriedades que são particularmente úteis quando estamos interessados na natureza biológica destas interações. Uma das razões é que, na prática, muitos delineamentos experimentais são baseados em comparações longitudinais de um fenótipo considerando um mesmo conjunto de indivíduos e seus correspondentes genótipos de marcadores. Como exemplo, pense na pressão sistólica de um indivíduo sendo avaliada antes e depois da exposição a uma dieta rica em sal. O genótipo do indivíduo não varia ao longo da condição de exposição, mas seu fenótipo sim. Devido à estrutura de covariância entre estes traços (pressão sistólica antes e depois da dieta de sal), a análise de múltiplos traços é mais apropriada na identificação dos QTL's responsáveis pela resposta ao tratamento e/ou sensibilidade ao sal do que a análise para cada traço realizada separadamente. A Figura 4.2 ilustra o esquema.

Suponha que vamos estudar a pressão sanguínea sistólica antes da exposição ao sal, denotada por SBP, e a pressão sanguínea sistólica pós-sal, denotada por SBPS. A busca por QTL's relacionados aos fenótipos citados pode ser feita para cada traço separadamente, baseada no Método Intervalar Composto [Haley & Knott, 1992]

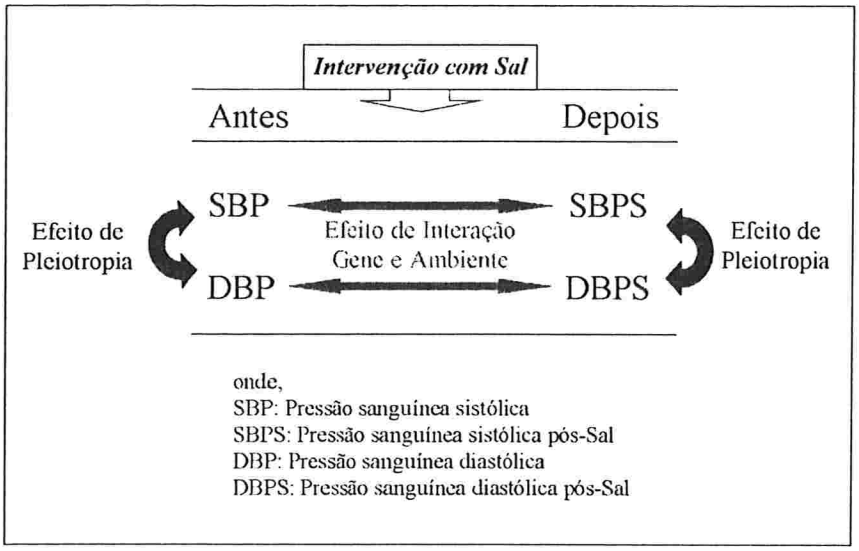


Figura 4.2: Exemplo esquemático do efeito de interação entre gene e ambiente e do efeito de pleiotropia.

[Zeng, 1994], ou conjuntamente, utilizando uma versão deste método proposta por Jiang & Zeng (1995) que leva em consideração modelos de regressão multivariada.

Para introduzir o modelo de regressão multivariada para finalidade de mapeamento genético, seja y_{SBP_j} e y_{SBPS_j} os valores para os fenótipos SBP e SBPS para o j-ésimo indivíduo, respectivamente. Se os fenótipos forem correlacionados, a análise univariada para cada um dos traços separadamente não levará em conta tal informação adicional de correlação. Uma análise bivariada, na qual a correlação fenotípica é explicitamente modelada, será capaz de explorar melhor as informações contidas nos dados e aumentará tanto o poder de detecção do QTL quanto a precisão na estimação dos parâmetros [Jiang & Zeng, 1995]. Neste contexto, a covariância entre os fenótipos:

$$\mathrm{Cov}(y_{1j}, y_{2j}) = \sigma_{12} = \rho_{12}\sigma_1\sigma_2$$

pode ser introduzida na análise.

De maneira geral, seja y_{kj} o valor do k-ésimo traço avaliado no j-ésimo indivíduo. O modelo considerando um QTL (fixado entre dois marcadores flanqueadores) afetando múltiplos traços, digamos dois para facilidade de formulação, é dado por:

$$
\begin{aligned}
y_{1j} &= \mu_1 + a_1^* X_{a_{1j}}^* + d_1^* X_{d_{1j}}^* + \sum_{i=1}^{M}(b_{1ij} X_{ij}) + e_{1j}; \\
y_{2j} &= \mu_2 + a_2^* X_{a_{2j}}^* + d_2^* X_{d_{2j}}^* + \sum_{i=1}^{M}(b_{2ij} X_{ij}) + e_{2j}.
\end{aligned}
\tag{4.3}
$$

onde,

j=1, 2,...,n;

$e_{kj} \sim N_2(0_{(2x1)}; V_{(2x2)})$ sob premissas clássicas;

μ_k é a média geral de resposta para o traço k;

$X_{a_{kj}}^*$ e $X_{d_{kj}}^*$ são as variáveis preditoras dos efeitos aditivo e de dominância do QTL, respectivamente, sobre o traço k, avaliada no indivíduo j;

a_k^* e d_k^* são os coeficientes de regressão associados aos efeitos aditivo e de dominância do QTL sobre o traço k. O asterisco é usado para diferenciar dos parâmetros do modelo de epistasia;

b_{kj} e X_{ij} são os coeficientes de regressão e as variáveis associadas ao efeito aditivo dos M marcadores incluídos no *background* genético, assumindo o mesmo conjunto de marcadores para ambos os traços;

e_{kj} são os resíduos do traço k para o j-ésimo indivíduo.

Pode-se assumir que $e_{kj} \sim N_2(0_{(2x1)}; V_{(2x2)})$ sob premissas clássicas, sendo a matriz de variância-covariância V definida por:

$$
\mathbf{V} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}
$$

Dado um suposto QTL, sob o mapeamento de múltiplos traços, as hipóteses a serem testadas são:

$H_0 : a_1^* = d_1^* = a_2^* = d_2^* = 0$

H_1 : pelo menos um dos efeitos é diferente de zero

Quando o mapeamento conjunto para um QTL influenciando dois traços é indicado, outros testes estatísticos podem ser feitos para testar se o QTL tem efeito em apenas um dos traços ou em ambos:
$H_{10} : a_1^* = d_1^* = 0, a_2^* \neq 0, d_2^* \neq 0$
$H_{11} : a_1^* \neq 0, d_1^* \neq 0, a_2^* \neq 0, d_2^* \neq 0$
e
$H_{20} : a_1^* \neq 0, d_1^* \neq 0, a_2^* = d_2^* = 0$
$H_{21} : a_1^* \neq 0, d_1^* \neq 0, a_2^* \neq 0, d_2^* \neq 0$

A flexibilidade do modelo de mapeamento genético multivariado se estende, para investigar o efeito de interação entre gene e ambiente, por exemplo, para testar efeitos genéticos que predispõem à sensibilidade a sal. Para este fim, usando o exemplo anterior citado, as hipóteses em teste podem ser definidas, por exemplo, como:
$H_{30} : a_{SBP} = a_{SBPS}, d_{SBP} = d_{SBPS}$
$H_{31} : a_{SBP} \neq a_{SBPS}, d_{SBP} = d_{SBPS}$
onde testa-se o efeito aditivo do QTL pleiotrópico na ausência de efito de dominância, ou, mais especificamente, se o efeito aditivo do QTL depende da presença do sal.

