Análise de associação aplicada ao mapeamento genético de doenças

Maria Jacqueline Batista

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

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

Durante a elaboração deste trabalho a autora recebeu apoio financeiro da FAPESP
Processo N° 04/04322-3

– São Paulo, 2006 –

Análise de associação aplicada ao mapeamento genético de doenças

Este exemplar corresponde à redação final da dissertação devidamente corrigida e defendida por Maria Jacqueline Batista e aprovada pela comissão julgadora.

São Paulo, 03 de março de 2006.

Comissão Julgadora:

- Titulares:
  - Profa. Júlia Maria Pavan Soler (Orientadora) - IME/USP
  - Prof. Dalton Francisco de Andrade - UFSC
  - Profa. Cibele Queiroz da Silva - UnB
- Suplentes:
  - Prof. Carlos Alberto de Bragança Pereira - IME/USP
  - Profa. Hildete Prisco Pinheiro - IMECC/UNICAMP

"O coração do homem planeja seu caminho, mas o Senhor lhe dirige os passos".

Provérbios 16,9.

A Deus,
meus pais, meu irmão,
minha avozinha querida
e ao Juvêncio.

Agradecimentos

Agradeço:

A Deus, pela saúde e oportunidade.

Ao meus pais: Socorro e Otacílio, que com muita dedicação me passaram amor e ensinamentos, dentre eles é que "estudar é sempre importante", ao meu irmão: Júnior, que sempre me apoia.

À minha avozinha: Ana, pelo carinho, por tanta saudade que sentimos neste período em que pela primeira vez ficamos tão distantes.

Ao meu noivo: Juvêncio (meu nego), pelo amor, companhia, compreensão, paciência nas inúmeras "crises", pela ajuda no desenvolvimento deste trabalho e principalmente pelas alegrias nos momentos em que eu precisava de mais apoio.

À minha sogra Gracilene, que sempre acompanha e torce muito pelas conquistas dos nossos sonhos.

À minhas primas Lindenira e Lourdinha pelo apoio quando vim para São Paulo.

À minha orientadora e amiga, professora Júlia Maria Pavan Soler, pela excelente recepção quando cheguei em São Paulo, pelo apoio e incentivo incondicional no decorrer deste trabalho, tanto no âmbito acadêmico quanto no pessoal.

Ao Laboratório de Cardiologia e Genética Molecular do InCor/USP, na pessoa do Dr. Alexandre Pereira, por todas as "discussões" sobre o trabalho e pela autorização na utilização dos dados reais, o qual agradeço também ao Dr. Eduardo Krieger.

À FAPESP (Processo N° 04/04322-3) pelo apoio financeiro no desenvolvimento deste trabalho.

Ao prof. João Maurício (UFC), pelo apoio, ensinamentos, por ser além de excelente professor, um grande amigo, que sempre estar disposto a ajudar, obrigada por tudo.

Ao professor João Welliandre (UFC), meu orientador de iniciação científica, uma das pessoas que mais me incentivou a continuar no meio acadêmico, obrigada.

Ao professor e amigo Dalton de Andrade (UFSC), que me acompanha desde a graduação, em alguns momentos do meu mestrado e com muita alegria e competência participou da minha banca, o qual contribuiu bastante com suas sugestões, agradeço por sempre ser solicito.

À professora Cibele Queiroz (UnB), pelas sugestões dadas ao presente trabalho, os quais foram muito importantes para complementá-lo.

Ao professor e amigo Julio Singer (USP), pelas palavras de incentivo e apoio.

Aos professores da graduação (UFC) em especial as professoras Ana Maria e Rosa Salani. Agradeço também aos professores Júlio Barros, Robson Medeiros, Alan Pereira, José Eduardo, Vicente de Paulo, Sílvia Freitas, José Lassance, Bitú Feitosa, Ronald Targino, Manoel Câmpelo e André Jalles.

Aos funcionários do DEMA (UFC), Luiza, Magerí, Mariluse e Margarida. Aos meus amigos da graduação, Chrystina, Cyntia, Velma, Erivan, Waleska, Wládia, Fábio, Saulo, Cleudimar, Marcos Aurélio e todos da turma 1999.2.

Aos professores que tive contato no mestrado (USP) em especial à professora Elisete Aubin e ao professor Marcos Magalhães, que confiaram em mim no primeiro semestre e por isto sou muito grata por este crédito. A professora Mônica Sandoval pela oportunidade de participar como monitora na disciplina MAE 5755. Também agradeço aos professores Wagner Borges, Gilberto Alvarenga, Carlos Bragança, Sílvia Nagib, Antônio Carlos Pedroso, Clélia Toloi e Elisabeti Kira.

Aos meus amigos: Michelli pelas palavras de amizade e todo carinho com que se preocupa comigo, juntamente com o seu marido Horácio; a Edijane, pela amizade e por ter me ajudado em muitos momentos, desde à situações referentes a este trabalho, bem como problemas "extras"; ao Marcelo & Lane, Patrícia & Raydonal, Iracema, Kelly, Elisa, Diana & Gustavo, Perseverando, Elier, Diana (Colômbia), Fred, Tatiane & Alessandro, Delli, Suely, Gilberto & Regele, Paulo Tadeu, Marcelo (Presidente Prudente), Luz Mery, Núbia, Adrilayne, Regina Albanese, Lourdes e Betsabé.

Aos meus amigos cearenses e cruspianos: Juvêncio, Caio pela solicitude, Márcio, Rafael, Alexandre, Patrícia e Jonny. À Coseas pela concessão da minha moradia no CRUSP, em especial minha assistente social Luiza.

Resumo

O mapeamento genético e a genética funcional de doenças são de grande importância na pesquisa médica e genômica. Para estas finalidades o estudo de associação entre fatores de risco genéticos e doença tem ganhado destaque na literatura. Neste trabalho disserta-se sobre a análise de associação aplicada ao mapeamento genético de doenças, caracterizando diferentes possibilidades de planejamentos experimentais e de utilização de modelos estatísticos de análise de dados. As formalizações estatísticas, como o tipo de delineamento experimental, a inclusão ou não de dados familiares, bem como a escolha do método estatístico de análise, que são decisivos na avaliação do poder dos testes obtidos e na sua aplicabilidade ao mapeamento genético, também são discutidas. Além disso, considera-se a análise de associação por meio de modelos de regressão logística em que, as análises de dados genéticos são abordadas via dados no nível genotípico e cromossômico. Finalmente, os conceitos supracitados são aplicados a conjuntos de dados reais, fornecidos pelo Laboratório de Cardiologia e Genética Molecular do InCor/USP, com o objetivo de ilustrar o problema teórico tratado e motivar a aplicação das metodologias estatísticas envolvidas.

Abstract

The genetic mapping and functional genetics have great importance in the genomics research. In order to conduct these researches the study of the association between genetic risk factors and disease has been becoming an important role in the literature. In this work we consider the association analyses applied to the genetic diseases mapping, charactering different possibilities of experimental designs and the use of statistical models to analyze data sets. The statistical concepts, as the kind of experimental design, the inclusion of familiar records or not, as well as the choice of the statistical analyze method, which are very important to the evaluation of the power of the tests obtained and to their applicability in the genetic mapping, are also discussed. Furthermore, we consider the association analysis at person level and chromosome data set. Finally, the latter concepts are applied to a real data set, provided by the Molecular Genetic and Cardiology Laboratory of InCor/USP, in order to illustrate the theoretical problem treated in this work and to motive the use of the involved statistical methodologies.

Índice

Lista de Tabelas

Lista de Figuras

Capítulo 1

Introdução

Estudos Epidemiológicos envolvem tipicamente a coleta, análise e interpretação de dados associados à saúde da população (Newman, 2001). A realização destes estudos, na maioria das vezes, tem o interesse em verificar a existência de associação entre fatores de risco e uma certa doença, por exemplo, o consumo de sal e hipertensão arterial, o tabagismo e câncer de pulmão. Para tais análises, se faz uso intenso de métodos estatísticos específicos de acordo com os objetivos a serem explorados em cada pesquisa. A literatura científica tem abordado amplamente tópicos nesta área.

Neste contexto a Epidemiologia Genética ganhou destaque em conseqüência do acelerado avanço do conhecimento em Genômica que, por sua vez, faz parte dos avanços da Biotecnologia. Deste modo, os fatores de risco para doenças passaram a ser estudados no nível molecular mais do que ambiental. Como exemplo, pode-se citar o gene BRCA1 identificado como fator de risco para o câncer de mama. Neste cenário, atualmente o grande desafio da Epidemiologia Genética é identificar todos os fatores genéticos que participam de alguma forma na etiologia e regulação de doenças comuns como obesidade, depressão, hipertensão, diabetes, entre outras.

A Epidemiologia Genética tem contribuído para estudos sobre a localização de genes associados a fatores que causam doenças, ou seja, estudos cujo objetivo é investigar a existência de associação entre um fator de risco genético e a doença. Para tais análises são tomadas informações de vários locos genéticos[1] passíveis de genotipagem, isto é, passíveis de se coletar informações moleculares, e testa-se a existência de associação entre loco genético e a doença sob estudo. Se uma medida de associação significativa é encontrada,

[1] Loco Genético: seguimento (localização) no cromossomo.

acredita-se que o próprio loco genético é o gene regulador da doença (gene funcional) e, neste caso, diz-se que a doença está mapeada geneticamente. Por outro lado, pode acontecer que o efeito do loco genético seja simplesmente um efeito "aparente", refletido pelo loco devido à sua proximidade (ligação) com o gene regulador. Neste último caso, a análise de associação também é útil para o mapeamento genético da doença, dado que obteve-se uma região genética candidata a conter o gene regulador. Porém, um problema possível com estas análises de associação é a ocorrência de conclusões falso-positivas em que, para certas situações de populações estruturadas (por exemplo, estratificação por etnias) pode-se ter locos associados com a doença, os quais não são genes reguladores e nem estão próximos a estes. Diz-se nestes casos que a associação é devida ao efeito de confundimento gerado por heterogeneidade populacional não controlada.

Deste modo, em Epidemiologia Genética a definição de associação se estende a várias situações. A dependência das ocorrências alélicas[2] de dois locos genéticos, por exemplo, é definida como associação alélica ou desequilíbrio de ligação. Sham (1998) explica que os estudos entre alelos pertencentes a locos genéticos diferentes, visam detectar um possível desequilíbrio (ou associação probabilística) na distribuição das frequências alélicas dos locos envolvidos. Tem crescido significativamente o interesse na literatura por estudos deste tipo, desde que estes se apresentam como uma ferramenta útil para o mapeamento genético de várias doenças e para a análise genética funcional (Jorde, 1995; Devlin et al. 1996; Collins & Morton, 1998; Lazzeroni, 1998; Kerstann et al. 2004).

Por outro lado, quando se estuda localização de genes no cromossomo, tem-se um outro conceito de associação, introduzido no contexto de análise de ligação (do inglês, *linkage*), que também tem sido extensivamente utilizado na identificação de regiões no genoma envolvidas na etiologia e regulação de certas doenças. Nestes casos, tem-se interesse em realizar inferências sobre a probabilidade de ocorrer recombinação (troca de material genético) entre dois locos genéticos, ou seja, estimar e testar hipóteses sobre um parâmetro denominado fração de recombinação entre, por exemplo, um gene e um certo marcador molecular (Terwilliger & Ott, 1994). A análise de ligação está fundamentada na seguinte

[2] Alelos: formas alternativas de um gene em um dado loco.

Lei da Genética Clássica: locos genéticos próximos tendem a se segregarem íntegros de geração a geração, sem a ocorrência de recombinações entre os alelos que os compõem. Esta lei tem sido explorada no mapeamento genético que avalia o sinal de ligação (associação) entre um gene envolvido na regulação de uma doença e locos de marcadores moleculares de localização e genotipagem conhecidas. Ligação entre locos é a dependência na forma de segregação (na origem) dos alelos nos locos, sendo relevante para o estudo de ligação se ter informação familiar, isto é, dados de indivíduos que compartilham algum grau de parentesco entre si, por exemplo, pais, filhos e netos.

Os conceitos de desequilíbrio de ligação (associação probabilística entre alelos) e ligação (associação na forma de segregação dos alelos) em Genética são implicitamente usados nos estudos de associação entre loco genético ("gene") e doença. Para esta última proposta existem diferentes alternativas de coleta de dados, como os delineamentos observacionais (transversal, prospectivo e retrospectivo) que podem incorporar ou não a informação genética de indivíduos e seus parentes.

Contudo, situações como migração entre populações com diferentes probabilidades alélicas em locos de interesse, efeito de fundadores, mutação, seleção, estratificação e miscigenação podem conduzir a associação entre locos genéticos não necessariamente próximos e nem mesmo relacionados funcionalmente. Um questionamento básico que surge é se dois locos ligados ("próximos") estão, ou não, em desequilíbrio de ligação ("em associação probabilística"). Observa-se que estudos com a população brasileira merecem cuidados já que ela é caracterizada por grande diversidade étnica e intensa miscigenação, o que pode ser verificado, por exemplo, nos censos demográficos realizados pelo Instituto Brasileiro de Geografia e Estatística (IBGE, 2002).

Deste modo, uma das preocupações nos estudos de associação é com a obtenção dos dados amostrais, pois a partir de como estes são obtidos pode existir um efeito de confundimento conduzindo a falsas conclusões resultantes de testes de hipóteses estatísticas de interesse. Uma maneira de minimizar isto é utilizar delineamentos experimentais apropriados à natureza do problema ou adotar metodologias robustas de análise de dados. Por exemplo, considerando delineamentos *caso-controle*, o TDT (do inglês, *Transmission*

Disequilibrium Test), é uma alternativa de teste de associação robusto para situações de populações miscigenadas. Este teste foi originalmente proposto por Spielman et al. (1993), o qual é baseado na análise de dados de trios (pai, mãe e filho afetado). No contexto de averiguar associação entre locos de marcadores moleculares e genes que influenciam a suscetibilidade de doenças, o TDT é uma alternativa de análise, uma vez que é não afetado por estratificação ou evidências falso-positivas. Na prática, este teste tem sido mais utilizado para doenças, em geral, avaliadas de forma dicotômica, contudo esta metodologia tem sido generalizada para mapeamento genético de doenças complexas avaliadas de forma quantitativa (Rabinowitz, 1997). Este teste foi também estendido para situações de múltiplos alelos, por exemplo, em Sham & Curtis (1995). Existem outros delineamentos que tendem a "proteger" as estatísticas de teste do efeito da estrutura populacional fazendo uso de dados familiares, como situações que comparam indivíduos *caso* e *controle* relacionados (parentes), isto é, que compartilham o mesmo "*background*" genético, por exemplo, pares de gêmeos sendo um afetado e outro não afetado. Porém, estes estudos levam mais tempo para serem conduzidos e são mais dispendiosos. Outra proposta é o uso de dados de pares de irmãos não necessariamente gêmeos, como exemplo, o método proposto por Spielman & Ewens (1998).

Os estudos caso-controle com indivíduos não relacionados têm suas vantagens em relação aos estudos com dados familiares, mas existe uma preocupação ao coletar os indivíduos relativamente ao "*background*" genético de cada grupo. *Caso* e *controle*, devem compartilhar alguma informação genética em comum, tornando os grupos comparáveis para os específicos fatores genéticos de interesse, o que minimiza o problema de falso-positivos ou de associações espúrias (Cardon & Palmer, 2003).