Se o modelo Mistura de Normais for assumido, os parâmetros de estimação do modelo podem ser obtidos pelos algoritmos ECM e as estatísticas de razão de verossimilhanças também podem ser calculadas para os testes do efeito do QTL no contexto multivariado. Tais procedimentos são bastante complexos e o cálculo da verossimilhança no contexto multivariado muitas vezes é feito utilizando propriedades de decomposição espectral [Mangin *et al.*, 1998].

É importante destacar algumas características importantes da análise conjunta (para múltiplos traços) de mapeamento de QTL's em relação à análise de um único traço [Jiang & Zeng, 1995]:

- Se a covariância entre dois traços for nula ($\rho = 0$), isto é, tratarem-se de dois traços não correlacionados, a estatística razão de verossimilhanças da análise conjunta (LR_{12}) será no mínimo igual a soma das estatísticas calculadas separadamente ($LR_{12} \geq LR_1 + LR_2$);

- Se o coeficiente aditivo de um dos traços for igual a zero (por exemplo, $a_2^* = 0$), temos que a estatística razão de verossimilhanças da análise conjunta será maior ou igual à estatística calculada para o outro traço isoladamente ($LR_{12} \geq LR_1$), se os traços forem correlacionados;

- A estatística razão de verossimilhanças da análise conjunta será sempre maior ou igual ao valor máximo encontrado para as estatísticas calculadas separadamente ($LR_{12} \geq máximo[LR_1, LR_2]$);

- Se a multiplicação dos efeitos aditivos ou de dominância (por exemplo, $a_1^* a_2^*$ ou $d_1^* d_2^*$) com a correlação (ρ) entre eles resultar em um sinal negativo, isto é, $(\rho a_1^* a_2) < 0$, então a estatística razão de verossimilhanças da análise conjunta é maior que a soma das estatísticas calculadas separadamente($LR_{12} > LR_1 + LR_2$). Neste caso, o poder da análise conjunta para mapeamento de QTL é maior que o de qualquer análise separada. Trata-se da situação mais favorável para a análise conjunta.

Vale comentar que, sob as premissas clássicas para a distribuição dos resíduos, um método de mapeamento de QTL's para múltiplos traços utilizando a teoria de Mínimos Quadrados no contexto matricial foi descrito. Ver [Knott & Haley, 2000].

Capítulo 5

Aplicações

Neste trabalho vamos utilizar dados de um cruzamento F2 entre ratos normotensos e ratos espontaneamente hipertensos, coletados no Laboratório de Genética e Cardiologia Molecular do InCor (Instituto do Coração, São Paulo).

Este banco de dados foi descrito e analisado parcialmente por Schork et al. (1995). O estudo examinou variações da pressão sanguínea antes e depois de exposição ao sal (NaCl) em 221 ratos, sendo que estes animais foram genotipados em 182 marcadores moleculares espalhados pelos 21 cromossomos. O banco de dados contém ainda informações para 23 traços quantitativos, sendo que nossa análise se concentrará na pressão sistólica.

As análises apresentadas neste capítulo foram realizadas utilizando os recursos do aplicativo *QTL Cartographer*, disponível na Internet no endereço *http://statgen.ncsu.edu/qtlcart*.

5.1 Mapa de Marcadores

Inicialmente, utilizando a metodologia descrita na seção 3.2, obtivemos os resultados apresentados na Tabela 5.1 para a pressão sistólica pós-sal, denominada SBPS com o objetivo de pesquisar o quão informativo é o marcador molecular. A coluna 1 indica o cromossomo, a coluna 2 apresenta o código do marcador, as colunas 3 e 4 correspondem às estimativas dos parâmetros b_0 e b_1, respectivamente, do modelo linear apresentado em 3.1, sendo que b_0 é o efeito geral e b_1 é o efeito aditivo do mar-

cador. As colunas 5 e 6 trazem os valores da estatística razão de verossimilhanças e a estatística F, respectivamente. A coluna 7 indica o valor do nível descritivo. Vale ressaltar que apresentaremos apenas os valores para aqueles marcadores que apresentaram um p-valor significante ao menos no nível de 5%.

Os asteriscos indicam o nível de significância do teste F: 5%, 1%, 0,1%, 0,01% estão indicados por um, dois, três ou quatro asteriscos, respectivamente.

Em seguida, apresentamos o gráfico na Figura 5.1 que contém o perfil da estatística *Lod score* para o mapeamento da pressão sistólica pós-sal com os ratos do InCor. Note que trata-se do mesmo gráfico apresentado na escala razão de verossimilhanças na Figura 3.1.

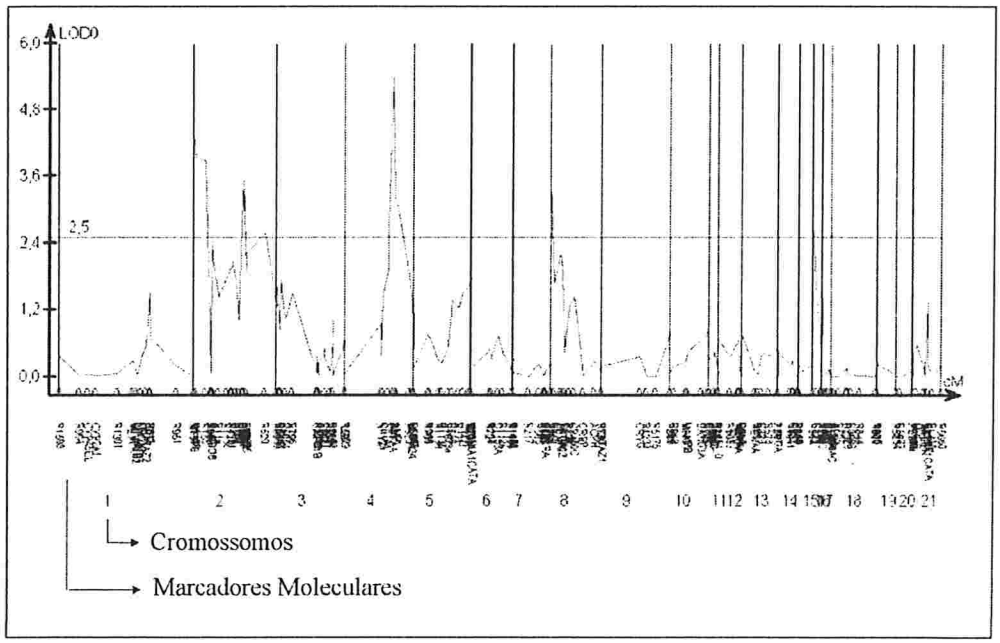


Figura 5.1: Gráfico de perfis da estatística *Lod score* para o mapeamento da pressão sistólica pós-Sal com os ratos F2 do InCor por meio da Análise de Regressão Simples.

Veja que o gráfico 5.1 apresenta o caso em que estamos testando quais marcadores parecem estar associados a variações na média do fenótipo (SBPS). Há cromossomos cujo número de marcadores é pequeno, por exemplo, os cromossomos 15, 19 e 20 possuem 2, 2 e 3 marcadores, respectivamente. Como na análise para genótipos

Cromossomo	Marcador	b_0	b_1	$-2ln(L_0/L_1)$	$F_{(1.n-2)}$	pr(F)	
1	M12	146,870	4,528	7,025	7,073	0,008	**
2	M1	147,408	7,457	19,792	20,518	0,000	****
2	M2	147,331	7,216	18,040	18,626	0,000	****
2	M3	147,693	6,872	17,999	18,583	0,000	****
2	M5	147,338	5,339	11,255	11,442	0,001	***
2	M6	147,423	5,328	9,874	10,007	0,002	**
2	M7	147,151	4,173	6,749	6,791	0,010	**
2	M8	147,172	4,234	6,552	6,590	0,011	*
2	M9	147,107	4,697	8,202	8,280	0,004	**
2	M10	147,081	4,923	9,162	9,270	0,003	**
2	M11	147,138	5,016	9,539	9,659	0,002	**
2	M12	147,099	4,672	7,994	8,067	0,005	**
2	M13	146,922	3,481	4,623	4,630	0,033	*
2	M14	147,402	6,668	15,442	15,849	0,000	****
2	M15	147,537	6,890	15,541	15,954	0,000	****
2	M16	147,353	6,273	13,797	14,108	0,000	***
2	M17	147,383	6,303	14,100	14,427	0,000	***
2	M18	147,397	6,585	16,247	16,707	0,000	****
2	M19	147,269	4,873	8,728	8,822	0,003	**
2	M20	147,288	5,318	10,190	10,334	0,002	**
2	M21	147,288	5,318	10,190	10,334	0,002	**
2	M22	147,633	5,744	11,880	12,094	0,001	***
2	M23	147,144	3,856	5,794	5,817	0,017	*
3	M1	147,009	4,400	7,425	7,483	0,007	**
3	M3	147,108	4,507	8,034	8,108	0,005	**
3	M4	147,256	3,740	4,637	4,644	0,032	*
3	M5	146,923	4,495	6,881	6,926	0,009	**
3	M15	147,265	-3,600	4,646	4,652	0,032	*
3	M16	146,886	-3,384	4,398	4,402	0,037	*
4	M2	147,114	3,423	4,367	4,370	0,038	*
4	M4	147,027	4,451	6,845	6,890	0,009	**
4	M5	146,858	4,895	8,852	8,950	0,003	**
4	M6	146,598	5,927	12,313	12,548	0,000	***
4	M7	146,331	8,641	24,859	26,073	0,000	****
4	M8	147,139	6,348	14,511	14,862	0,000	***
4	M9	147,199	6,239	14,378	14,722	0,000	***
4	M10	147,419	5,234	8,434	8,519	0,004	**
4	M11	147,217	4,021	5,916	5,941	0,016	*
4	M14	147,138	3,761	5,354	5,370	0,021	*
5	M8	146,878	4,242	6,357	6,391	0,012	*
5	M9	146,997	4,016	5,707	5,729	0,018	*
5	M10	147,071	4,580	7,073	7,123	0,008	**
5	M11	146,977	5,358	7,472	7,531	0,007	**
5	M12	147,075	4,887	8,196	8,274	0,004	**

Cromossomo	Marcador	b_0	b_1	$-2ln(L_0/L_1)$	$F_{(1,n-2)}$	pr(F)	
8	M1	146,827	6,699	14,981	15,360	0,000	***
8	M2	147,004	4,470	7,595	7,657	0,006	**
8	M3	147,021	5,341	10,217	10,362	0,001	**
8	M4	146,930	4,986	9,048	9,152	0,003	**
8	M6	146,970	4,030	5,745	5,767	0,017	*
8	M7	147,263	4,572	6,626	6,666	0,010	*
16	M1	147,107	5,491	12,055	12,278	0,001	***
16	M2	147,142	4,135	6,499	6,536	0,011	*
16	M5	146,897	4,775	6,469	6,506	0,011	*
21	M8	147,076	4,191	6,168	6,199	0,014	*

Tabela 5.1: Resultados da Análise de Regressão Simples (para dados genotípicos conhecidos) para o traço SBPS (pressão sistólica pós-sal).

conhecidos calculamos a estatística *Lod* apenas para os marcadores, tal fato dificulta a detecção de um QTL, pois os intervalos entre estes marcadores são muito grandes. Note que no gráfico desta análise, a curva é definida pela simples união dos valores das estatísticas determinadas pelos marcadores.

De qualquer forma, o gráfico mostra que os cromossomos 2, 4, 8 e 16 possuem marcadores que podem estar associados à pressão sistólica pós-sal. Lembramos que os sinais apresentados podem tratar-se do efeito do marcador ou de um efeito aparente.

5.2 Mapeamento Intervalar da Pressão Sistólica

Observe agora a Figura 5.2 que contém o perfil da estatística *Lod score* para o mapeamento intervalar da pressão sistólica pós-sal com os ratos do InCor.