Como ilustração dos estudos caso-controle, considere o estudo descrito por Dunning et al. (1997) e discutido por Clayton (2003), com o objetivo de verificar associação de um polimorfismo "comum" no gene BRCA1 com o câncer de mama. Na Tabela 1.1 (a) tem-se a distribuição do genótipo do marcador molecular Pro871Leu em 800 amostras de *casos* e 572 amostras *controles*. Observa-se que a análise desta tabela procederá no nível genotípico (indivíduo), enquanto que para a Tabela 1.1 (b) tem-se uma análise no

nível cromossômico[3] em que os dados serão duplicados em relação ao nível genotípico. Para as amostras *casos* serão consideradas 1600 mulheres e para amostras *controles* 1144 mulheres. Nestes dois cenários de análise pode-se avaliar o risco de câncer de mama para mulheres que carregam os genótipos *LeuLeu*, *LeuPro* e *ProPro*, ou o risco de câncer de mama para mulheres que carregam os alelos *Leu* e *Pro*, ou seja, os riscos de câncer de mama podem ser definidos em função dos genótipos ou dos alelos. Este estudo será melhor explorado no Capítulo 4.

Tabela 1.1 *Distribuição do polimorfismo Pro871Leu no gene BRCA1 em casos e controles de câncer de mama.*

(a)

	Genótipo			
	LeuLeu	LeuPro	ProPro	Total
Caso	89	369	342	800
Controle	56	250	266	572

(b)

	Alelo		
	Leu	Pro	Total
Caso	547	1053	1600
Controle	362	782	1144

O objetivo deste trabalho é contribuir para a formalização da teoria envolvida nos estudos de associação genética, por meio de uma revisão da literatura e descrição de várias formulações destes estudos aplicados ao mapeamento genético de doenças. Dependendo do possível plano experimental adotado, o qual pode utilizar amostras com indivíduos indepedentes ou relacionados (parentes), são aplicados métodos estatísticos apropriados a cada tipo de estudo. Deste modo, um primeiro passo na análise de associação genética é conhecer a estrutura do delineamento amostral no qual os dados foram gerados.

Baseado no conteúdo supracitado, no Capítulo 2, são descritos alguns conceitos em

[3] As análises no nível genotípico e cromossômico serão detalhadas no Capítulo 2.

genética como fração de recombinação, equilíbrio de Hardy-Weinberg, análise de ligação e desequilíbrio de ligação. No Capítulo 3 serão descritos vários métodos estatísticos na análise de associação genética como, por exemplo, os estudos transversais e retrospectivos (caso-controle), incluindo o modelo de regressão logística via análise no nível genotípico e cromossômico e o TDT, que têm como principal propósito verificar associação entre um loco candidato e a ocorrência de uma doença. No Capítulo 4 são apresentadas aplicações desta metodologia, motivadas pela análise dos dados apresentados na Tabela 1.1 e de bancos de dados reais fornecidos pelo Laboratório de Cardiologia e Genética Molecular do InCor - USP. Estes são analisados para ilustrar os conceitos introduzidos neste trabalho e alguns planejamentos experimentais adotados para finalidade de mapeamento genético. No Capítulo 5 são apresentadas considerações finais.

Capítulo 2

Medidas de Associação Alélica

O interesse recente no entendimento do padrão de desequilíbrio de ligação entre locos genéticos na população humana é devido, em grande parte, ao crescimento dos estudos em Epidemiologia Genética de doenças complexas (Pritchard & Przeworski, 2001). Primeiramente, seguem alguns conceitos os quais são de interesse para o desenvolvimento deste trabalho.

O **cromossomo** é formado por uma molécula de DNA (*Ácido Desoxiribonucléico*) muito longa que se dispõe de forma "empacotada" dentro desta organela. **Marcador** molecular é uma seqüência de DNA, um loco genético, identificável no cromossomo, sendo possível genotipar indivíduos para tais locos, isto é classificá-los como AA, Aa ou aa, por exemplo. Os **genes** distribuem-se nos cromossomos de forma linear e correspondem a seqüências de DNA que codificam proteínas. Cada gene tem uma posição definida em um determinado cromossomo, a qual é chamada de **loco** (Farah, 1997). Gene é um termo geral que significa a entidade física transmitida de pai para filho durante o processo de reprodução que influência características hereditárias (Andrade & Pinheiro, 2002). Formas alternativas de um gene em um dado loco são chamados de **alelos**, em que estes presentes em um determinado loco constituem o **genótipo** (constituição genética do indivíduo), enquanto as características que se observam em um indivíduo, representam o **fenótipo**. Um loco é considerado **polimórfico** quando a freqüência do genótipo mais raro é de no mínimo 1% (Farah, 1997). Alelos (de diferentes genes) recebidos de um dos pais do indivíduo são chamados de **haplótipo** (Ott, 1991).

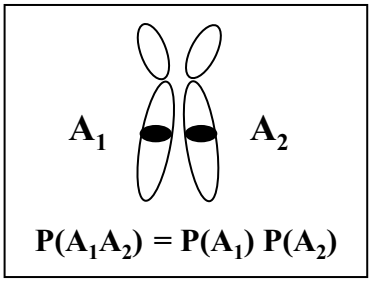
Neste capítulo será apresentado o coeficiente de desequilíbrio de ligação, um parâmetro importante no estudo de associação alélica. Serão também considerados outros conceitos

fundamentais para o desenvolvimento do presente trabalho, como equilíbrio de Hardy-Weinberg, fração de recombinação e análise de ligação.

2.1 Equilíbrio de Hardy-Weinberg

Para um único loco genético, um conceito importante em genética é o equilíbrio de Hardy-Weinberg (EHW), em que este descreve as probabilidades genotípicas em termos de probabilidades alélicas, supondo independência na combinação dos alelos paternos que definem o genótipo (vide, Figura 2.1). Sob condições de cruzamentos aleatórios e ausência de processos como migração, mutação e seleção, a população é dita estar em equilíbrio de Hardy-Weinberg.

Figura 2.1 *Ilustração do Equilíbrio de Hardy-Weinberg.*



Um sistema genético com r alelos em um loco ocorrendo com probabilidades p_i, i= 1, 2,...,r, diz-se em equilíbrio de Hardy-Weinberg se as $(r+1)r/2$ probabilidades genotípicas p_{ij}, $i \leq j$ são dadas por:

$$p_{ij} = p_i p_j [\mathbb{1}(i = j) + 2\mathbb{1}(i < j)], \tag{2.1}$$

onde $\mathbb{1}(C)$ representa a função indicadora do conjunto C.

Considerando locos dialélicos, segue na Tabela 2.1 a relação entre probabilidades dos alelos e genótipos sob EHW (Falconer & Mackay, 1996). De maneira geral, pode-se dizer

que o EHW é uma medida de associação entre alelos em um único loco. Existem possibilidades de desvios deste equilíbrio, uma delas é a estratificação da população, em que os cruzamentos são não aleatórios, ou seja, os cruzamentos entre indivíduos de estratos diferentes são menos prováveis de ocorrer do que cruzamentos entre indivíduos de mesmo estrato (Sham, 1998). Weir et al. (2004), fazem uma análise de um conjunto denso de SNPs[1] em que se observa desvios do EHW, possivelmente devido a erros de genotipagem.

Tabela 2.1 *Relação entre probabilidades dos alelos e genótipos.*

	Alelos		**Genótipos**		
	A_1	A_2	A_1A_1	A_1A_2	A_2A_2
Probabilidades	p	q	p^2	$2pq$	q^2

Na análise de associação genética de uma população descreve-se seus possíveis genótipos, e pode-se considerar duas abordagens de análise de dados: no nível de genótipo (indivíduo) e cromossômico (alélico). Quando se considera os dados de marcadores na forma de genótipo, (A_1A_1, A_1A_2, A_2A_2, por exemplo) não se assume o EHW. Tem-se uma amostra de tamanho n indivíduos (n genótipos) e, assim o risco da doença será definido em função do genótipo. Na análise no nível cromossômico, o tamanho da amostra será dobrado ($2n$ cromossomos, que são o número de alelos) e, na construção desta amostra aumentada assume-se o EHW, isto é, os alelos são considerados independentes. A partir dos dados genotípicos, por exemplo, em um estudo caso-controle para uma tabela 2×3, pode-se obter os dados cromossômicos que serão dispostos em uma tabela 2×2, como visto anteriormente, na ilustração do Capítulo 1, Tabela 1.1. Estas duas análises podem também ser estendidas na construção do TDT.

[1] SNPs: do inglês "*Single Nucleotide Polymorphisms*", que são polimorfismos de um único nucleotídeo que ocorrem na população e são utilizados como marcadores polimórficos.

2.2 Análise de Ligação

Considerando pares de locos genéticos os conceitos de ligação e desequilíbrio de ligação são bastante importantes. Na análise de ligação estuda-se os eventos de recombinação entre dois locos, sejam eles genes, marcadores moleculares, aberrações cromossômicas, etc, em que utiliza-se a fração de recombinação (denotada por θ) definida como a probabilidade de que ocorra um evento de recombinação (*crossing-over*)[2] entre dois locos quaisquer. O parâmetro θ está diretamente relacionado à distância citogenética entre locos num mesmo cromossomo (Ott, 1991), e seu valor varia no intervalo [0, 0,5]. Verifica-se, que quando θ se aproxima de **0,5**, não existe ligação genética entre os locos, os quais são considerados como geneticamente distantes no cromossomo; se θ está próximo de **0**, isto indica que há ligação genética entre dois locos e, portanto, os locos são considerados como geneticamente próximos no cromossomo.

Alguns métodos para análise de ligação envolvendo dois locos foram desenvolvidos (Mather, 1951) utilizando as probabilidades genéticas de duplos heterozigotos. Um dos métodos mais conhecidos foi desenvolvido por Morton (1955), sendo baseado em dados familiares e em uma estatística de teste de ligação, a estatística Lod Score, definida como o logaritmo na base 10 da razão de verossimilhanças:

$$\text{LOD} = Z(\widehat{\theta}) = \log_{10} \frac{L(\widehat{\theta})}{L(0,5)}, \tag{2.2}$$

onde $L(\widehat{\theta})$ é a função de verossimilhança maximizada sob todo o espaço paramétrico e $L(0,5)$ é a função de verossimilhança calculada sob a hipótese nula, H_0: θ=0,5. Para a maioria das aplicações L é definida em função do modelo de probabilidades Binomial onde, por exemplo, em famílias de 5 filhos observa-se o número daqueles que são recombinantes para dois locos em estudo. De maneira geral, quando a estatística Lod Score excede um valor crítico, pode-se dizer que há evidência a favor da ligação entre locos, ou seja, em situações de mapeamento genético conclui-se que o loco da doença se encontra

[2] *Crossing-over*: troca recíproca entre segmentos correspondentes de cromossomos homólogos, a qual ocorre na primeira divisão da meiose.

nas proximidades (ligado) do loco marcador. Baseado na teoria de testes da razão de verossimilhanças para grandes amostras tem-se que, sob a hipótese nula, 4,6 $\times$ $Z(\widehat{\theta})$ segue aproximadamente uma distribuição Qui-Quadrado com 1 grau de liberdade. O valor crítico LOD=3, o qual tem sido adotado como um critério de decisão, está associado a um nível descritivo[3] menor que 10^{-4}.

Pode ser estabelecida uma relação entre a fração de recombinação (θ) e distância citogenética ($\mathbf{d}$). Esta relação funcional nem sempre é clara e direta, sendo necessário assumir pressupostos referentes à distribuição de probabilidades para o número de *crossovers* em intervalos ao longo do genoma. Existem várias funções de distância citogenética propostas na literatura, por exemplo, em Lange (1997), Ott (1991) e Schuster & Cruz (2004).

A função de distância citogenética mais simples (função identidade $\theta = \mathbf{d}$) foi proposta por Morgan (1928), apropriada para intervalos curtos. Haldane (1919), expressou a função de distância quando assume-se que os *crossovers* se distribuem no cromossomo segundo uma distribuição Poisson. Outra função de distância citogenética é vista em Kosambi (1944), em que supõe-se a restrição $\theta < 0{,}5$, ou seja, $\theta = 0{,}5$ não pode ser considerado. Quando assume-se N *crossovers* finitos ocorrendo no genoma de acordo com a distribuição binomial tem-se a função desenvolvida por Karlin (1984), (ver Tabela 2.2). Como ilustração, para um θ= 0,20 tem-se $\mathbf{d}$= 0,20M= 20cM, $\mathbf{d}$= 26cM, $\mathbf{d}$= 21cM e para N=2, $\mathbf{d}$= 23cM, para as funções Morgan, Haldane, Kosambi e Karlin, respectivamente[4].

Tabela 2.2 *Funções de distância citogenética.*

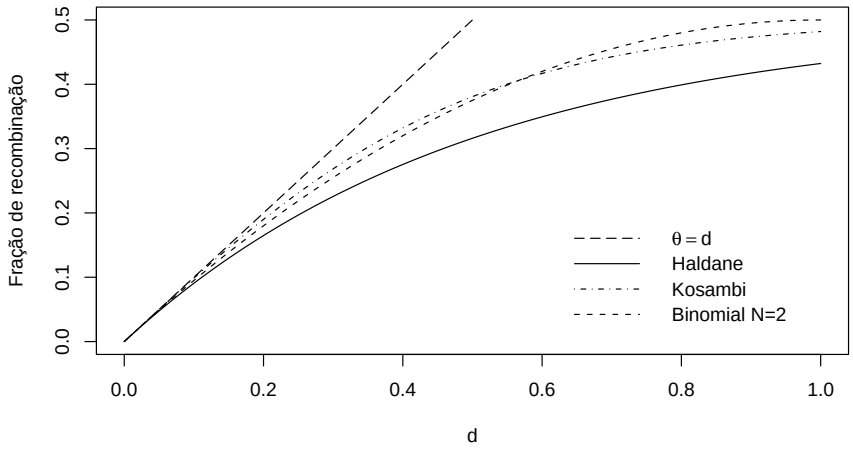
	Morgan	**Haldane**	**Kosambi**	**Karlin**
d	θ	$\begin{cases} -\frac{1}{2}\ln(1-2\theta), & 0 \le \theta \le 0{,}5 \\ \infty, & \text{caso contrário} \end{cases}$	$\frac{1}{2}tgh^{-1}(2\theta) = \frac{1}{4}\ln\frac{1+2\theta}{1-2\theta}$	$\frac{1}{2}N[1-(1-2\theta)^{\frac{1}{N}}]$
θ	d	$\frac{1}{2}[1-\exp(-2\lvert d\rvert)]$	$\frac{1}{2}tgh(2d) = \frac{1}{2}\frac{\exp(4d)-1}{\exp(4d)+1}$	$\begin{cases} \frac{1}{2}[1-(1-2d/N)^N], & d < \frac{N}{2} \\ \frac{1}{2}, & \text{caso contrário} \end{cases}$

[3] O nível descritivo associado a LOD=3 é igual a 0,0002016645.

[4] Para as funções de distância citogenética a unidade de medida M denota Morgan e cM denota centiMorgan.

Segue na Figura 2.2 o gráfico de algumas funções de distância segundo o parâmetro fração de recombinação θ. Outras funções de distância foram também propostas, veja, por exemplo, Felsenstein (1979) e Sturt (1976).

Figura 2.2 *Funções de distância citogenética.*



2.3 Desequilíbrio de Ligação

Os estudos de associação entre locos genéticos, visam detectar um possível desequilíbrio na distribuição das probabilidades alélicas dos locos envolvidos, que é uma dependência probabilística na distribuição dos alelos nos haplótipos. Medidas de desequilíbrio de ligação entre locos se apresentam como uma ferramenta útil para o mapeamento genético. O desequilíbrio de ligação entre locos ligados (próximos) tem vantagem em relação à análise de ligação, pois resulta num mapa de escala mais refinada. Contudo, o desequilíbrio de ligação não acontece somente em locos ligados, podendo ser gerado por algumas fontes como: estratificação ou miscigenação, mutação, efeito do fundador, seleção. Para a finalidade de mapeamento de variantes genéticas funcionais, isto é, mapeamento de genes funcionalmente associados a doenças, é essencial obter evidências de desequilíbrio de ligação entre locos ligados.

Considere o coeficiente de desequilíbrio de ligação (gamético) entre dois locos definido como

$$\mathrm{D} = p_{A_iB_k} - p_{A_i}p_{B_k}, \tag{2.3}$$

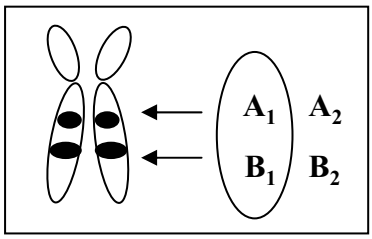
onde $p_{A_iB_k}$ representa a probabilidade do haplótipo A_iB_k; A_i e B_k são alelos de dois locos diferentes em um mesmo haplótipo, p_{A_i} e p_{B_k} representam as probabilidades dos alelos A_i e B_k, respectivamente. Então, quando D=0 tem-se o chamado equilíbrio de ligação, isto é, a independência (probabilística) na segregação dos locos envolvidos.

Uma das primeiras medidas de desequilíbrio comumente usada foi desenvolvida por Lewontin (1964), conforme citado em Jorde (2000). Considerando o modelo de desequilíbrio baseado em dois locos (Figura 2.3), tem-se os alelos A_1, A_2 no loco A, os alelos B_1 e B_2 no loco B, e os possíveis haplótipos A_1B_1, A_1B_2, A_2B_1 e A_2B_2. Considerando o haplótipo A_1B_1, a medida de desequilíbrio gamético é dada por:

$$D_{A_1B_1} = p_{A_1B_1} - p_{A_1}p_{B_1}, \tag{2.4}$$

onde $D_{A_1B_1} \in [D_{min}, D_{max}]$, tal que,
$D_{min} = -\max[p_{A_1}p_{B_1}, (1-p_{A_1})(1-p_{B_1})]$ e $D_{max} = \min[p_{A_1}(1-p_{B_1}), (1-p_{A_1})p_{B_1}]$, de tal forma que $D_{min} \geq -1$ e $D_{max} \leq 1$. D pode ser padronizado obtendo-se D', que é definido como $D' = \frac{D}{D_{max}}$.

Figura 2.3 *Composição dos alelos em dois locos genéticos.*



O coeficiente D pode ser interpretado como uma medida de covariância entre alelos em um haplótipo, isto é,

$$\mathrm{D} = \mathrm{Cov}[\mathbb{1}(A_1), \mathbb{1}(B_1)], \tag{2.5}$$

onde:

$\mathrm{Cov}[\mathbb{1}(A_1), \mathbb{1}(B_1)]$ é a covariância entre as variáveis $\mathbb{1}(A_1)$ e $\mathbb{1}(B_1)$, que é dado por $\mathrm{Cov}[\mathbb{1}(A_1), \mathbb{1}(B_1)] = I\!E[\mathbb{1}(A_1)\mathbb{1}(B_1)] - I\!E[\mathbb{1}(A_1)]I\!E[\mathbb{1}(B_1)]$ e

$$\mathbb{1}(A_1) = \begin{cases} 1, & \text{se } A_1 \text{ está presente} \\ 0, & \text{caso contrário} \end{cases} \quad \mathbb{1}(B_1) = \begin{cases} 1, & \text{se } B_1 \text{ está presente} \\ 0, & \text{caso contrário} \end{cases}$$

Tem-se:

$I\!E[\mathbb{1}(A_1)] = p(\mathbb{1}(A_1) = 1) = p_{A_1}$,

$I\!E[\mathbb{1}(B_1)] = p(\mathbb{1}(B_1) = 1) = p_{B_1}$ e

$I\!E[\mathbb{1}(A_1)\mathbb{1}(B_1)] = p(\mathbb{1}(A_1) = 1, \mathbb{1}(B_1) = 1) = p_{A_1B_1}$.

Após várias gerações, o gene da doença e somente aqueles locos muito ligados a ele no cromossomo original permanecem juntos, estas ligações representam exemplos de desequilíbrio de ligação (Liu, 1998). A Figura 2.4 mostra o decréscimo do desequilíbrio de ligação por gerações para diferentes frações de recombinação entre os locos.

Quando dois locos estão ligados com uma fração de recombinação θ, os coeficientes de desequilíbrio decrescem por um fator $(1 - \theta)$ a cada geração em cruzamentos aleatórios. A expressão do desequilíbrio na *t-ésima* geração é dada por:

$$D^t = (1 - \theta)^t D^0, \tag{2.6}$$

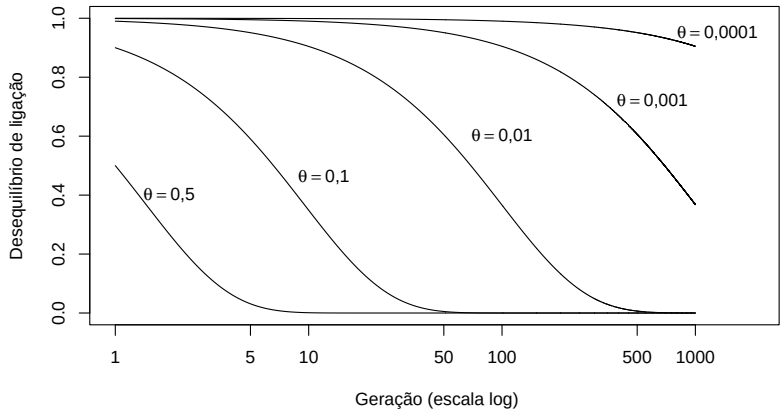
em que t é a geração atual; D^t é o valor do desequilíbrio para a geração t; D^0 é o valor do desequilíbrio para a geração 0.

A equação 2.7 a seguir (Hill & Robertson, 1968), especifica que o desequilíbrio torna-se muito pequeno com o número de gerações e/ou torna-se grande com a fração de recombinação (Jorde, 1995).

$$D^2 = \frac{1}{4N_e\theta + 1}, \tag{2.7}$$

onde N_e é o tamanho amostral efetivo.

Figura 2.4 *Decréscimo do desequilíbrio de ligação por gerações.*
$\theta = 0{,}5 \Rightarrow$ não existe ligação genética entre os locos;
$\theta = 0 \Rightarrow$ existe ligação genética entre os locos.



Quando na população são coletados dados de genótipos é possível verificar associação entre locos, por uma medida definida como *desequilíbrio de ligação composto* dada por (Weir, 1996):

$$\Delta_{A_1B_1} = D_{A_1B_1} + D_{A_1/B_1}, \tag{2.8}$$

onde $D_{A_1/B_1} = p_{A_1/B_1} - p_{A_1}p_{B_1}$ (desequilíbrio de ligação não gamético), com $p_{A_1/B_1} = p_{A_1A_1B_1B_1} + \frac{1}{2}[p_{A_1A_1B_1B_2} + p_{A_1A_2B_1B_1} + p_{A_2A_1B_1B_2}]$, ou seja, $\Delta_{A_1B_1}$ é a soma dos coeficientes gaméticos e não gaméticos, então: $\Delta_{A_1B_1} = p_{A_1B_1} + p_{A_1/B_1} - 2p_{A_1}p_{B_1}$.

Para realizar inferências sobre o parâmetro Δ, suponha nove classes genotípicas numeradas como na Tabela 2.3. Um estimador para Δ é dado por (Weir et al., 2004):

$$\widehat{\Delta} = \frac{1}{n}[n_{A_1A_1B_1B_1} + n_{A_1A_1B_1B_2} + n_{A_1A_2B_1B_1} + \frac{1}{2}n_{A_1A_2B_1B_2}] - 2\hat{p}_{A_1}\hat{p}_{B_1}, \tag{2.9}$$

onde $\hat{p}_{A_1}$ e $\hat{p}_{B_1}$ são as probabilidades amostrais para os alelos A_1 e B_1, respectivamente.

Tabela 2.3 *Classes genotípicas.*

	B_1B_1	B_1B_2	B_2B_2
A_1A_1	$n_{A_1A_1B_1B_1}$	$n_{A_1A_1B_1B_2}$	$n_{A_1A_1B_2B_2}$
A_1A_2	$n_{A_1A_2B_1B_1}$	$n_{A_1A_2B_1B_2}$	$n_{A_1A_2B_2B_2}$
A_2A_2	$n_{A_2A_2B_1B_1}$	$n_{A_2A_2B_1B_2}$	$n_{A_2A_2B_2B_2}$

O coeficiente de desequilíbrio composto é também uma medida de covariância entre locos genéticos. Defina x_{ij} e y_{ij}, $i = 1, ..., n$ (*i-ésimo* indivíduo amostrado), $j = 1, 2$ (*j-ésimo* haplótipo), como variáveis aleatórias com, os $x's$ valendo 0 se o haplótipo carrega A_1 e 1 se o haplótipo carrega A_2 e os $y's$ valendo 0 se o haplótipo carrega B_1 e 1 se o haplótipo carrega B_2. Seja $X_i = (x_{i1} + x_{i2})/2$ e $Y_i = (y_{i1} + y_{i2})/2$, tem-se, $\Delta = \text{Cov}(X_i, Y_i)$ sendo o coeficiente de correlação dado por:

$$\rho = \frac{\Delta_{A_1B_1}}{\sqrt{(\pi_{A_1} + D_{A_1})(\pi_{B_1} + D_{B_1})}}, \tag{2.10}$$

onde, $\pi_{A_1} = p_{A_1}(1 - p_{A_1})$; $\pi_{B_1} = p_{B_1}(1 - p_{B_1})$, com $D_{A_1} = p_{A_1A_1} - {p^2}_{A_1}$ e $D_{B_1} = p_{B_1B_1} - {p^2}_{B_1}$ que são os desvios do EHW. Sob $H_0 : \Delta = 0$, denotando o estimador de máxima verossimilhança de (2.10) por $\hat{\rho}$, é possível mostrar que a distribuição assintótica de $n\hat{\rho}^2$ é χ^2_1. Em mapeamento genético, note que o coeficiente de correlação, ρ, assume o valor 1 somente quando o polimorfismo do loco marcador é o próprio polimorfismo funcional, isto é, o loco de marcador é o gene regulador.

Para a proposta de mapeamento genético por meio de estudos de associação, existem diferentes alternativas de coletas de dados, como os delineamentos observacionais que podem incorporar ou não a informação genética de indivíduos e seus parentes. Na próxima subseção será discutida a estrutura populacional, pois esta precisa ser contemplada na escolha do processo de amostragem.

2.4 Estrutura Populacional

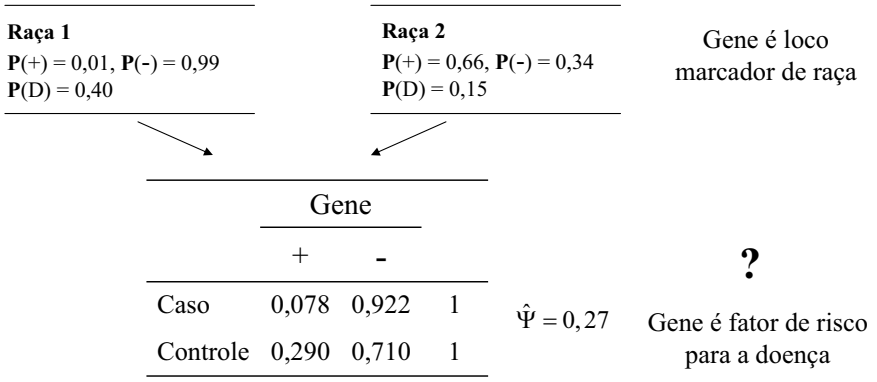
Em estudos de associação genética, as diferentes formas de coleta de dados dependem do conhecimento da estrutura da população, isto é, como esta se apresenta, se é homogênea ou possui estratos (variáveis que podem influenciar no resultado das análises). Por conseguinte, os delineamentos experimentais utilizados para coleta de dados genéticos, como os estudos transversais e retrospectivos, devem ser bem planejados tendo-se o cuidado de observar se a população é heterogênea, isto é, está estruturada. Existem delineamentos amostrais que "protegem" as estatísticas de teste do efeito de confundimento devido a variáveis que deveriam ser controladas no planejamento amostral. Por exemplo, o uso de dados familiares em estudos genéticos (Pritchard & Donnelly, 2001; Marchini et al., 2004).

Considere situações em que os dados provêm de diferentes populações ou de diferentes subdivisões de uma mesma população. Assim, tais amostras terão indivíduos pertencentes a estratos distintos. Em uma situação extrema considere um estudo caso-controle, em que a amostra do grupo *caso* está definida por indivíduos da raça 1 e a do grupo *controle* pela raça 2. Situações como estas podem causar problemas nos estudos de associação, já que é esperado que a constituição genética de indivíduos pertencentes a raças distintas seja diferente. Logo, os grupos de *caso* e *controle* se diferenciam não somente pelo estado da afecção, mas também por sua constituição genética. Deste modo, um loco genético específico de raça pode, erroneamente, ser identificado como fator de risco para a doença, como em um exemplo mostrado em Cardon & Palmer (2003), no qual observa-se que houve um desbalanceamento na constituição dos grupos (Figura 2.5).

Em Epidemiologia Genética é mais comum que fatores como etnia conduzam ao confundimento nos estudos de associação, mas podem ocorrer outras possibilidades. Por exemplo, um gene associado com a tendência do tabagismo pode estar associado com o câncer de pulmão. Logo, em estudos objetivando a identificação de locos genéticos diretamente associados com a biologia do câncer, o hábito de fumar pode ser considerado um fator de confundimento.

A chamada associação espúria, bastante enfatizada na literatura (vide Cardon & Palmer,

Figura 2.5 *Efeito de confundimento nos resultados de associação entre gene (+ e -) devido a raça e doença (D).*



2003), pode também ocorrer quando a probabilidade da doença varia nas subpopulações, implicando, por exemplo, no aumento da probabilidade de que indivíduos afetados sejam amostrados. Várias propostas para proteger os dados deste tipo de contaminação existem na literatura, por exemplo, os indivíduos devem ter tido a mesma oportunidade de serem expostos ao fator de risco genético, como é o caso de membros da mesma família. Como ilustração, considere que filhos afetados constituem o grupo *caso* e pais não afetados o grupo *controle*. A formação de grupos *caso* e *controle* que possuem certas características em comum (devido ao grau de parentesco) é feita de modo a evitar o impacto do confundimento de possíveis fatores genéticos, além de outros, quando se estuda populações estruturadas. Outras possibilidades de controle da heterogeneidade (genética) entre grupos pelo uso de pares de parentes na constituição dos grupos é proposto por Boehnke & Langefeld (1998) que constroem os testes de associação baseados em pares de irmãos discordantes, DSPs (do inglês, *discordant sib pairs*), ou seja, no lugar de utilizar pais não afetados como controles, coleta-se pares de irmãos, um afetado e outro não para constituir os grupos de *caso* e *controle*, respectivamente.

Como comentado por Pritchard & Donnelly (2001), em populações estruturadas pode existir uma alta taxa de associações significantes até mesmo entre locos genéticos não ligados (distantes). Uma forma de minimizar este problema é coletar dados sobre etnia dos membros da população e então obter a amostra de forma estratificada e realizar a análise

de acordo com grupos étnicos, na tentativa de assegurar que as probabilidades alélicas de locos do "*background*" genético[5] entre os grupos sejam as mesmas, reduzindo assim a ocorrência de associações falso-positivas. Contudo, só o controle da raça muitas vezes pode não resolver o problema de confundimento. Uma alternativa é genotipar indivíduos *caso* e *controle* para um mapa de marcadores moleculares e constituir o estudo por pareamento individual de tal forma que a distribuição dos genótipos (para o mapa adotado) não varie entre os genes. Tal balanceamento genético, apesar do alto custo, oferece maior segurança contra associações falso-positivas.