Note que a tendência do gráfico mostrado na Figura 5.2 é a mesma que aquela da Figura 5.1, afinal o mapeamento intervalar é uma "suavização" do gráfico dos marcadores, apresentando os mesmos valores para as estatísticas calculadas nos pontos definidos para os próprios marcadores moleculares, como na análise para genótipos conhecidos. Contudo, o mapeamento intervalar é capaz de estimar efeitos

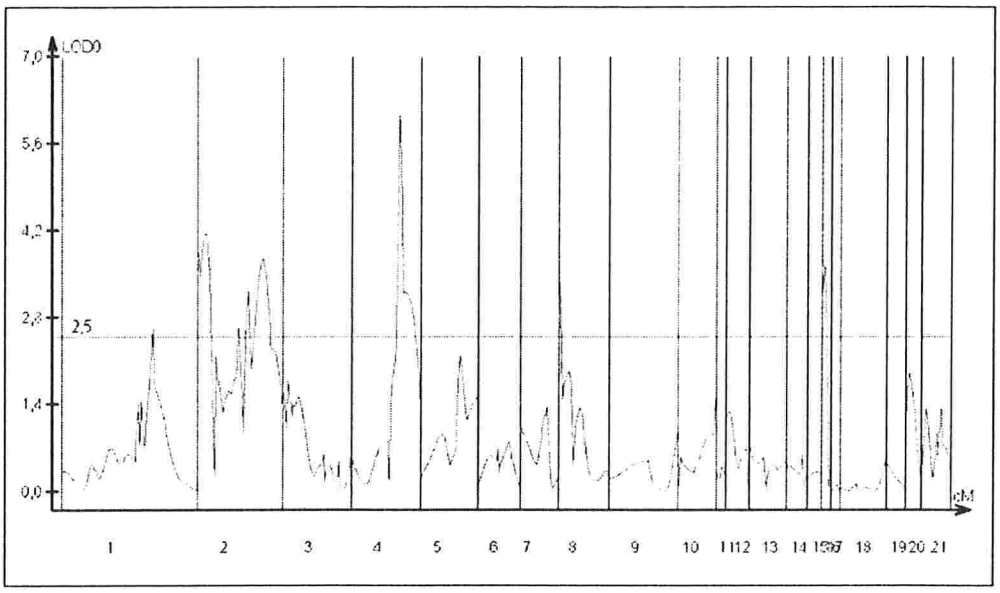


Figura 5.2: Gráfico de perfis da estatística *Lod score* para o mapeamento da pressão sistólica pós-Sal com os ratos F2 do InCor por meio da Análise de Mapeamento Intervalar.

e posições de QTL's fazendo uso de sub-intervalos entre os marcadores, o que não era possível na análise anterior. Deste modo, a curva deixa de ser definida pela simples união dos valores calculados para os marcadores e passa a ser definida pelos valores da estatística *Lod* obtidos por meio de cada sub-intervalo, neste caso definido em 1cM.

Veja que além de obtermos sinais de presença de QTL's nas regiões dos cromossomos 2, 4, 8 e 16, obtivemos também um sinal no cromossomo 1, que não aparecia na análise anterior.

5.3 Mapeamento Intervalar Composto

Vamos apresentar agora a Figura 5.3 que contém o perfil da estatística *Lod score* para o mapeamento intervalar composto da pressão sistólica pós-sal com os ratos do InCor.

Sabemos que a grande vantagem da Análise de Mapeamento Intervalar Composto

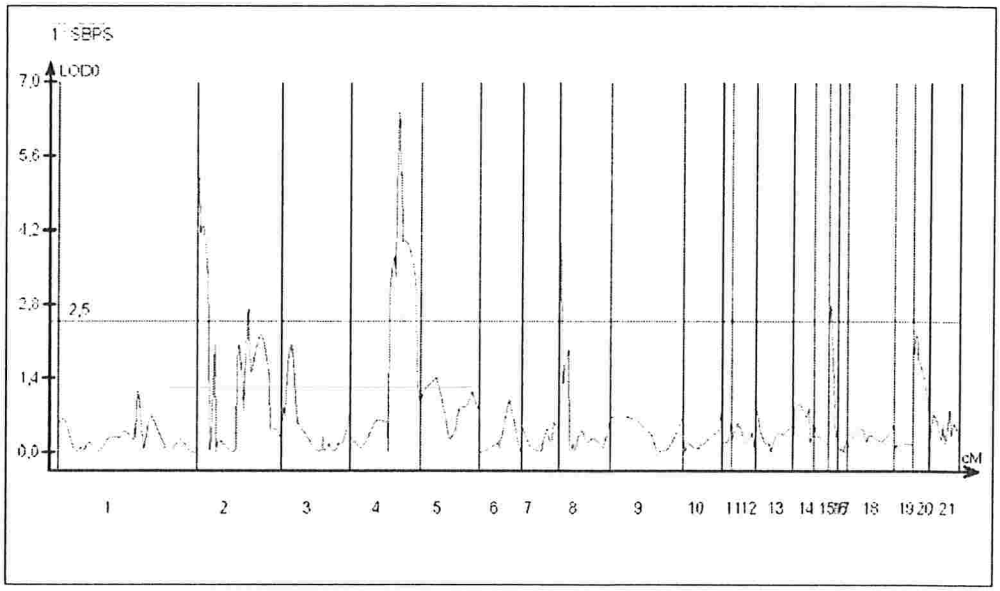


Figura 5.3: Gráfico de perfis da estatística *Lod score* para o mapeamento da Pressão Sistólica Pós-Sal com os ratos F2 do InCor por meio da Análise de Mapeamento Intervalar Composto.

é que pode-se selecionar um conjunto de marcadores (covariáveis) para formar parte do *background* genético, e assim ganhar precisão por diminuir o resíduo. Por meio deste método, temos os sinais de presença de QTL's novamente nos mesmos cromossomos apontados pela análise intervalar, exceto o sinal no cromossomo 1, que desaparece, possivelmente por tratar-se de um falso positivo.

Vale comentar que a seleção *Stepwise*, implementada no programa *Cartographer*, identificou 5 marcadores e os colocou no *background* genético no ajuste do modelo de mapeamento intervalar composto.

5.4 Modelo com Epistasia

O modelo com Epistasia foi ajustado por meio de um procedimento de regressão com seleção de QTL's pelo método *Forward*, já comentado na seção 4.2. É importante comentar que a busca pelos QTL's e efeitos de interação é implementada no *Cartographer* da seguinte forma: uma vez ajustados os efeitos principais, começa a busca pelos efeitos de interação condicionada aos efeitos principais encontrados no

modelo.

A Tabela 5.2 apresenta os resultados obtidos considerando os locos que deram maior sinal de efeito de interação. A coluna 1 indica o QTL ou o par de QTL's incluídos no modelo. A coluna 2 indica o tipo do efeito estimado: aditivo (a) ou de dominância (d). As colunas 3, 4 e 5 apresentam o cromossomo, o marcador em questão e a estimativa da posição, respectivamente. A coluna 6 traz o valor da estatística *Lod score*. Por fim, as colunas 7 e 8 apresentam as estimativas dos referidos efeitos e a proporção da variância do traço SBPS explicada por estes efeitos.

QTL (ou par de QTL's)	Tipo	Cromossomo	Marcador	Posição	Lod ($H_3 : H_0$)	Efeito	Efeito (%)
1	a	4	6	204,0100	4,59	12,6686	33,4
1	d	4	6	204,0100	0,42	-5,0446	3,8
2	a	2	2	22,0099	1,90	6,3145	9,1
2	d	2	2	22,0099	0,02	0,8663	0,1
3	a	8	1	0,0100	1,79	5,5146	6,0
3	d	8	1	0,0100	0,23	-2,6526	1,1
2x3	ad				0,001	0,3219	0,0
1x2	aa				0,4420	4,2674	2,0
2x3	aa				0,2550	3,4846	1,6

Tabela 5.2: Resultados da Análise do Modelo com Epistasia para o traço SBPS (pressão sistólica pós-sal).