De maneira geral, o termo confundimento refere-se ao efeito que uma terceira variável (em geral, um fator de estratificação) pode exercer no estudo de associação entre duas variáveis. No caso de Genética, considere um estudo caso-controle em que se coletam indivíduos com diabetes (grupo *caso*) e sem diabetes (grupo *controle*), e observa-se em cada grupo a ocorrência dos alelos A_1 e A_2. Ao analisar os dados verifica-se a não existência de associação, isto é, o risco de diabetes para indivíduos que estão na categoria alélica A_1 é o mesmo que para a categoria A_2. Ao estratificar os grupos por raça (por exemplo), conclui-se pela associação, caracterizando assim a variável raça como um fator de "confundimento". Combinar dados de diferentes tabelas exige cuidados, pois pode ocorrer o chamado Paradoxo de Simpson, sob o qual o padrão de associação varia segundo os níveis de uma variável confundidora (veja, por exemplo, Paulino & Singer (2004), Agresti (2002)).

Devido à complexidade dos modelos de regulação molecular das doenças, vários estudos experimentais e observacionais são conduzidos para se atacar o problema de mapeamento genético de doenças. Observa-se na literatura que existem estudos que não apresentam reprodutibilidade, isto é, grupos de pesquisa diferentes adotando delineamentos experimentais equivalentes para avaliar a mesma hipótese de estudo chegam a conclusões divergentes. Isto pode ocorrer se o critério de coleta de dados e de análise não for rigoroso o suficiente, por exemplo, se os conjuntos de dados forem heterogêneos para fontes de variação não conhecidas que não puderam ser controladas de forma apropriada nos dife-

[5] *Background* genético: conjunto de muitos genes.

rentes estudos (Cardon & Palmer, 2003). Causas comuns para a não reprodutibilidade dos estudos são a heterogeneidade da constituição genética das amostras, os procedimentos laboratoriais de genotipagem e a categorização dos *status* da doença. Em particular algumas soluções para o problema de estratificação/confundimento nos estudos de associação genética decorrem de um rigoroso planejamento experimental e aplicação apropriada das estatísticas de teste envolvidas na análise dos dados. Estes pontos serão retomados no próximo capítulo.

Capítulo 3

Métodos Estatísticos na Análise de Associação Genética

Existem várias alternativas de planejamento experimental em que os estudos de associação genética podem ser conduzidos. Neste capítulo são descritos alguns delineamentos de interesse, bem como algumas metodologias de análise estatística.

3.1 Estudo Transversal

Associação entre Pares de Locos

Os estudos de associação entre pares de locos genéticos são relevantes na construção de mapas de marcadores moleculares, e estão implicitamente envolvidos com análises de associação entre um fator de risco genético (um loco genético) e uma doença. Nestes estudos espera-se que, se o marcador mostra um efeito significativo, o próprio loco sob estudo é o gene regulador ou que ele está ligado (próximo) ao gene regulador, mostrando assim um efeito "aparente" decorrente de uma proximidade e associação com o gene regulador.

Qualquer que seja a motivação para a análise de associação entre pares de locos, os estudos transversais representam o tipo de delineamento mais comumente utilizado, em que considera-se apenas o tamanho total da amostra como fixado. As unidades amostrais são então classificadas em categorias de interesse. Maiores detalhes sobre tais delineamentos podem ser encontrados, por exemplo, em Agresti (2002) e Fleiss, et al. (2003).

Neste tipo de estudo, seguindo a abordagem proposta por Sham (1998) e Ewens & Spielman (2003), sejam os locos A e B de marcadores autossômicos localizados no mesmo cromossomo, com alelos A_1, A_2, ..., A_m e B_1, B_2, ..., B_n, respectivamente. Para uma

amostra aleatória de indivíduos considere os genótipos $A_iA_jB_kB_l$ (i, j=1,2,..., m e k, l=1,2,..., n). Existem $m(m+1)/2$ e $n(n+1)/2$ possíveis genótipos para os locos A e B, respectivamente, então o número total de genótipos conjuntos é $[m(m+1)/2][n(n+1)/2]$. Os dados assim gerados podem ser dispostos em um formato de tabela de contingência.

Para a finalidade dos estudos de associação entre fator de risco genético, cujo fator observável é um marcador e uma doença, implicitamente pesquisa-se a existência de associação (desequilíbrio de ligação) entre o marcador (loco A, por exemplo) e o gene regulador (loco B, por exemplo).

Para uma amostra de $n_{....}$ indivíduos, seja a contagem do genótipo $A_iA_jB_kB_l$ denotada por n_{ijkl}. Então para dois locos A e B dialélicos os dados podem ser representados no formato da Tabela 3.1. Considerando $n_{....}$ indivíduos fixados e não relacionados, tem-se que o modelo de probabilidades Multinomial é adequado para expressar os dados:

$$P(N_{1111} = n_{1111}, ..., N_{2222} = n_{2222}) = \frac{n....!}{\prod_{i,j,k,l} n_{ijkl}!} \prod_{i,j,k,l} p_{ijkl}^{n_{ijkl}}, \tag{3.1}$$

onde, $p_{ijkl} = P(A_iA_jB_kB_l)$ são as probabilidades genotípicas, tal que $\sum_{i,j,k,l} p_{ijkl} = 1$ e $\sum_{i,j,k,l} n_{ijkl} = n_{....}$.

Portanto, o logaritmo da função de verossimilhança para dados deste tipo pode ser escrito como

$$\ln L = \ln L(p_{ijkl}) \propto \sum n_{ijkl} \ln(p_{ijkl}), \tag{3.2}$$

sendo o estimador de máxima verossimilhança de p_{ijkl} dado por $\widehat{p}_{ijkl} = n_{ijkl}/n_{....}$.

Sejam as mn probabilidades dos haplótipos dos locos A e B denotadas por h_{11}, h_{12}, ...,h_{mn}, com $h_{ik} = P(A_iB_k)$; as correspondentes probabilidades dos genótipos denotadas por p_{1111}, p_{1112}, ...,p_{mmnn}; as m probabilidades alélicas de A denotadas por p_1, p_2, ...,p_m e as n probabilidades alélicas de B, denotadas por q_1, q_2, ...,q_n. As probabilidades dos

Tabela 3.1 *Distribuição de freqüências genotípicas entre pares de locos.*

	Loco B			
Loco A	B_1B_1	B_1B_2	B_2B_2	Total
A_1A_1	n_{1111}	n_{1112}	n_{1122}	$n_{11..}$
A_1A_2	n_{1211}	n_{1212}	n_{1222}	$n_{12..}$
A_2A_2	n_{2211}	n_{2212}	n_{2222}	$n_{22..}$
Total	$n_{..11}$	$n_{..12}$	$n_{..22}$	$n_{....}$

genótipos conjuntos p_{ijkl} podem ser escritas da seguinte forma, sob equilíbrio de Hardy-Weinberg e equilíbrio de ligação (Sham, 1998):

$$\begin{aligned} p_{iikk} &= p_ip_iq_kq_k = p_i^2q_k^2, \\ p_{iikl} &= p_ip_iq_kq_l + p_ip_iq_lq_k = 2p_i^2q_kq_l, \\ p_{ijkk} &= p_ip_jq_kq_k + p_jp_iq_kq_k = 2p_ip_jq_k^2, \\ p_{ijkl} &= p_ip_jq_kq_l + p_ip_jq_lq_k + p_jp_iq_kq_l + p_jp_iq_lq_k = 4p_ip_jq_kq_l, \end{aligned} \tag{3.3}$$

para $i \neq j$, $k \neq l$, i e j=1,2,..., m e k e l=1,2,..., n.

Considere que a função de log-verossimilhança dada em (3.2) descreve o modelo saturado. As condições dadas em (3.3) podem ser impostas a (3.2) obtendo-se o logaritmo da verossimilhança do modelo restrito denotado por $\ln L_0$. Assim, tem-se a estatística da razão de verossimilhanças, dada por:

$$2(\ln L - \ln L_0), \tag{3.4}$$

em que (3.4), sob equilíbrio de ligação e equilíbrio de Hardy-Weinberg, segue assintoticamente uma distribuição Qui-Quadrado com $\{[m(m+1)/2][n(n+1)/2] - (m+n) + 1\}$ graus de liberdade. O teste assim definido, investiga conjuntamente as hipóteses de equilíbrio de ligação e equilíbrio de Hardy-Weinberg, considerando dados de um estudo transversal.

Afim de testar desequilíbrio de ligação sob equilíbrio de Hardy-Weinberg, é necessário considerar um modelo intermediário entre o modelo saturado ($\ln L$) e o restrito ($\ln L_0$). Neste caso, as expressões das probabilidades genotípicas se reduzem a:

$$\begin{aligned} p_{iikk} &= h_{ik}h_{ik} = h_{ik}^2, \\ p_{iikl} &= h_{ik}h_{il} + h_{il}h_{ik} = 2h_{il}h_{ik}, \\ p_{ijkk} &= h_{ik}h_{jk} + h_{jk}h_{ik} = 2h_{ik}h_{jk}, \\ p_{ijkl} &= h_{ik}h_{jl} + h_{jl}h_{ik} + h_{il}h_{jk} + h_{jk}h_{il} = 2(h_{ik}h_{jl} + h_{il}h_{jk}), \end{aligned} \tag{3.5}$$

para $i \neq j$ e $k \neq l$.

Substituindo as restrições dadas em (3.5) na função de log-verossimilhança (3.2), obtém-se $\ln L_1$. Neste caso, tem-se a estatística da razão de verossimilhanças, dada por:

$$2(\ln L_1 - \ln L_0), \tag{3.6}$$

em que (3.6), sob equilíbrio de ligação segue assintoticamente uma distribuição Qui-Quadrado com $(m-1)(n-1)$ graus de liberdade. A estatística dada em (3.6) pode ser adotada no teste condicional de equilíbrio de ligação entre dois locos, sob a suposição de equilíbrio de Hardy-Weinberg (Sham, 1998).

3.2 Estudo Caso-Controle

Associação entre Doença e um Fator de Risco Genético

A comparação das probabilidades dos alelos de marcadores moleculares em amostras de genótipos ou haplótipos, entre grupos de indivíduos normais e afetados é chamada de teste de associação genética (Weir, 1996). Os delineamentos caso-controle são mais comumente adotados, sendo também denominados de estudos retrospectivos, onde pessoas diagnosticadas no início do experimento como tendo uma certa característica (*caso*) são comparadas com pessoas que não têm a característica (*controle*). Pode-se definir, por

exemplo, os grupos de *caso* e *controle* como constituídos por indivíduos *com* e *sem* uma determinada doença de interesse como diabetes, hipertensão, obesidade, depressão entre outras, respectivamente. Os tamanhos amostrais dos dois grupos são fixados por planejamento e, em cada grupo, é observada uma resposta de interesse (a ocorrência de uma mutação ou estado alélico e genotípico de um loco genético, por exemplo).

Para este estudo, pela abordagem proposta por Terwilliger & Ott (1994), com o objetivo de testar se existe associação entre a doença e um fator de risco genético, coleta-se uma amostra de indivíduos não relacionados afetados com uma certa doença e compara-se a probabilidade de ocorrência de certos genótipos ou alelos com uma amostra de indivíduos normais não relacionados, sendo os indivíduos afetados e normais amostrados da mesma população. A seguir considere duas abordagens de análises: no nível genotípico e no nível cromossômico.

3.2.1 Analíse no Nível Genotípico (indivíduo)

Quando se considera os dados de marcadores dialélicos na forma de genótipos, a análise é feita no nível de indivíduo (amostra de tamanho n), sendo o risco da doença definido em função do genótipo. Neste caso, não é necessário assumir EHW entre locos que compõem o genótipo. Para a análise no nível genotípico (indivíduo) considere a Tabela 3.2.

Tabela 3.2 *Estudo caso-controle no nível genotípico.*

	Genótipo			
	A_1A_1	A_1A_2	A_2A_2	Total
Caso	$n^{(1)}_{A_1A_1}$	$n^{(1)}_{A_1A_2}$	$n^{(1)}_{A_2A_2}$	$n^{(1)}_{\cdot}$
Controle	$n^{(0)}_{A_1A_1}$	$n^{(0)}_{A_1A_2}$	$n^{(0)}_{A_2A_2}$	$n^{(0)}_{\cdot}$

Note que, considerando a Tabela 3.2, a estatística Qui-Quadrado clássica para o estudo de associação pode ser adotada:

$$\chi^2 = \sum_{i=1}^{3}\sum_{j=0}^{1} \frac{(n_i^{(j)} - e_i^{(j)})^2}{e_i^{(j)}}, \tag{3.7}$$

em que (3.7), sob H_0: P(Doença|Genótipo)= P(Não Doença|Genótipo), tem distribuição Qui-Quadrado com 2 graus de liberdade, sendo $n_i^{(j)}$ a freqüência observada na amostra, $e_i^{(j)} = \frac{(n_\cdot^{(j)} \times n_i^{(\cdot)})}{n_\cdot^{(\cdot)}}$ com i índice de genótipo, j índice de grupo e $n_i^{(\cdot)}$, $n_\cdot^{(j)}$ e $n_\cdot^{(\cdot)}$ indicando o tamanho amostral da linha, coluna e total, respectivamente.

Uma alternativa mais apropriada de verificar associação nestes estudos é calcular o parâmetro "razão de chances" (*Odds Ratio*) denotado por Ψ. Para a Tabela 3.3, têm-se os dois *Odds Ratios* que descrevem associação entre doença e marcador nos níveis genotípicos (com referência ao A_2A_2), dados por:

$$\Psi_{A_1A_1|A_2A_2} = \frac{p^{(1)}_{A_1A_1} p^{(0)}_{A_2A_2}}{p^{(0)}_{A_1A_1} p^{(1)}_{A_2A_2}}, \ \Psi_{A_1A_2|A_2A_2} = \frac{p^{(1)}_{A_1A_2} p^{(0)}_{A_2A_2}}{p^{(0)}_{A_1A_2} p^{(1)}_{A_2A_2}}, \tag{3.8}$$

em que, $\Psi_{A_1A_1|A_2A_2}$ representa o risco relativo da doença para indivíduos que carregam o genótipo A_1A_1 e $\Psi_{A_1A_2|A_2A_2}$ representa o risco relativo da doença para indivíduos que carregam o genótipo A_1A_2, considerando o genótipo A_2A_2 como referência.

Tabela 3.3 *Estudo caso-controle no nível genotípico (probabilidades).*

	Genótipo			
	A_1A_1	A_1A_2	A_2A_2	Total
Caso	$p^{(1)}_{A_1A_1}$	$p^{(1)}_{A_1A_2}$	$p^{(1)}_{A_2A_2}$	1
Controle	$p^{(0)}_{A_1A_1}$	$p^{(0)}_{A_1A_2}$	$p^{(0)}_{A_2A_2}$	1

Os estimadores de máxima verossimilhança destes parâmetros assumindo o modelo produto de Trinomiais são dados por:

$$\hat{\Psi}_{A_1A_1|A_2A_2} = \frac{n^{(1)}_{A_1A_1} n^{(0)}_{A_2A_2}}{n^{(0)}_{A_1A_1} n^{(1)}_{A_2A_2}}, \ \hat{\Psi}_{A_1A_2|A_2A_2} = \frac{n^{(1)}_{A_1A_2} n^{(0)}_{A_2A_2}}{n^{(0)}_{A_1A_2} n^{(1)}_{A_2A_2}}. \tag{3.9}$$

Tendo em vista esses conceitos um modelo útil para o teste de associação entre fator de risco genético e doença é o de regressão logística. A flexibilidade deste modelo para análise de associação se estende, por exemplo, para permitir a análise no nível de dados genotípicos (indivíduo) ou alélicos (cromossômico). No caso de dados genotípicos considere

a variável Y definida como, Y=1 para *casos* e Y=0 para *controles* e pode-se categorizar as variáveis X_1 e X_2 como:

$$X_1 = \begin{cases} 1, & \text{se o indivíduo carrega o genótipo } A_1A_1 \\ 0, & \text{caso contrário} \end{cases}$$

$$X_2 = \begin{cases} 1, & \text{se o indivíduo carrega o genótipo } A_1A_2 \\ 0, & \text{caso contrário.} \end{cases}$$

Deste modo, pode-se considerar o seguinte modelo de regressão logística:

$$\ln\left(\frac{p_1(\mathbf{x})}{1-p_1(\mathbf{x})}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2, \tag{3.10}$$

com $p_j(\mathbf{x})$ a probabilidade de um indivíduo do grupo genotípico $\mathbf{x}$ pertencer ao *j-ésimo* nível da variável Y.

Logo o risco para o genótipo A_1A_1 é dado por:

$$\frac{p_1(x_1=1, x_2=0)}{1-p_1(x_1=1, x_2=0)} = e^{\alpha+\beta_1}. \tag{3.11}$$

Da mesma maneira, o risco para o genótipo A_1A_2 é dado por:

$$\frac{p_1(x_1=0, x_2=1)}{1-p_1(x_1=0, x_2=1)} = e^{\alpha+\beta_2}. \tag{3.12}$$

Enquanto que o risco para o genótipo A_2A_2 é dado por:

$$\frac{p_1(x_1=0, x_2=0)}{1-p_1(x_1=0, x_2=0)} = e^{\alpha}. \tag{3.13}$$

Deste modo, os *Odds Ratios* são dados por:

$$\begin{aligned} \Psi_{A_1A_1|A_2A_2} &= e^{\beta_1}, \ \text{ para } \ x_1=1 \ \text{ e } \ x_2=0; \\ \Psi_{A_1A_2|A_2A_2} &= e^{\beta_2}, \ \text{ para } \ x_1=0 \ \text{ e } \ x_2=1. \end{aligned} \tag{3.14}$$

Sob a parametrização (3.10), os parâmetros β_1 e β_2 podem ser escritos como:

$$\begin{aligned}\beta_1 &= \ln(\Psi_{A_1A_1|A_2A_2});\\ \beta_2 &= \ln(\Psi_{A_1A_2|A_2A_2}).\end{aligned} \tag{3.15}$$

Observa-se que β_1 e β_2 podem também ser escritos via "*logitos*":

$$\begin{aligned}\text{logito}_{A_1A_1} &= \ln\frac{p_1(x_1=1,x_2=0)}{1-p_1(x_1=1,x_2=0)} = \alpha+\beta_1;\\ \text{logito}_{A_1A_2} &= \ln\frac{p_1(x_1=0,x_2=1)}{1-p_1(x_1=0,x_2=1)} = \alpha+\beta_2;\\ \text{logito}_{A_2A_2} &= \ln\frac{p_1(x_1=0,x_2=0)}{1-p_1(x_1=0,x_2=0)} = \alpha.\end{aligned} \tag{3.16}$$

Portanto: $\beta_1 = \text{logito}_{A_1A_1} - \text{logito}_{A_2A_2}$ e $\beta_2 = \text{logito}_{A_1A_2} - \text{logito}_{A_2A_2}$.

É bem conhecida na literatura a formalização dos modelos como (3.10) por meio de notação matricial (ver Paulino & Singer, 2004; Agresti, 2002; Grizzle et al., 1969). Neste caso o modelo logístico linear fica dado por:

$$\mathbf{F}(\mathbf{p}) = \mathbf{X}\boldsymbol{\beta},$$

tal que,

$$\mathbf{F}(\mathbf{p}) = \begin{pmatrix}\text{logito}_{A_1A_1}\\ \text{logito}_{A_1A_2}\\ \text{logito}_{A_2A_2}\end{pmatrix},\ \mathbf{X} = \begin{pmatrix}1&1&0\\1&0&1\\1&0&0\end{pmatrix},\ \boldsymbol{\beta} = \begin{pmatrix}\alpha\\ \beta_1\\ \beta_2\end{pmatrix},$$

onde $\mathbf{F}(\mathbf{p})$ representa um vetor contendo as funções logísticas definidas para os dados;

$\mathbf{X}$ é a matriz de delineamento;

$\boldsymbol{\beta}$ é o vetor de parâmetros.

A hipótese de não existência de associação é H_0: $\beta_1 = \beta_2 = 0$, que, da teoria de modelos logísticos lineares, pode ser testada pelo método de máxima verossimilhança (não condicional e condicional) ou pelo método de mínimos quadrados ponderados. Maiores

detalhes ver Paulino & Singer (2004), Agresti (2002), Breslow & Day (1980), Prentice & Pyke (1979), Bishop et al. (1975), Grizzle et al. (1969).

Pode-se fazer uma restrição na equação (3.10), ou seja, assumindo que o modelo genético é multiplicativo, isto é, que o risco relativo da doença para o genótipo A_1A_1 (que carrega 2 alelos A_1) é duas vezes maior (na escala logarítmica) que o risco relativo para o genótipo A_1A_2 (que carrega uma cópia do alelo A_1). Neste caso, impõe-se a restrição de efeito linear do número de alelos no genótipo. Deste modo é possível obter um modelo de regressão logística reduzido. Tal modelo é obtido via a seguinte restrição:

$$\Psi_{A_1A_1|A_2A_2} = \Psi^2_{A_1A_2|A_2A_2}. \tag{3.17}$$

Aplicando a transformação logarítmica na equação (3.17) tem-se:

$$\ln \Psi_{A_1A_1|A_2A_2} = 2 \ln \Psi_{A_1A_2|A_2A_2}. \tag{3.18}$$

Como visto, a partir da equação (3.10), tem-se (3.15). Logo de (3.18) obtém-se:

$$\beta_1 = 2\beta_2, \tag{3.19}$$

e, portanto, sob o modelo genético multiplicativo:

$$\begin{aligned} \text{logito}_{A_1A_1} &= \alpha + \beta_1 = \alpha + 2\beta_2; \\ \text{logito}_{A_1A_2} &= \alpha + \beta_2; \\ \text{logito}_{A_2A_2} &= \alpha. \end{aligned} \tag{3.20}$$

Portanto, em notação matricial o modelo genético multiplicativo é da forma:

$$\mathbf{F}(\mathbf{p}) = \begin{pmatrix} \text{logito}_{A_1A_1} \\ \text{logito}_{A_1A_2} \\ \text{logito}_{A_2A_2} \end{pmatrix}, \ \mathbf{X} = \begin{pmatrix} 1 & 2 \\ 1 & 1 \\ 1 & 0 \end{pmatrix}, \ \boldsymbol{\beta} = \begin{pmatrix} \alpha \\ \beta_2 \end{pmatrix}.$$

Observa-se, que quando $\beta_2 = 0$, tem-se a não existência de associação entre o fator de

risco genético (marcador) e a doença. Para testar H_0: $\beta_2 = 0$ pode-se utilizar a teoria de máxima verossimilhança ou mínimos quadrados ponderados.

Nota-se que para o modelo dado em (3.10) o teste de associação é realizado considerando 2 graus de liberdade porém, ao fazer a restrição (3.17), o teste é realizado considerando 1 grau de liberdade, ou seja, houve redução da quantidade de parâmetros ao assumir o efeito linear do número de alelos no genótipo.

3.2.2 Análise no Nível Cromossômico

A Tabela 3.4 apresenta um formato geral considerando dados de um estudo caso-controle onde se observa a ocorrência do alelo A_1 ou A_2 nos dois grupos, *caso* e *controle*. Neste caso, a hipótese de não associação é dada por H_0: P(Doença|Alelo marcador)= P(Não Doença|Alelo marcador), ou seja, a igualdade das probabilidades de incidência ou não da doença nos indivíduos portadores do alelo marcador.

Tabela 3.4 *Estudo caso-controle no nível cromossômico.*

	Alelo Marcador		
	A_1	A_2	Total
Caso	n_{11}	n_{12}	$n_{1.}$
Controle	n_{21}	n_{22}	$n_{2.}$

No caso de tabelas 2 × 2 para o estudo de associação a estatística Qui-Quadrado clássica (como dada em 3.7) também pode ser adotada em que, sob H_0, tem distribuição Qui-Quadrado com 1 grau de liberdade. Note que, fazendo um paralelo com os dados de genótipo (Tabela 3.2), na análise alélica tem-se $n_{11} = 2n^{(1)}_{A_1A_1} + n^{(1)}_{A_1A_2}$, $n_{12} = 2n^{(1)}_{A_2A_2} + n^{(1)}_{A_1A_2}$, $n_{21} = 2n^{(0)}_{A_1A_1} + n^{(0)}_{A_1A_2}$ e $n_{22} = 2n^{(0)}_{A_2A_2} + n^{(0)}_{A_1A_2}$, portanto $n_{1.} = 2n^{(1)}_{\cdot}$ e $n_{2.} = 2n^{(0)}_{\cdot}$, indicando o tamanho amostral de *casos* e *controles*, respectivamente.

Outra alternativa de verificar associação nestes estudos é calcular o parâmetro razão de chances (*Odds Ratio*) denotado por Ψ. Para uma tabela 2 × 2 (ver Tabela 3.5) tem-se o *Odds Ratio*, dado pela seguinte razão:

$$\Psi = \Psi_{A_1|A_2} = \frac{p_{11}/p_{21}}{p_{12}/p_{22}} = \frac{p_{11}p_{22}}{p_{21}p_{12}}. \tag{3.21}$$

Para dados dispostos em uma tabela 2×2, independentemente do tipo do estudo (transversal, prospectivo ou retrospectivo), o estimador do parâmetro Ψ é dado por (Breslow & Day, 1980):

$$\widehat{\Psi} = \frac{n_{11}n_{22}}{n_{21}n_{12}}. \tag{3.22}$$

O parâmetro Ψ é também conhecido como razão de produtos cruzados. Note que $\Psi = 1$ representa não associação entre doença e fator de risco genético, $\Psi > 1$ associação positiva e $\Psi < 1$ associação negativa. A vantagem de se adotar Ψ como medida de associação em tabelas de contingência é devido à sua invariância relativamente ao tipo de estudo observacional. Maiores detalhes podem ser encontrados em Agresti (2002) e Paulino & Singer (2004), por exemplo.

Tabela 3.5 *Estudo caso-controle no nível cromossômico (probabilidades).*

	Alelo Marcador		
	A_1	A_2	Total
Caso	p_{11}	p_{12}	1
Controle	p_{21}	p_{22}	1

A teoria de regressão logística aplicada aos estudos de associação genética estende-se também para a análise dos dados no nível cromossômico. Para sua descrição considere a variável Y definida como, Y=1 para *casos* e Y=0 para *controles*. Considere a variável preditora X, categorizada como:

$$X = \begin{cases} 1, & \text{se o indivíduo carrega o alelo } A_1 \\ 0, & \text{caso contrário} \end{cases}$$

Neste caso, seja $p_j(x)$ a probabilidade de indivíduos do grupo alélico x pertencer ao j-ésimo nível da variável Y. Assim, a probabilidade do indivíduo que carrega o alelo no nível x ter a doença (Y=1) é dada por:

$$p_1(x) = P(Y = 1|x). \tag{3.23}$$

A seguinte parametrização pode ser adotada:

$$p_1(x) = \frac{1}{1 + e^{-(\alpha+\beta_2 x)}}, \tag{3.24}$$

onde α e β_2 são parâmetros desconhecidos. Deste modo, obtém-se o seguinte modelo de regressão logística:

$$\ln\left(\frac{p_1(x)}{1 - p_1(x)}\right) = \alpha + \beta_2 x. \tag{3.25}$$

Logo, o risco de um indivíduo ter a doença quando carrega o alelo A_2 (no nível x=1) é dado por:

$$\frac{p_1(x = 1)}{1 - p_1(x = 1)} = e^{\alpha+\beta_2}, \tag{3.26}$$

enquanto que o risco do indivíduo ter a doença quando carrega o alelo A_1 (no nível x=0) é:

$$\frac{p_1(x = 0)}{1 - p_1(x = 0)} = e^{\alpha}. \tag{3.27}$$

Então o *Odds Ratio* é dado por:

$$\Psi = \frac{P(Y = 1|x = 1)/P(Y = 0|x = 1)}{P(Y = 1|x = 0)/P(Y = 0|x = 0)} = \frac{p_1(1)/p_0(1)}{p_1(0)/p_0(0)} = \frac{p_1(1)p_0(0)}{p_0(1)p_1(0)} = e^{\beta_2}. \tag{3.28}$$

Observa-se que β_2 pode também ser escrito via "*logitos*":

$$\begin{aligned} \text{logito}_{A_1} &= \ln \frac{p_1(x=1)}{1-p_1(x=1)} = \alpha + \beta_2 \text{ e} \\ \text{logito}_{A_2} &= \ln \frac{p_1(x=0)}{1-p_1(x=0)} = \alpha, \end{aligned} \tag{3.29}$$

tal que $\beta_2 = \ln \Psi$.

Em notação matricial o modelo logístico linear neste caso é dado por:

$$\mathbf{F}(\mathbf{p}) = \begin{pmatrix} \text{logito}_{A_1} \\ \text{logito}_{A_2} \end{pmatrix}, \ \mathbf{X} = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}, \ \boldsymbol{\beta} = \begin{pmatrix} \alpha \\ \beta_2 \end{pmatrix}.$$

Observa-se que quando $\beta_2 = 0$ ($\Psi = 1$) tem-se a não existência de associação entre o fator de risco genético (marcador) e a doença, isto é, o risco da doença é o mesmo independentemente do estado alélico. Para testar H_0: $\beta_2 = 0$ pode-se utilizar resultados da teoria de máxima verossimilhança ou mínimos quadrados ponderados. Qualquer que seja a alternativa a estatística do teste tem distribuição Qui-Quadrado com 1 grau de liberdade.

Fazendo um paralelo entre as análises de associação no nível genotípico e cromossômico, considere que pela análise no nível cromossômico, na qual estima-se riscos alélicos da doença, também é possível obter estimativas de riscos genotípicos da doença, o que também pode ser obtido diretamente via a análise genotípica como visto na seção anterior. Seja o parâmetro Ψ definido como em 3.21, assumindo que o modelo genético é multiplicativo, impõe-se as seguintes restrições:

$$\begin{aligned} \Psi_{A_1A_2|A_2A_2} &= \Psi_{A_1|A_2} = \Psi; \\ \Psi_{A_1A_1|A_2A_2} &= \Psi^2_{A_1|A_2} = \Psi^2. \end{aligned} \tag{3.30}$$

A Tabela 3.6 apresenta os *Odds Ratios* para os três modelos logísticos discutidos. É importante salientar o efeito linear do número de alelos no genótipo sob o modelo multiplicativo. Observa-se que o teste de associação para o nível genotípico é realizado considerando 2 graus de liberdade, porém ao fazer a restrição do modelo genético multiplicativo

o teste será realizado considerando 1 grau de liberdade, ou seja, houve a redução de um parâmetro. Nota-se também que para as estimativas no nível cromossômico é assumido o EHW.

Tabela 3.6 *Parâmetros dos modelos genéticos.*

Modelo	*gl*	Medidas de risco	
Nível genotípico	2	$\Psi_{A_1A_1\|A_2A_2}$	$\Psi_{A_1A_2\|A_2A_2}$
Nível genotípico (multiplicativo)	1		$\Psi_{A_1A_2\|A_2A_2}$ *
Nível cromossômico (EHW)	1		$\Psi_{A_1\|A_2}$ **

* $\Psi_{A_1A_1|A_2A_2} = \Psi^2_{A_1A_2|A_2A_2}$, sob (3.17)

** $\Psi_{A_1A_2|A_2A_2} = \Psi_{A_1|A_2}$ e $\Psi_{A_1A_1|A_2A_2} = \Psi^2_{A_1|A_2}$, sob (3.30)

Vale ressaltar que na prática os dados obtidos são efetivamente genotípicos, dado que a amostra refere-se a indivíduos (que são diplóides). Logo, as estimativas de risco genotípico são obtidas diretamente. Já na análise cromossômica, apesar do tamanho amostral ser duplicado, garantindo estimativas mais precisas, assume-se adicionalmente que o EHW está satisfeito.