Os efeitos de epistasia apresentados não foram significantes, entretanto foram os maiores encontrados. Deste modo conclui-se que o modelo de efeitos aditivos é apropriado para explicar o modo de regulação dos 5 QTL's sobre a pressão SBPS. Como já salientado no mapeamento de traços complexos é esperado que efeitos de interação entre genes (epistasia) estejam envolvidos. Acredita-se que até o momento não se identificou nenhum efeito de interação importante nem entre os 5 QTL's e nem mesmo entre eles e outros locos do *background* genético do animal F2 por limitação da metodologia de análise estatística utilizada ou por limitações do delineamento experimental usado (baixa densidade de marcadores, poucos indivíduos). Note que o estudo da interação entre 3 QTL's codominantes, por exemplo implicaria em classificar a amostra em 3^3 categorias genotípicas.

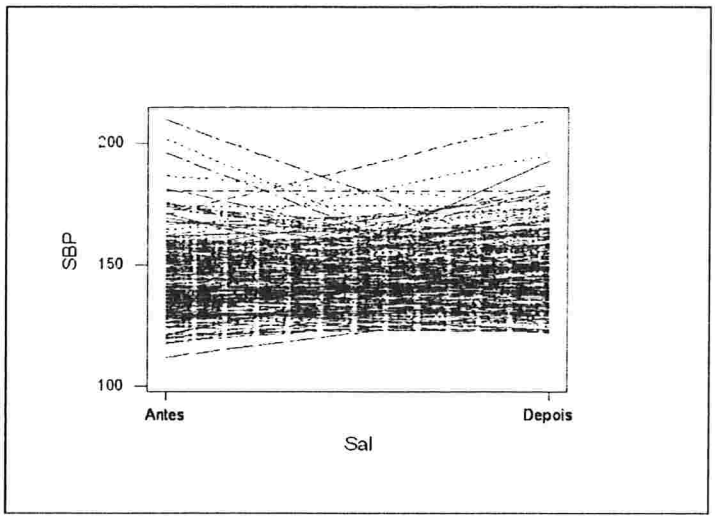


Figura 5.4: Perfis individuais dos ratos F2 do Projeto InCor quanto à pressão sanguínea sistólica antes e depois da exposição ao sal.

5.5 Modelo com Pleiotropia - Sensibilidade ao Sal

No caso do modelo com pleiotropia utilizaremos os traços SBP e SBPS, que são a pressão sanguínea sistólica antes e depois da exposição ao sal, respectivamente.

Inicialmente, vamos apresentar a matriz de variância-covariância V dos traços:

$$\mathbf{V} = \begin{pmatrix} 234,5904 & 99,6601 \\ 0,4233 & 236,2651 \end{pmatrix}$$

onde na diagonal estão indicadas a variância amostral do SBP e SBPS, respectivamente. Os valores indicados na diagonal superior e inferior referem-se à covariância e correlação amostral entre os traços, respectivamente.

Como é notado a variância dos traços é praticamente a mesma, porém, se levarmos em conta o mapeamento genético de cada traço separadamente obtemos que a herdabilidade de SBPS é muito maior do que a do traço SBP (dados não apresentados).

A Figura 5.4 apresenta o gráfico de perfis individuais para o SBP e para o SBPS.

Há indicação de uma sutil tendência da pressão sistólica pós-sal. Contudo verifica-se para alguns animais com valores de pressão basal muito alta houve uma queda drástica no valor da pressão pós-sal, o que pode ser explicado pela presença de algum gene regulador da pressão.

Essa resposta ao efeito do sal é vista por meio dos perfis individuais, que é o que se pretende modelar por meio do mapeamento conjunto.

Note que as análises univariadas para cada traço não consideram qualquer informação da história do animal sob a exposição ao sal.

Primeiramente, apresentaremos os gráficos dos perfis da estatística *Lod score* da análise marginal derivada do modelo multivariado com pleiotropia de cada traço, SBP e SBPS. Veja as Figuras 5.5 e 5.6.

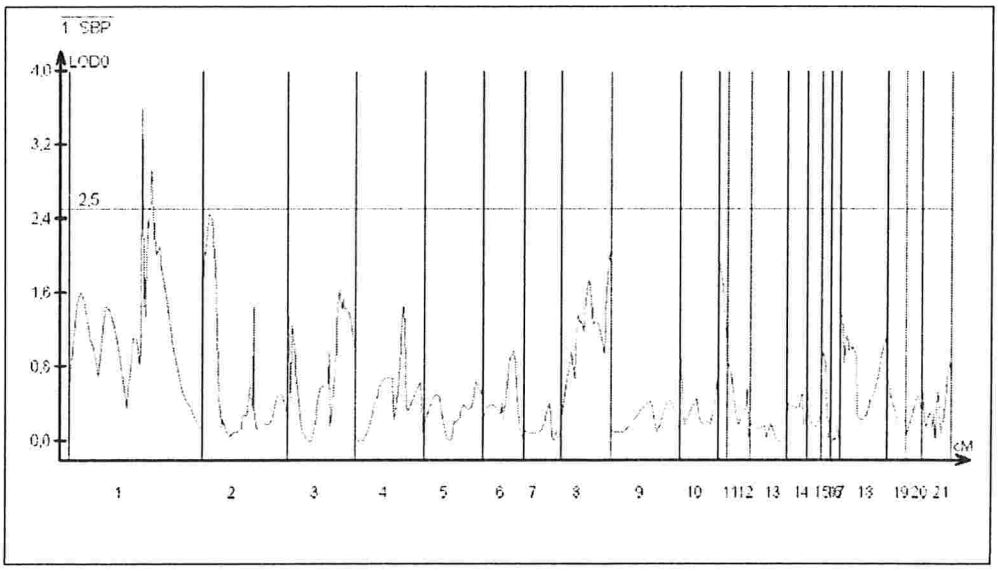


Figura 5.5: Gráfico de perfis da estatística *Lod score* para o mapeamento da pressão sanguínea sistólica antes da exposição ao sal (SBP) com os ratos F2 do InCor por meio da análise marginal do modelo com pleiotropia.

De acordo com a Figura 5.5, temos que há um sinal de presença de QTL para o traço SBP na região do cromossomo 1.