3.2.3 Estrutura Populacional

Como já salientado anteriormente, um dos problemas em estudos de associação genética em populações estruturadas é a ocorrência de conclusões falso-positivas. Nos delineamentos caso-controle, por exemplo, uma variável chamada de fator de estratificação pode influenciar nos resultados das análises, o que tem sido extensivamente considerado na literatura da área de Epidemiologia bem como da Estatística (Newman, 2001). Neste contexto existem algumas técnicas para minimizar tais limitações. Nesta seção serão discutidas algumas delas.

A técnica de Cochran-Mantel-Haenszel é apropriada para a análise de associação em situações de populações estratificadas. A idéia é combinar as tabelas dispostas nos vários

estratos, propondo uma medida de associação comum, se for o caso. Hall et al. (2000) discutem o teste de Cochran-Mantel-Haenszel considerando desde tabelas $2 \times 2 \times K$ até tabelas $2 \times$ J $\times K$, onde J é o número de categorias de resposta e K é o número de estratos.

Para a análise no nível genotípico pode-se ter K níveis da variável de confundimento para uma tabela 2×3, como ilustrado na Tabela 3.7, ou seja, combinam-se as K tabelas e calcula-se a estatística de Cochran-Mantel-Haenszel.

Tabela 3.7 *Relação da variável de confundimento com caso e controle (2 × 3).*

Variável de confundimento		Genótipo Marcador			
		A_1A_1	A_1A_2	A_2A_2	Total
1	Caso	n_{111}	n_{112}	n_{113}	$n_{11.}$
	Controle	n_{121}	n_{122}	n_{123}	$n_{12.}$
2	Caso	n_{211}	n_{212}	n_{213}	$n_{21.}$
	Controle	n_{221}	n_{222}	n_{223}	$n_{22.}$
$\vdots$		$\vdots$	$\vdots$	$\vdots$	
K	Caso	n_{K11}	n_{K12}	n_{K13}	$n_{K1.}$
	Controle	n_{K21}	n_{K22}	n_{K23}	$n_{K2.}$

Da mesma maneira, mas para análise no nível cromossômico, pode-se dividir a população e combinar os parâmetros Ψ de cada estrato específico controlando o efeito de confundimento. Uma técnica deste tipo é o método de Mantel-Haenszel. Agresti (2002), por exemplo, explica a proposta de Mantel & Haenszel (1959) para a estimação do *Odds Ratio* comum. A Tabela 3.8 apresenta a distribuição de *casos* e *controles* nos níveis da variável de confundimento. A equação (3.31) refere-se à estimativa do *Odds Ratio* para tabelas $2 \times 2 \times K$ de acordo com Mantel-Haenszel.

$$\widehat{\Phi}_{MH} = \frac{\sum_{k=1}^{K}(n_{k11}n_{k22}/n_{k..})}{\sum_{k=1}^{K}(n_{k12}n_{k21}/n_{k..})}, \tag{3.31}$$

Tabela 3.8 *Relação da variável de confundimento com caso e controle (2 × 2).*

Variável de confundimento		Alelo Marcador A_1	Alelo Marcador A_2	Total
1	Caso	n_{111}	n_{112}	$n_{11.}$
	Controle	n_{121}	n_{122}	$n_{12.}$
2	Caso	n_{211}	n_{212}	$n_{21.}$
	Controle	n_{221}	n_{222}	$n_{22.}$
⋮		⋮	⋮	⋮
K	Caso	n_{K11}	n_{K12}	$n_{K1.}$
	Controle	n_{K21}	n_{K22}	$n_{K2.}$

em que $n_{k..}= n_{k1.}+n_{k2.}=n_{k11}+n_{k12}+n_{k21}+n_{k22}$.

A estatística de Mantel-Haenszel deve ser aplicada com cuidado, testando-se previamente se a medida de associação nos vários estratos não é conflitante, pois se for, a medida de associação comum não é informativa. O teste de associação conflitante pode ser feito por meio da estatística de pseudo homogeneidade. Maiores detalhes ver Agresti (2002).

O estudo de associação genética em populações estruturadas pode também ser realizado por meio de modelos logísticos lineares. A seguir, alguns destes modelos são considerados levando em conta a análise no nível genotípico e cromossômico.

3.2.3.1 Análise no nível genotípico

Modelo 1: A seguinte parametrização pode ser considerada para os dados dispostos no formato da Tabela 3.7:

$$\mathbf{F_1(p)} = \mathbf{X_1}\boldsymbol{\beta_1},$$

onde, para o *k-ésimo* estrato tem-se:

$$\mathbf{F}_{1k}(\mathbf{p}) = \begin{pmatrix} \text{logito}_{A_1A_{1k}} \\ \text{logito}_{A_1A_{2k}} \\ \text{logito}_{A_2A_{2k}} \end{pmatrix}, \ \mathbf{X}_{1k} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}, \ \boldsymbol{\beta}_{1k} = \begin{pmatrix} \alpha_k \\ \beta_{1k} \\ \beta_{2k} \end{pmatrix}.$$

Os logitos sendo dados por:

$$\begin{aligned} \text{logito}_{A_1A_{1_k}} &= \alpha_k + \beta_{1k}; \\ \text{logito}_{A_1A_{2_k}} &= \alpha_k + \beta_{2k}; \\ \text{logito}_{A_2A_{2_k}} &= \alpha_k. \end{aligned} \tag{3.32}$$

Logo, no modelo completo para os K estratos, a matriz $\mathbf{X}$ é bloco diagonal.

Neste caso, para o estrato k os *Odds Ratios* são dados por:

$$\Psi_{A_1A_1|A_2A_{2_k}} = e^{\beta_{1k}} \text{ e } \Psi_{A_1A_2|A_2A_{2_k}} = e^{\beta_{2k}}, \tag{3.33}$$

e, deste modo, as medidas de risco relativo da doença para cada genótipo variam de acordo com os estratos. Várias hipóteses de interesse podem ser testadas via contrastes do tipo $\mathbf{C}\boldsymbol{\beta} = \mathbf{0}$, por exemplo, a uniformidade dos *Odds Ratios* nos estratos, bem como a homogeneidade das medidas de risco genotípico dentro de cada estrato. Modelos mais simplificados podem ainda considerar que o logito de referência ($\text{logito}_{A_2A_2}$) não varia nos estratos.

Modelo 2: Supondo a restrição de modelo multiplicativo, isto é, efeito linear do número de alelos no genótipo, pode-se formular o seguinte modelo:

$$\mathbf{F_2(p)} = \mathbf{X_2\boldsymbol{\beta}_2},$$

onde, para o *k-ésimo* estrato tem-se:

$$\mathbf{F}_{2k}(\mathbf{p}) = \begin{pmatrix} \text{logito}_{A_1A_{1_k}} \\ \text{logito}_{A_1A_{2_k}} \\ \text{logito}_{A_2A_{2_k}} \end{pmatrix}, \ \mathbf{X}_{2k} = \begin{pmatrix} 1 & 2 \\ 1 & 1 \\ 1 & 0 \end{pmatrix}, \ \boldsymbol{\beta}_{2k} = \begin{pmatrix} \alpha_k \\ \beta_{2k} \end{pmatrix}.$$

Os logitos sendo dados por:

$$\begin{aligned} \text{logito}_{A_1A_{1k}} &= \alpha_k + 2\beta_{2k}; \\ \text{logito}_{A_1A_{2k}} &= \alpha_k + \beta_{2k}; \\ \text{logito}_{A_2A_{2k}} &= \alpha_k. \end{aligned} \tag{3.34}$$

Neste caso, os *Odds Ratios* são dados por:

$$\Psi_{A_1A_2|A_2A_2k} = e^{\beta_{2k}}, \tag{3.35}$$

tendo-se o interesse, por exemplo, em testar a uniformidade de medida de risco relativo ao longo dos estratos. Como ressaltado anteriormente, sob o modelo linear, o risco relativo da doença para o genótipo A_1A_1 é obtido como:

$$\Psi_{A_1A_1|A_2A_2k} = \Psi^2_{A_1A_2|A_2A_2k}. \tag{3.36}$$

Modelo 3: Um modelo alternativo pode ser formulado introduzindo-se diretamente o efeito dos fatores genéticos, das variáveis de confundimento e de possíveis termos de interações. Assumindo o modelo multiplicativo tem-se, por exemplo:

$$\text{logito}_{g_k} = \alpha + \beta_2 X_g + \gamma_k V_k + \delta_k (X_g * V_k), \tag{3.37}$$

com as restrições $\sum_{k=1}^{K} \gamma_k = 0$; $\sum_{k=1}^{K} \delta_k = 0$, onde $X_g = 0, 1, 2$ para os genótipos A_2A_2, A_1A_2 e A_1A_1, respectivamente, e V_k são variáveis associadas ao fator de confundimento tal que, $(X_g * V_k)$ mede o efeito de interação nos dados. Neste caso a medida de risco relativo da doença para o genótipo g é dada por:

$$\Psi_{g|A_2A_2k} = e^{\beta_2 X_g + \delta_k (X_g V_k)}, \tag{3.38}$$

que é função das variáveis de confundimento, isto é:

$$\begin{aligned} \Psi_{A_1A_1|A_2A_2k} &= e^{2\beta_2+2\delta_k V_k} \\ \Psi_{A_1A_2|A_2A_2k} &= e^{\beta_2+\delta_k V_k}. \end{aligned} \tag{3.39}$$

Se $\delta_k = 0$, $\forall$ k= 1,...,K, não existe interação entre o fator de risco genético e a variável de confundimento.

3.2.3.2 Análise no nível cromossômico

Considerando os dados dispostos como na Tabela 3.8, um paralelo pode ser estabelecido diretamente com os Modelos 1 e 3, formalizados no contexto cromossômico, isto é, sob medidas de risco relativo da doença para os alelos A_1 e A_2. Por exemplo, o Modelo 1 fica definido alternativamente como:

$$\mathbf{F_I(p)} = \mathbf{X_I}\boldsymbol{\beta_I},$$

onde, para o *k-ésimo* estrato tem-se:

$$\mathbf{F}_{Ik}(\mathbf{p}) = \begin{pmatrix} \text{logito}_{A_{1k}} \\ \text{logito}_{A_{2k}} \end{pmatrix}, \ \mathbf{X}_{Ik} = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}, \ \boldsymbol{\beta}_{Ik} = \begin{pmatrix} \alpha_k \\ \beta_{2k} \end{pmatrix}.$$

Os logitos sendo dados por:

$$\begin{aligned} \text{logito}_{A_{1k}} &= \alpha_k + \beta_{2k}; \\ \text{logito}_{A_{2k}} &= \alpha_k. \end{aligned} \tag{3.40}$$

No caso do Modelo 3 definido para dados cromossômicos, a única alteração na expressão 3.37 reside em X_g= 0, 1 para a ocorrência dos alelos A_2 e A_1, respectivamente.

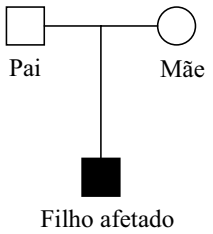
Modelos logísticos lineares são bastante flexíveis para acomodar efeitos de variáveis de confundimento, as quais, em genética, em geral, correspondem a etnias ou a locos de

marcadores moleculares de um mapa de referência usado para controle da heterogeneidade genética entre grupos. Contudo, na prática, é difícil garantir a homogeneidade dos grupos *caso* e *controle* para variáveis moleculares, que não são diretamente observáveis e, portanto não podem ser controladas nem por delineamento e nem via modelagem estatística. Uma alternativa é o uso de dados familiares na definição dos grupos *caso* e *controle*, como é visto na próxima seção.

3.3 Teste de Desequilíbrio de Transmissão (TDT)

O teste de desequilíbrio de transmissão (TDT) é outra ferramenta para a análise de associação genética. O TDT não é afetado pelo efeito de variáveis de confundimento, como no caso de populações estruturadas, que podem induzir a evidências falso-positivas (Ewens & Spielman, 2003). Para o TDT os dados amostrais possuem o mesmo "*background*" genético, pois neste teste as amostras são baseadas em trios: pequenos núcleos familiares compostos do pai, mãe e filho afetado (Figura 3.1).

Figura 3.1 *Amostra de trio.*



Em estudos caso-controle de associação genética uma das preocupações é se as amostras para os grupos de *caso* e *controle* são coletadas de uma mesma população genética. Contudo, como é difícil obter tais amostras e algumas soluções são propostas como, por exemplo, os delineamentos chamados de Risco Relativo do Haplótipo no Nível Genotípico denotado por $GHRR$, propostos por Rubinstein et al. (1981) e os sugeridos por Terwilliger & Ott (1992 e 1994) denominados Risco Relativo do Haplótipo no Nível Cromossômico,

indicado por $HHRR$. As seções a seguir consideram tais propostas que, basicamente, se utilizam de diferentes leituras dos dados genéticos dispostos em trios.

3.3.1 Risco Relativo do Haplótipo no Nível Genotípico ($GHRR$)

A idéia para este método é coletar uma amostra aleatória de indivíduos afetados juntamente com seus pais (não afetados), ou seja, a base da análise é o estudo da segregação de alelos nos trios. O intuito é amostrar *casos* e *controles* da mesma população genética (o núcleo familiar trio) e avaliar o risco relativo do haplótipo.

Considerando um marcador sob estudo, o genótipo do filho afetado é considerado como um ponto amostral do grupo "*caso*" e os dois alelos paternos que não foram transmitidos para o filho afetado são considerados um ponto amostral do grupo "*controle*". Desta maneira tem-se as amostras de uma mesma população genética.

Como ilustração considere a Figura 3.2 e o *Trio 1*, os genótipos dos pais são GH e HJ e do filho é HH, então o genótipo do filho afetado é considerado como um ponto amostral "*caso*" (alelos transmitidos) e os dois alelos paternos que não foram trasmitidos para o filho são considerados na amostra "*controle*" (alelos não transmitidos), neste caso os alelos G e J.

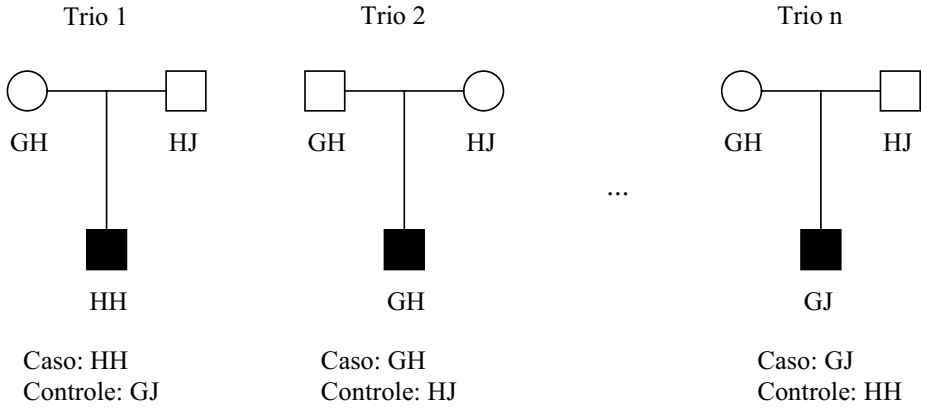
Para a construção da Tabela 3.9, considere novamente a Figura 3.2. Observa-se que o *Trio 1* contribuirá com uma observação na casela W e outra na casela Z, o *Trio 2* contribuirá com uma observação na casela W e outra na casela Y e o *Trio n* contribuirá com uma observação na casela X e outra na casela Y.

O teste de associação genética será dado por um teste Qui-Quadrado clássico de independência em tabelas de contingência, definido em termos das freqüências dos alelos transmitidos e não transmitidos (Tabela 3.9). A estatística do teste é dada por:

$$\chi^2 = \frac{2N(WZ - XY)^2}{(W+X)(W+Y)(X+Z)(Y+Z)}, \tag{3.41}$$

em que (3.41), sob a hipótese nula (3.42), ou seja, não associação entre o fator de risco

Figura 3.2 *Amostra de n trios.*



genético e a doença D, segue uma distribuição assintótica Qui-Quadrado com 1 grau de liberdade. Com N sendo o número total de trios.

$$H_0 : P(D|\text{transmitiu}H) = P(ND|\text{transmitiu}H), \tag{3.42}$$

Tabela 3.9 *Transmissão de alelos - GHRR.*

	H	$\overline{H}$	Total
Transmitido	W	X	$W+X$
Não Transmitido	Y	Z	$Y+Z$
	$W+Y$	$X+Z$	$2N$

Seguindo esta abordagem mas, alternativamente, considerando dados pareados (Tabela 3.10), cada trio contribuirá com uma única observação e será classificado em termos de genótipos transmitidos e não transmitidos. Sob esta leitura dos dados, pela Figura 3.2 o *Trio 1* contribuirá com uma única observação na casela B, o *Trio 2* contribuirá com uma observação na casela A e o *Trio n* contribuirá com uma observação na casela C.

Neste caso, a estatística TDT é dada por:

$$\text{TDT} = \frac{(B-C)^2}{(B+C)}, \tag{3.43}$$

Tabela 3.10 *Transmissão de alelos - Amostra pareada GHRR.*

	Não Transmitidos		
Transmitidos	H	$\overline{H}$	Total
H	A	B	W
$\overline{H}$	C	D	X
Total	Y	Z	N

em que (3.43), sob a hipótese de não associação, segue uma distribuição assintótica Qui-Quadrado com 1 grau de liberdade. Além do mais, tem-se que (3.43) é o tradicional teste de McNemar (veja Sham, 1998, por exemplo).

3.3.2 Risco Relativo do Haplótipo no Nível Cromossômico ($HHRR$)

Avaliando agora o risco relativo do haplótipo no nível cromossômico, denotado por $HHRR$, considera-se para a Tabela 3.11 um total amostral de $4N$ (Lange, 1997), ou seja, os alelos transmitidos e não transmitidos de cada um dos pais fornecem quatro observações por família, N sendo considerado como número de trios.

Considerando novamente a Figura 3.2 e a notação disposta na Tabela 3.11, o *Trio 1* contribuirá com duas observações na casela w e duas na casela z, as quatro observações do *Trio 2* serão em cada uma das caselas e o *Trio n* contribuirá com duas observações na casela x e duas na casela y. A hipótese nula e o teste a ser considerado são os referidos anteriormente para a Tabela 3.9.

Tabela 3.11 *Transmissão de alelos - HHRR.*

	H	$\overline{H}$	Total
Transmitido	w	x	$w+x$
Não Transmitido	y	z	$y+z$
	$w+y$	$x+z$	$4N$

Para este caso, mas com dados pareados o TDT considera n_{ij} definido como o número

de vezes em que os pais transmitem o alelo i e não transmitem o alelo j e p_{ij} definido como a probabilidade de pais transmitirem o alelo i e não transmitirem o alelo j (Spielman et al., 1993; Sham, 1998). Neste caso, os dados seguem o formato da Tabela 3.12.

Tabela 3.12 *Transmissão de alelos - Amostra pareada HHRR.*

	Não Transmitidos		
Transmitidos	H	$\overline{H}$	Total
H	n_{11}	n_{12}	$n_{1.}$
$\overline{H}$	n_{21}	n_{22}	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	$2N$

Para a Figura 3.2 o *Trio 1* contribuirá com duas observações na casela n_{12}, o *Trio 2* contribuirá com uma observação na casela n_{21} e outra na casela n_{12} e o *Trio n* contribuirá com duas observações na casela n_{21}. A estatística TDT será dada pelo teste de McNemar, como descrito anteriormente. Sob a hipótese H_0: $p_{12} = p_{21}$, tal teste também é comumente chamado na literatura por "teste de simetria" (Agresti, 2002).

Questiona-se qual destas abordagens, $HHRR$ e $GHRR$, é a mais apropriada para identificar locos genéticos associados à doença. Terwilliger & Ott (1992; 1994) mostraram que $HHRR$ fornece mais poder do que $GHRR$ devido à estrutura de pareamento dos alelos transmitidos e não transmitidos.

3.3.3 Outras Considerações do TDT

O TDT foi estendido para locos com múltiplos alelos (Sham & Curtis, 1995). Quando o loco marcador contém m alelos, $m > 2$ (ver Tabela 3.13), a generalização é dada pelo teste de "simetria":

$$\text{TDT} = \sum_{i<j} \frac{(n_{ij} - n_{ji})^2}{(n_{ij} + n_{ji})}, \tag{3.44}$$

onde n_{ij} é o número de vezes em que os pais transmitem o alelo i e não transmitem o

alelo j e a estatística (3.44), sob a hipótese nula de não associação, tem assintoticamente uma distribuição Qui-Quadrado com $m(m-1)/2$ graus de liberdade.

Tabela 3.13 *Distribuição dos dados de transmissão de alelos.*

	Não Transmitidos				
Transmitidos	Alelo 1	Alelo 2	...	Alelo m	Total
Alelo 1	n_{11}	n_{12}	...	n_{1m}	$n_{1.}$
Alelo 2	n_{21}	n_{22}	...	n_{2m}	$n_{2.}$
			...		
Alelo m	n_{m1}	n_{m2}	...	n_{mm}	$n_{m.}$
Total	$n_{.1}$	$n_{.2}$	...	$n_{.m}$	$n_{..}$

Apesar da vasta aplicação do TDT como um teste de associação genética entre um fator de risco genético e uma doença, existem controvérsias se este é um teste de ligação ou desequilíbrio de ligação entre os locos da doença e marcador sob estudo. Na prática, o principal propósito do TDT é que este é um teste de associação entre um loco marcador e a doença, sendo o efeito "aparente" do marcador no risco da doença decorrente do fato do loco de marcador estar em desequilíbrio de ligação com o loco da doença ou do próprio marcador ser o loco funcional. O que é a situação ideal para o mapeamento de doenças.

De acordo com Sham (1998) e Clayton (2003) a resposta para esta questão está na natureza da amostra. Restrigindo a hipótese nula para dados de famílias que possuem somente um filho afetado, pode ser mostrado que a estatística TDT testa, simultaneamente, associação e ligação, livre do efeito de confundimento, sendo portanto um teste de desequilíbrio de ligação (na presença de associação e ligação) na população da qual a amostra foi extraída. Segue na Seção A.1 do Apêndice A uma descrição das probabilidades para a Tabela 3.12 em função de parâmetros associados a possíveis efeitos de estratificação, fração de recombinação (ligação) e associação (Clayton, 2003). Tal parametrização tem sido usada na comprovação de que o TDT é um teste de desequilíbrio de ligação na presença de associação e ligação.

Outro método chamado "Sib TDT" (ou S-TDT), permite o uso dos princípios do TDT

para dados de pares de irmãos sem de dispor dos dados dos pais (Spielman & Ewens, 1998). Neste caso, os dados consistem em genótipos de marcadores de irmãos, sendo requerido no mínimo um irmão afetado e outro não afetado e estes não devem ter o mesmo genótipo.

3.3.4 Poder e Probabilidade de Falso-positivo

Considerações sobre a probabilidade de conclusões falso-positivas e o poder de testes em estudos de associação em genética têm merecido atenção da literatura (Schlesselman, 1982; Schork, 2002; Duncan et al., 2005). Algumas fontes de falso-positivos são, por exemplo, situações em que o loco de marcador pode estar em desequilíbrio de ligação (associação) com o gene regulador, mas não estar ligado (próximo), isto pode ocorrer devido a estrutura populacional; outro exemplo, é identificar um loco marcador como significativo ao risco da doença, quando na "verdade" o loco é identificador (marcador) de raça e os grupos estão desbalanceados quanto à raça.

Na construção de um teste de hipóteses $H_0 \times H_1$, em estatística duas medidas de erros são avaliadas:

$$\alpha = P(\text{Erro Tipo I}) = P(\text{Rejeitar } H_0 | H_0 \text{ Verdadeiro}) \Rightarrow \text{Falso-positivo}; \quad (3.45)$$

$$\beta = P(\text{Erro Tipo II}) = P(\text{Não Rejeitar } H_0 | H_0 \text{ Falso}) \Rightarrow \text{Falso-negativo}. \quad (3.46)$$

Ao considerar *Poder* de um teste, um conceito a ser discutido é o falso-negativo dado em (3.46), ou seja, $Poder = 1-\beta$. Observa-se que para um nível de significância α fixado, quanto maior o poder "melhor" é o teste. Algumas fontes de falso-negativos também podem ser destacadas, como exemplo, amostragem não representativa (baixa resolução) do genoma em estudo; a não identificação de um loco por "falta de precisão" no ajuste do modelo, onde vale ressaltar que o componente molecular (genes) tem, em geral, efeito esperado pequeno sobre o risco da doença comparado a outros fatores ambientais; ao incluir covariáveis no modelo, estas podem competir ou cancelar o efeito do loco (raça, por exemplo).

O problema de falso-positivo e baixo poder dos estudos de associação têm levado a

uma não convergência na literatura das evidências de associação. Vários esforços têm sido feitos no sentido de encontrar delineamentos ou metodologias de análises de dados "ótimos". Lander & Schork (1994) concluem que populações isoladas (não estruturadas, isto é, homogêneas geneticamente) correspondem ao tipo ideal de universo de amostragem. Contudo, no caso de pesquisas em populações miscigenadas as alternativas de dados familiares, como trios, por exemplo, são recomendadas para evitar falso-positivos. Ainda, optando-se por dados de trios a metodologia de análise via dados pareados é mais poderosa se comparada com amostras independentes (Ewens & Spielman, 1995). Limitações na coleta de dados de trios envolvem a ocorrência de trios não informativos e a dificuldade de coleta de tamanho amostral adequado para se garantir poder na análise.

Morton & Collins (1998), demonstram que os testes TDT com trios completamente informativos são mais poderosos que os testes baseados em estudos caso-controle (indivíduos não relacionados). Contudo, estes últimos são mais fáceis de serem realizados. Schork (2002) mostra que os delineamentos caso-controle podem ter seu poder aumentado quando aplicados apropriadamente, por exemplo, garantindo-se o balanceamento genético entre os grupos e pela seleção de fenótipos (estado da doença) extremos. Além disso, pode-se diminuir a ocorrência de falso-positivos por adotar uma graduação nas interpretações dos níveis descritivos de um loco a ser considerado, tal como, loco "sugestivo", loco "candidato" e loco "significativo" de estar asssociado ao risco da doença. Apesar de pouco viável, a replicação dos estudos de associação também tem sido recomendada como forma de garantir conclusões acertadas.

Barcellos et al. (1997) e Long & Langley (1999) recomendam tamanhos amostrais de 500-1000 para cada grupo em estudos de associação para garantir poder e reduzir a ocorrência de associações falso-positivas. Estratégias como delineamentos multi-estágios (estudo de ligação seguidos de estudo de associação) e avaliações de um grande número de marcadores são também recomendadas.

Schork et al. (2001) evitando usar estudos de simulação ou desenvolvimentos teóricos que envolvem muitos parâmetros genéticos, propõem realizar a avaliação do poder considerando o particular conjunto de dados efetivamente observado. Nesta avaliação *post-hoc* os

autores desenvolveram uma estratégia empírica que preserva a estrutura de associação observada nos dados e não assume qualquer suposição teórica sobre os parâmetros genéticos da população.

Rao & Gun (2001) discutem o problema de falso-positivo e falso-negativo no contexto de estudos de associação que consideram um grande número de marcadores sendo pesquisado. Estes autores salientam o fato de que pouca atenção tem sido dada ao controle de falso-negativos em prol de se controlar falso-positivos. Neste sentido os autores propõem estatísticas que realizam um balanceamento entre os dois tipos de erros e recomendam aumentar o tamanho amostral, possivelmente por meio de estratégias de meta-análise que combinam diferentes estudos. Neste sentido, vale lembrar que as análises de associação no nível cromossômico duplicam o tamanho amostral e podem conduzir a ganho de poder e a proteção de falso-positivos se comparados com as análises no nível genotípico.

Capítulo 4

Aplicação

No presente capítulo tem-se a aplicação das propostas de alguns métodos estatísticos discutidos no Capítulo 3. Para este trabalho, discute-se a análise de um conjunto de dados genéticos apresentados em Clayton (2003) além de dois conjuntos de dados reais que foram disponibilizados pelo Laboratório de Cardiologia e Genética Molecular do InCor/USP, "População de Vitória" e "Trios de São Paulo".

4.1 Dados de Câncer de Mama

Considere os dados apresentados na Tabela 1.1, descritos por Dunning et al. (1997) e discutidos por Clayton (2003), com o objetivo de verificar associação de um polimorfismo "comum" no gene BRCA1 com o câncer de mama. Na Tabela 1.1 (a) tem-se a distribuição do genótipo do polimorfismo Pro871Leu em 800 amostras de *casos* e 572 amostras *controles*, observa-se que esta análise procederá no nível genotípico (indivíduo), enquanto que para a Tabela 1.1 (b) a análise será conduzida no nível cromossômico.

O programa utilizado nestas análises foi o **R** (http://www.R-project.org). Os códigos dos programas utilizados encontram-se na Seção D.1 do Apêndice D.

4.1.1 Análise no Nível Genotípico (indivíduo)

Neste caso, primeiramente calculou-se os *Odds Ratios*, como descrito na Subseção 3.2.1 do Capítulo 3. A medida de risco de câncer de mama foi definida para os níveis genotípicos $LeuLeu$ e $LeuPro$ do polimorfismo Pro871Leu relativamente ao nível genotípico $ProPro$. Na Tabela 4.1 têm-se as estimativas dos *Odds Ratio* obtidas e os correspondentes valores p

baseados no teste de associação via estatística de Wald. Note que $\hat{\Psi}_{LeuLeu|ProPro} = 1{,}2361$ indica que o risco do câncer de mama para indivíduos que carregam o genótipo *LeuLeu* é 24% maior que para o genótipo *ProPro*. Considerando o genótipo heterozigoto *LeuPro* o risco é aproximadamente 15% maior que para *ProPro*. Porém, estes valores de risco aumentado para mulheres que carregam 2 ou 1 cópia do alelo *Leu* em relação àquelas que carregam nenhum (*ProPro*) não são significativos.

Tabela 4.1 *Estimativas dos Odds Ratios no nível genotípico: polimorfismo Pro871Leu.*

Polimorfismo	$\hat{\Psi}$		Valor p
Pro871Leu	$\hat{\Psi}_{LeuLeu\|ProPro}$	1,2361	0,2624
	$\hat{\Psi}_{LeuPro\|ProPro}$	1,1480	0,2330

Segue na Tabela 4.2 as estimativas dos parâmetros e seus respectivos erros padrão considerando o modelo logístico dado por:

$$\ln\left(\frac{p_1(\mathbf{x})}{1-p_1(\mathbf{x})}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2, \tag{4.1}$$

onde $p_1(\mathbf{x})$ representa a probabilidade de uma mulher do grupo genotípico $\mathbf{x}$ ter câncer de mama, e

$$x_1 = \begin{cases} 1, & \text{se o indivíduo carrega o genótipo } LeuLeu \\ 0, & \text{caso contrário} \end{cases}$$

$$x_2 = \begin{cases} 1, & \text{se o indivíduo carrega o genótipo } LeuPro \\ 0, & \text{caso contrário} \end{cases}$$

As estimativas dos parâmetros para o modelo (4.1) são baseadas no processo iterativo do tipo "mínimos quadrados reponderados" (Paula, 2004; McCullagh & Nelder, 1989).