Quanto ao gráfico da Figura 5.6, temos que há sinais de presença de QTL para

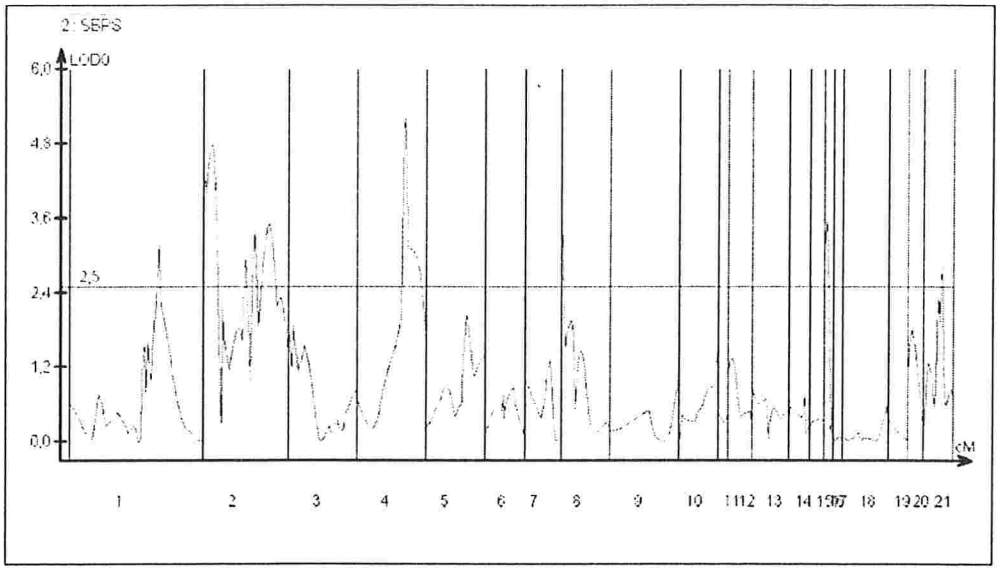


Figura 5.6: Gráfico de perfis da estatística *Lod score* para o mapeamento da pressão sanguínea sistólica depois da exposição ao sal (SBPS) com os ratos F2 do InCor por meio da análise marginal do modelo com pleiotropia.

o traço SBPS nas regiões dos cromossomos 1, 2, 4, 8 e 16, já identificados, além do 21.

Note que a análise marginal do modelo multivariado não é idêntica à análise univariada. Isto porque nas análises univariadas testamos para um dos traços os efeitos aditivo e de dominância do fenótipo SBPS na ausência de qualquer informação sobre o fenótipo SBP, enquanto que, na análise marginal do modelo multivariado, testamos os efeitos aditivo e de dominância no modelo que, apesar de considerar nulos os efeitos do outro fenótipo, inclui a informação da covariância entre os traços.

A Figura 5.7 apresenta os perfis da estatística *Lod score* da análise conjunta do modelo com pleiotropia juntamente com os resultados das análises marginais já apresentadas anteriormente para efeito de comparabilidade. Note que $(LR_{12} \geq máximo[LR_1, LR_2])$, como pode ser demonstrado analiticamente. Além disso, vale ressaltar que os sinais encontrados nos cromossomos 2, 4, 8 e 16 para nossos dados têm obtido uma maior concordância na literatura, isto é, outros estudos também têm identificado estes mesmos sinais. Contudo, há questionamentos sobre a validação do sinal no cromossomo 1, pois tal resultado não tem sido encontrado em outros

estudos. Quanto ao cromossomo 21, são necessárias maiores investigações a respeito do sinal identificado, pois trata-se do cromossomo sexual.

De acordo com a mesma Figura 5.7, podemos perceber que a curva apresentada pela análise conjunta apresenta uma tendência mais parecida com a mostrada pela análise marginal do SBPS do que com a mostrada pela análise marginal do SBP. Deste modo, pode-se dizer que o fenótipo SBP deve ser menos influenciado por genes do que o SBPS ou a identificação de genes que controlam o SBP exige modelos mais precisos. Possivelmente, a herdabilidade do SBPS é maior que a do SBP.

Ainda, é possível detectar um sinal de presença de QTL em uma região do cromossomo 5 que não havia sido identificado até o momento por meio das demais análises. Deste modo, apresentaremos na Tabela 5.3 a estatística razão de verossimilhanças e as estimativas dos efeitos aditivo e de dominância do suposto QTL presente na região do cromossomo 5 sob o modelo multivariado.

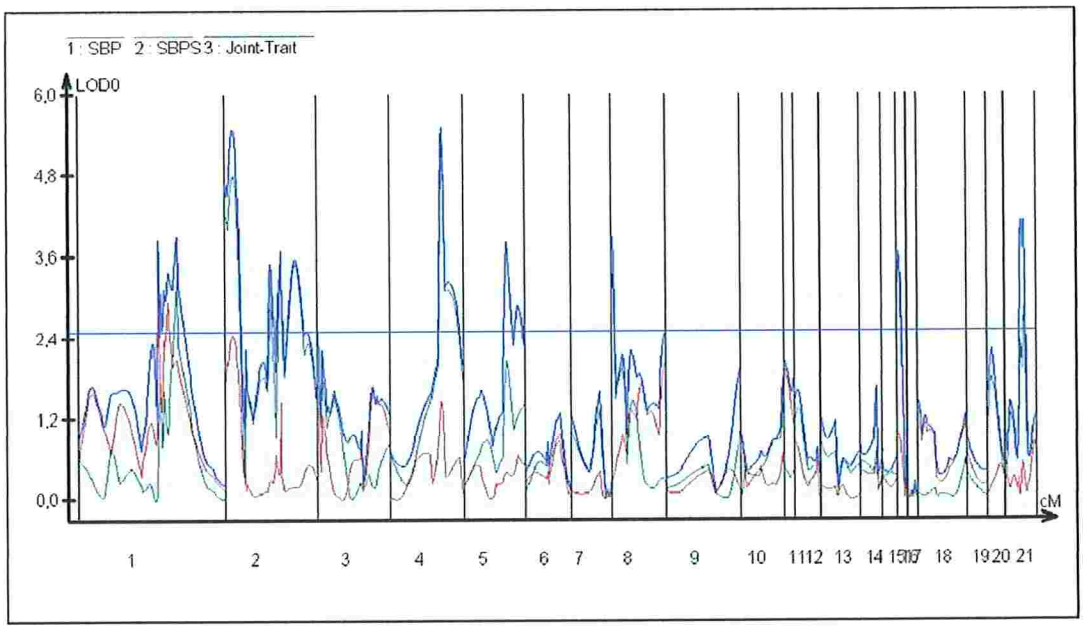


Figura 5.7: Gráfico de perfis da estatística *Lod score* para o mapeamento da SPB e da SBPS marginal e conjuntamente com os dados de ratos F2 do InCor por meio do modelo com pleiotropia.

De acordo com o que foi discutido na seção 4.3, temos que a situação apresentada

Fenótipo	Lod($H_3 : H_0$)	Efeito aditivo	Efeito de dominância
SBP	0,4069	0,0040140	3,7544535
SBPS	2,1107	3,6036658	-4,4117278
Análise conjunta	3,9124		

Tabela 5.3: Estatística razão de verossimilhanças e as estimativas dos efeitos aditivo e de dominância do suposto QTL presente na região do cromossomo 5 sob o modelo multivariado.

pelo suposto QTL encontrado na região do cromossomo 5 é a mais favorável para o ajuste do modelo multivariado. Veja que os dados expostos na Tabela 5.3 indicam que o suposto QTL, na situação basal aumenta (apesar de não ser significativo) em aproximadamente 3,8 unidades a pressão sanguínea sistólica devido ao efeito de dominância deste loco genético, enquanto que, depois da exposição ao sal, o suposto gene atua na diminuição da pressão sistólica em aproximadamente 4,4 unidades, isto é, seu efeito de dominância é -4,4. Deste modo, o critério $(\rho d^*_{SBP} d^*_{SBPS}) < 0$ é encontrado neste caso.

Note que o QTL identificado no cromossomo 5 tem efeito protetor à exposição ao sal. Podemos argumentar que os perfis individuais apresentados na figura 5.4 com uma queda drástica do valor da pressão podem ser pontos de influência neste resultado. Estudos mais aprofundados na direção de análise diagnóstico em modelos de mapeamento genético são necessários.

Capítulo 6

Considerações Finais

É possível ressaltar vários pontos apresentados neste trabalho que merecem ser discutidos e estudados com maior aprofundamento visando uma maior precisão na busca pela identificação de genes controladores de traços quantitativos.

Um dos problemas no mapeamento genético que foi levantado ao longo deste trabalho está relacionado ao nível de significância para uma família de comparações. Sabe-se que múltiplos testes simultâneos implicam na necessidade de correção do nível de significância global. A correção mais indicada ainda é polêmica dentre as diversas abordagens existentes na literatura e no caso específico de mapeamento genético.

Outro ponto está relacionado à densidade de pontos nos mapas de marcadores. Nosso estudo baseou-se em um mapa com 182 marcadores moleculares distribuídos nos 21 cromossomos do rato e, é importante salientar que atualmente um mapa com este número de marcadores é considerado pouco denso senão pouco informativo. Os estudos mais recentes publicados baseiam-se em mapas de marcadores que possuem cerca de 2.000 marcadores moleculares visando um maior poder para os testes realizados.

Vimos também que o Mapeamento Intervalar Composto propõe a inclusão de covariáveis genéticas no modelo com o intuito de ganhar precisão em relação ao Mapeamento Intervalar simples. Tal ganho é esperado baseado na diminuição do resíduo e na hipótese de isolamento no efeito estimado do QTL principal. Entretanto, o fato de desconhecermos o padrão de dependência entre o suposto QTL e as supostas

covariáveis faz com que exista a possibilidade deste subconjunto de locos acabar competindo com o efeito principal do QTL, o que não é desejado. Tal problema é verdadeiro também para a inclusão de quaisquer covariáveis, mesmo as de natureza ambiental; não genéticas, como por exemplo: sexo ou peso.

No que se refere à epistasia, podemos comentar que acredita-se que modelos de regulação celular sejam muito complexos e devem envolver alta ordem de interações genéticas e ambientais [Cordell, 2002], [Asíns, 2002]. Sendo assim, os modelos aditivos não devem ser os mais adequados para seu estudo. A literatura é polêmica neste assunto e, na tentativa de descobrir se a interação que não foi detectada em um determinado modelo não existe mesmo ou trata-se de uma limitação estatística inerente ao modelo proposto, surgem alternativas de estudo como os delineamentos experimentais com perturbações multifatoriais [Jansen, 2003]. Estes delineamentos originam linhagens de animais congênicos cujo genoma recebe fragmentos de diferentes linhagens de animais e deste modo hipóteses complexas de interação entre genes podem ser testadas. Suponha que a conclusão de nosso estudo baseado no modelo seja de que existem 5 locos com efeitos aditivos que controlam a hipertensão, e não existe efeito de interação entre eles. Em laboratório, para verificar se a hipótese de inexistência de epistasia entre os locos é verdadeira, criam-se vários animais com a constituição dos 5 locos diversificada de forma que se possa verificar nestes animais congênicos se a conclusão obtida por meio do modelo estatístico procede, de acordo com a manifestação do fenótipo (que poderá ser quantificada empiricamente).

No caso da pleiotropia, os modelos que a envolvem são extremamente úteis porque não há dúvidas de que o sistema biológico é complexo e interconectado [Mangin *et al.*, 1998]. Desta forma, modelos que envolvem vários fenótipos, como por exemplo os modelos longitudinais, são ferramentas valiosas a serem exploradas no mapeamento de genes.

Finalmente, vale comentar que este trabalho abordou os modelos de regressão no mapeamento genético considerando que os efeitos do QTL com base nos marcadores moleculares são variáveis de efeitos fixos (como se estas fossem conhecidas, observáveis). Possivelmente modelos apropriados a esta situação sejam os modelos com erros nas variáveis. A literatura na área de mapeamento genético tem explorado pouco alternativas desta natureza.

Apêndice A

Banco de Dados do Delineamento F2 (InCor)

A.1 Banco de dados dos Marcadores Moleculares e Variáveis Genotípicas

Para o cromossomo 1, denominado Chr01:

	Marcador 1	Marcador 2	Marcador 3	$\cdots$	Marcador 15
Distância (cM)	3	26	35	$\cdots$	152
animal 1	1	1	-1	$\cdots$	2
animal 2	-1	-1	-1	$\cdots$	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
animal 221	-1	-1	1	$\cdots$	-1

Tabela A.1: Localização dos Marcadores Moleculares e dados genotípicos dos ratos para o cromossomo 1.

A codificação utilizada para os genótipos dos Marcadores Moleculares da Tabela A.1está descrita na Tabela A.2.

Genótipo do Marcador	Código
AA	2
Aa	1
aa	0
A-	12
a-	10
–	-1

Tabela A.2: Códigos utilizados no Banco de Dados de Genótipos de Marcadores.

Para cada um dos 21 cromossomos temos uma tabela semelhante à apresentada, sendo que cada cromossomo tem uma quantidade diferente de marcadores moleculares. Veja Tabela A.3.

Cromossomo	No. de Marcadores
1	15
2	23
3	18
4	14
5	12
6	7
7	8
8	10
9	5
10	8
11	5
12	8
13	6
14	7
15	2
16	5
17	4
18	9
19	2
20	3
21	11

Tabela A.3: Número de Marcadores Moleculares em cada um dos 21 cromossomos.

A.2 Banco de dados das Variáveis Fenotípicas

	Traço 1	Traço 2	Traço 3	$\cdots$	Traço 23
animal 1	122	119	67	$\cdots$	missing
animal 2	126	119	90	$\cdots$	missing
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
animal 221	124	107	48	$\cdots$	179

Tabela A.4: Dados fenotípicos para os 221 ratos.

Referências Bibliográficas

[Amabis & Martho, 1994] Amabis, J. M., & Martho, G. R. 1994. *Biologia das populações.* São Paulo: Ed. Moderna.

[Asíns, 2002] Asíns, M. J. 2002. Preview and future of quantitative trait locus analysis in plant breeding. *Plant breeding*, **121**, 281–291.

[Balding *et al.*, 2003] Balding, D.J., Bishop, M., & Cannings, C. 2003. *Handbook of statistical genetics.* England: John Wiley & Sons Ltd.

[Benjamini & Hocheberg, 1995] Benjamini, Y., & Hocheberg, Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistics society*, **57**, 289–300.

[Churchill & Doerge, 1994] Churchill, G. A., & Doerge, R. W. 1994. Empirical threshold values for quantitative trait mapping. *Genetics*, **138**, 967–971.

[Cordell, 2002] Cordell, H. J. 2002. Epistasis: what it means, what it doesn´t mean, and statistical methods to detect it in humans. *Human molecular genetics*, **11**, 2463–2468.

[Draper & Smith, 1981] Draper, N. R., & Smith, H. 1981. *Applied regression analysis.* New York: John Wiley Sons.

[Fagundes & Taha, 2004] Fagundes, D. J., & Taha, M. O. 2004. Modelo animal de doença: critérios de escolha e espécies de animais de uso corrente. *Acta cirúrgica brasileira*, **19**, 59–65.

[Falconer, 1964] Falconer, D. S. 1964. *Introduction to quantitative genetics.* London: Oliver and Boyd Ltd.

[Farah, 1997] Farah, S. B. 1997. *Dna segredos & mistérios.* São Paulo: Sarvier Editora de Livros Médicos Ltda.

[Hair *et al.*, 1998] Hair, F. H., Anderson, R. E., Tatham, R. L., & Black, W. C. 1998. *Análise multivariada de dados.* São Paulo: Bookman.

[Haley & Knott, 1992] Haley, C. S., & Knott, S. A. 1992. A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity*, **69**, 315–324.

[Jansen, 1992] Jansen, R. C. 1992. A general mixture model for mapping quantitative trait loci by using molecular markers. *Theoretical and applied genetics*, **85**, 252–260.

[Jansen, 2003] Jansen, R.C. 2003. Studying complex biological systems using multifactorial perturbation. *Nature reviews genetics*, **4**, 145–151.

[Jiang & Zeng, 1995] Jiang, C., & Zeng, Z. B. 1995. Multiple trait analysis of genetic mapping for quantitative trait loci. *Genetics*, **140**, 1111–1127.

[Kao & Zeng, 2002] Kao, C. H., & Zeng, B. Z. 2002. Modeling epistasis of quantitative trait loci using cockerham's model. *Genetics*, **160**, 1243–1261.

[Kao *et al.*, 1999] Kao, C. H., Zeng, Z. B., & Teasdale, R. D. 1999. Multiple interval mapping for quantitative trait loci. *Genetics*, **152**, 1203–1216.

[Kempthorne, 1957] Kempthorne, O. 1957. *An introduction to genetic statistics.* New York: John Wiley & Sons, Inc.

[Knott & Haley, 2000] Knott, S. A., & Haley, C. S. 2000. Multitrait least squares for quantitative trait loci detection. *Genetics*, **156**, 899–911.

[Lander & Botstein, 1989] Lander, E. S., & Botstein, D. 1989. Mapping mendelian factors underlying quantitative traits using rflp linkage maps. *Genetics*, **121**, 185–199.

[Liu, 1998] Liu, B. H. 1998. *Statistical genomics.* Florida: CRC Press LLC.

[Lynch & Walsh, 1998] Lynch, M., & Walsh, B. 1998. *Genetics and analysis of quantitative traits.* Sunderland: Sinauer Associates.

[Mangin *et al.*, 1998] Mangin, B., Touquet, P., & Grimsley, N. 1998. Pleiotropic qtl analysis. *Biometrics*, **54**, 88–99.

[Meng & Rubin, 1993] Meng, X. L., & Rubin, D. B. 1993. Maximum likelihood estimation via the ecm algorithm: a general framework. *Biometrika*, **80**, 267–268.

[Neter *et al.*, 1996] Neter, J., H., Kutner M., Nachtsheim, C. J., & Wasserman, W. 1996. *Applied linear statistical models.* Chicago: McGraw-Hill.

[Ott, 1991] Ott, J. 1991. *Analysis of human genetics linkage.* Baltimore: The Johns Hopkins University Press.

[Pereira & Krieger, 2001] Pereira, A. C., & Krieger, J. E. 2001. Biologia e genética molecular aplicadas ao diagnóstico e tratamento da hipertensão novos paradigmas, antigos problemas. *Rev bras hipertens*, **8**, 105–113.

[Reiner *et al.*, 2003] Reiner, A., Yekutieli, D., & Benjamini, Y. 2003. Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics*, **19**, 368–375.

[Schork *et al.*, 1995] Schork, N. J., Krieger, J. E., Trolliet, M. R., Franchini, K. G., George, K., Krieger, E. M., Lander, E. S., Dzau, V.J., & Jacob, H.J. 1995. A biometrical genome search in rats reveals the multigenic basis of blood pressure variation. *Genome research*, **5**, 164–172.

[Sen & Churchill, 2001] Sen, S., & Churchill, G. A. 2001. A statistical framework for quantitative trait mapping. *Genetics*, **159**, 371–387.

[Sham, 1998] Sham, P. 1998. *Statistics in human genetics.* London: Arnold Applications of Statistics.

[Silva & Coelho, 2005] Silva, H. D., & Coelho, A. S. G. 2005. Métodos biométricos aplicados a análise de qtl's. *Anais do seagro.*

[Silva Jr & Sasson, 1990] Silva Jr, C., & Sasson, S. 1990. *Biologia 3.* São Paulo: Atual Editora.

[Zeng, 1993] Zeng, Z. B. 1993. Theoretical basis os separation of multiple linked gene effects on mapping quantitative trai loci. *Proc. natl. acad. sci. usa*, **90**, 10972–10976.

[Zeng, 1994] Zeng, Z. B. 1994. Precision mapping of quantitative trait loci. *Genetics*, **136**, 1457–1468.

[Zeng *et al.*, 1999] Zeng, Z. B., Kao, C. H., & Basten, C. J. 1999. Estimating the genetic architecture of quantitative traits. *Genetical research*, **74**, 279–289.