Para o modelo multiplicativo, ou seja, realizando a análise no nível genotípico e considerando a restrição de que o risco relativo do genótipo *LeuLeu* é duas vezes maior (na

Tabela 4.2 *Estimativas dos parâmetros do modelo no nível genotípico: polimorfismo Pro871Leu.*

Parâmetro	Estimativa	Erro padrão	Valor p
α	0,2513	0,0818	0,0021
β_1	0,2120	0,1892	0,2624
β_2	0,1380	0,1157	0,2330

AIC: 25,069

escala logaritmica) que o risco relativo para o genótipo $LeuPro$, como descrito na Subseção 3.2.1 do Capítulo 3, tem-se a estimativa do $\hat{\Psi}_{LeuPro|ProPro}$ =1,1247 ($p = 0{,}1591$). A partir desta estimativa ao quadrado obtém-se $\hat{\Psi}_{LeuLeu|ProPro}$ =1,2650. Segue na Tabela 4.3 as estimativas e os seus respectivos erros padrão, para o modelo multiplicativo:

$$\ln\left(\frac{p_1(x)}{1-p_1(x)}\right) = \alpha + \beta x_2, \tag{4.2}$$

onde

$$x = \begin{cases} 2, & \text{se o indivíduo carrega o genótipo } LeuLeu \\ 1, & \text{se o indivíduo carrega o genótipo } LeuPro \\ 0, & \text{caso contrário} \end{cases}$$

Tabela 4.3 *Estimativas dos parâmetros no nível genotípico sob o modelo multiplicativo: polimorfismo Pro871Leu.*

Parâmetro	Estimativa	Erro padrão	Valor p
α	0,2582	0,0773	0,0008
β_2	0,1175	0,0834	0,1591

AIC: 23,134

Deste modo, também sob o modelo multiplicativo, que inclui um menor número de parâmetros, não é encontrada evidência de que o polimorfismo $Pro871Leu$ esteja associado com o câncer de mama. Vale salientar que o câncer de mama é considerado uma

doença rara e, portanto, não há problema em assumir que os *Odds Ratios* calculados são estimativas de risco da doença.

4.1.2 Análise no Nível Cromossômico

Considerando a Tabela 1.1 (b), ou seja, o estudo no nível cromossômico, calculou-se a estimativa do *Odds Ratio* : $\hat{\Psi}_{Leu|Pro}$ =1,1222 (p= 0,1630). Assim, o risco de câncer de mama para mulheres que carregam uma cópia do alelo *Leu* é 12% maior que para aquelas que carregam 1 cópia do alelo *Pro*. Segue na Tabela 4.4 as estimativas do modelo logístico dado por:

$$\ln\left(\frac{p_1(x)}{1-p_1(x)}\right) = \alpha + \beta x_2, \tag{4.3}$$

onde $p_1(x)$ representa a probabilidade de uma mulher do grupo alélico x ter câncer de mama, e

$$x = \begin{cases} 1, & \text{se o indivíduo carrega o alelo } Leu \\ 0, & \text{caso contrário} \end{cases}$$

Tabela 4.4 *Estimativas dos parâmetros no nível cromossômico: polimorfismo Pro871Leu.*

Parâmetro	Estimativa	Erro padrão	Valor p
α	0,2975	0,0472	$<$ 0,0001
β_2	0,1153	0,0826	0,1630

AIC: 19,167

Tem-se na Tabela 4.5 uma comparação das estimativas dos *Odds Ratios* vistos para este exemplo. Observe que as estimativas $\hat{\Psi}_{LeuLeu|ProPro}$ e $\hat{\Psi}_{LeuPro|ProPro}$ para os três modelos são próximas e que em todos os casos a mesma conclusão foi obtida. comparando-se os modelos pelo critério de Akaike, 1974 (AIC) a análise cromossômica pode ser considerada mais parcimoniosa.

Tabela 4.5 *Estimativas dos Odds Ratios (erro padrão) para cada modelo.*

Modelo	gl	$\hat{\Psi}_{LeuLeu\|ProPro}$	$\hat{\Psi}_{LeuPro\|ProPro}$
Nível genotípico	2	1,2361(0,2339)	1,1480(0,1328)
Nível genotípico (multiplicativo)	1	1,2650(0,2110)	1,1247(0,0938)
Nível cromossômico (EHW)	1	1,2593(0,2080)	1,1222(0,0927)

4.2 População de Vitória

Esse conjunto de dados consiste em um grupo de 1577 indivíduos não relacionados amostrados da Cidade de Vitória - ES, de forma sistemática por unidades domiciliares, respeitando-se a estratificação pelo nível sócio econômico. Estes indivíduos participaram de um estudo epidemiológico sobre fatores de risco cardiovasculares. Além de dados genotípicos para alguns marcadores genéticos, fazem parte deste conjunto de dados caracterizações fenotípicas de pressão arterial, peso, altura, etnia, colesterol total, LDL-colesterol, HDL-colesterol, triglicérides e glicemia de jejum. Para a análise deste conjunto de dados foram considerados 09 marcadores moleculares (Tabela 4.6): B1BK, NHPS2.5, BAR-2(16) Cat, BAR-2(27) Cat, ECA Cat, ANGO Cat, ENOS Cat, p22 PHOX, GNB3 e as variáveis diabetes, hipertensão e obesidade, anotadas de forma categorizada, isto é, como grupos de *caso* e *controle*, por exemplo, indivíduos hipertensos (grupo *caso*) e normotensos (grupo *controle*). O interesse é verificar se o risco da doença está associado a fatores de risco genético (marcadores moleculares). Com isto considerou-se a análise por genótipos (tamanho amostral n), ou seja, o risco da doença depende do genótipo, e também a análise por haplótipos (tamanho amostral 2n), ou seja, no nível cromossômico onde tem-se que os dois alelos de cada indivíduo são amostrados independentemente da amostra de uma população de alelos, isto é, o equilíbrio de Hardy-Weinberg é assumido.

Para a análise deste banco de dados algumas ressalvas se fazem necessárias. Primeiramente considerando o plano amostral de coleta dos dados, ve-se que trata-se de um estudo transversal avaliando-se simultaneamente vários fenótipos relacionados a problemas cardiovasculares, anotados em sua maioria de forma quantitativa. Logo, o estudo apesar de

Tabela 4.6 *Marcadores da "População de Vitória".*

Marcador	Níveis Genotípicos	Níveis Alélicos
B1BK	AA, AG e GG	A e G
NHPS2.5	AA, GA e GG	A e G
BAR-2(16) Cat	AA, AG e GG	A e G
ECA Cat	DD, DI e II	D e I
ANGO Cat	TT, MT e MM	T e M
ENOS Cat	TT, GT e GG	T e G
p22 PHOX	TT, TC e CC	T e C
GNB3	TT, CT e CC	T e C
BAR-2(27) Cat	$GlnGln$, $GlnGlu$ e $GluGlu$	Gln e Glu

não ser do tipo caso-controle será analisado como tal, por categorização dos fenótipos de interesse e condicionamento dos totais observados de cada grupo. Ainda, trata-se da análise de doenças comuns (obesidade, hipertensão e diabetes) que não são de ocorrência rara na população e, portanto, os *Odds Ratios* calculados não estimam o risco relativo da doença em termos de probabilidades de incidência. Finalmente, vale ressaltar que para estes dados o controle da homogeneidade genética entre os grupos foi feita por meio da classificação dos indivíduos por raça, categorizada por branco, negro, mulato e outros.

O programa utilizado nesta análise foi o **R** (http://www.R-project.org). Os códigos dos programas utilizados encontram-se na Seção D.1 do Apêndice D.

4.2.1 Análise no Nível Genotípico (indivíduo)

Para cada marcador, foram feitas tabelas de contingência avaliando os grupos de *caso* e *controle*, para diabetes, hipertensão e obesidade (ver Tabela B.1 na Seção B.1 do Apêndice B). Observa-se que o número total da amostra $n = 1577$ varia, pois existem dados faltantes, incompletos, tanto para a variável fenotípica como para a variável preditora genotípica.

Para o cálculo dos *Odds Ratios*, como descrito na Subseção 3.2.1 do Capítulo 3, os seguintes níveis genotípicos foram definidos para cada marcador: B1BK (AA e AG),

NHPS2.5 (AA e GA), BAR-2(16) Cat (AA e AG), BAR-2(27) Cat ($GlnGln$ e $GlnGlu$), ECA Cat (DD e DI), ANGO Cat (TT e MT), ENOS Cat (TT e GT), p22 PHOX (TT e TC), GNB3 (TT e CT).

Tem-se na Tabela 4.7 as estimativas dos *Odds Ratios*[1] obtidas e os correspondentes valores p, baseados no teste de Wald, para o estudo da associação. Adotando-se um nível de significância de 5%, observa-se que, para hipertensão, há evidência de associação significativa para os marcadores: ANGO Cat no genótipo TT, B1BK em relação ao genótipo AG e para o marcador p22PHOX nos genótipos TT e TC. Para o marcador BAR-2(16) Cat em relação ao genótipo AG nota-se que há evidência de associação com a hipertensão e obesidade. Estes resultados indicam que o modelo de regulação genético da hipertensão envolve múltiplos locos (considerando apenas 9 marcadores sob análise) e que existe um mesmo loco genético controlando, simultâneamente, a hipertensão e obesidade, o que pode explicar uma possível covariância entre estes fenótipos devido a fatores genéticos.

Segue nas Tabelas 4.8, 4.9 e 4.10 as estimativas dos parâmetros[2] e seus respectivos erros padrão considerando o modelo logístico dado por:

$$\ln\left(\frac{p_1(\mathbf{x})}{1-p_1(\mathbf{x})}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2. \tag{4.4}$$

onde $p_1(\mathbf{x})$ representa a probabilidade de um indivíduo do grupo genotípico $\mathbf{x}$ pertencer ao grupo *caso*. Para o marcador B1BK os genótipos estão definidos como:

$$x_1 = \begin{cases} 1, & \text{se o indivíduo carrega o genótipo } AA \\ 0, & \text{caso contrário} \end{cases}$$

$$x_2 = \begin{cases} 1, & \text{se o indivíduo carrega o genótipo } AG \\ 0, & \text{caso contrário} \end{cases}$$

Para o marcador ANGO Cat:

$$x_1 = \begin{cases} 1, & \text{se o indivíduo carrega o genótipo } TT \\ 0, & \text{caso contrário} \end{cases}$$

[1] Para o marcador NHPS2.5 não foi possível calcular todas as estimativas *Odds Ratios* devido à casela $n_{AA}^{(1)}$ da Tabela B.1 ser igual a zero.

[2] Baseados no processo iterativo "mínimos quadrados reponderados".

Tabela 4.7 *Estimativas dos Odds Ratios no nível genotípico.*

Marcador	$\hat{\Psi}$	Diabetes	valor p	Hipertensão	valor p	Obesidade	valor p
B1BK	$\hat{\Psi}_{AA/GG}$	0,7388	0,2471	0,7534	0,0765	0,9881	0,9401
	$\hat{\Psi}_{AG/GG}$	0,6067	0,0564	0,6770	**0,0130**	0,9285	0,6365
NHPS2.5	$\hat{\Psi}_{AA/GG}$	-	-	-	-	-	-
	$\hat{\Psi}_{GA/GG}$	0,5718	0,2840	1,1895	0,4520	0,8961	0,6380
BAR-2(16) Cat	$\hat{\Psi}_{AA/GG}$	1,0868	0,7660	0,7509	0,0787	1,3081	0,0908
	$\hat{\Psi}_{AG/GG}$	0,9672	0,8870	0,7348	**0,0198**	1,3341	**0,0277**
BAR-2(27) Cat	$\hat{\Psi}_{GlnGln/GluGlu}$	1,5820	0,2620	1,2212	0,3300	1,4725	0,0556
	$\hat{\Psi}_{GlnGlu/GluGlu}$	1,4405	0,3810	1,2061	0,3700	1,1496	0,5003
ECA Cat	$\hat{\Psi}_{DD/II}$	1,1181	0,7030	1,0835	0,6310	1,0248	0,8795
	$\hat{\Psi}_{DI/II}$	1,0293	0,9150	1,1026	0,5230	0,9408	0,6796
ANGO Cat	$\hat{\Psi}_{TT/MM}$	1,1884	0,5270	1,4839	**0,0108**	1,2959	0,0820
	$\hat{\Psi}_{MT/MM}$	1,1006	0,7190	1,2776	0,1045	1,0660	0,6590
ENOS Cat	$\hat{\Psi}_{TT/GG}$	1,1063	0,7960	1,3912	0,1310	0,7768	0,2710
	$\hat{\Psi}_{GT/GG}$	1,1280	0,5350	1,1499	0,2100	0,9961	0,9710
p22 PHOX	$\hat{\Psi}_{TT/CC}$	1,0787	0,8160	0,7879	**0,0469**	1,0700	0,6990
	$\hat{\Psi}_{TC/CC}$	1,2750	0,2260	0,7532	**0,0104**	0,9774	0,8370
GNB3	$\hat{\Psi}_{TT/CC}$	0,9130	0,7270	1,2586	0,1250	0,8398	0,2410
	$\hat{\Psi}_{CT/CC}$	0,8333	0,3940	1,0849	0,5120	0,9604	0,7360

$$x_2 = \begin{cases} 1, & \text{se o indivíduo carrega o genótipo } MT \\ 0, & \text{caso contrário} \end{cases}$$

E assim por diante para os demais marcadores.

Tabela 4.8 *Estimativas dos parâmetros (erro padrão) do modelo no nível genotípico para Diabetes.*

	Diabetes		
Marcador	Intercepto	$\hat{\beta}_1$	$\hat{\beta}_2$
B1BK	-2,1203(0,2160)	-0,3028(0,2616)	-0,4997(0,2618)
BAR-2(16) Cat	-2,4773(0,1812)	0,0832(0,2799)	-0,0333(0,2353)
BAR-2(27) Cat	-2,8499(0,3887)	0,4587(0,4089)	0,3650(0,4168)
ECA Cat	-2,5023(0,2386)	0,1116(0,2927)	0,0288(0,2709)
ANGO Cat	-2,5686(0,2264)	0,1727(0,2729)	0,0959(0,2664)
ENOS Cat	-2,4989(0,1290)	0,1010(0,3912)	0,1204(0,1941)
p22 PHOX	-2,5860(0,1513)	0,0758(0,3257)	0,2430(0,2007)
GNB3	-2,3514(0,1615)	-0,0910(0,2605)	-0,1823(0,2138)

Considerando o modelo multiplicativo, ou seja, a análise no nível genotípico, mas im-

Tabela 4.9 *Estimativas dos parâmetros (erro padrão) do modelo no nível genotípico para Hipertensão.*

	Hipertensão		
Marcador	Intercepto	$\hat{\beta}_1$	$\hat{\beta}_2$
B1BK	-0,3761(0,1354)	-0,2832(0,1599)	**-0,3901(0,1571)**
BAR-2(16) Cat	-0,5459(0,1003)	-0,2865(0,1630)	**-0,3081(0,1322)**
BAR-2(27) Cat	-0,8473(0,1914)	0,1998(0,2051)	0,1874(0,2091)
ECA Cat	-0,7472(0,1349)	0,0802(0,1672)	0,0977(0,1529)
ANGO Cat	-0,9233(0,1282)	**0,3946(0,1548)**	0,2450(0,1509)
ENOS Cat	-0,7442(0,0730)	0,3302(0,2189)	0,1397(0,1114)
p22 PHOX	-0,3624(0,0783)	**-0,3567(0,1795)**	**-0,2834(0,1107)**
GNB3	-0,7587(0,0973)	0,2300(0,1500)	0,0815(0,1242)

Tabela 4.10 *Estimativas dos parâmetros (erro padrão) do modelo no nível genotípico para Obesidade.*

	Obesidade		
Marcador	Intercepto	$\hat{\beta}_1$	$\hat{\beta}_2$
B1BK	-0,4686(0,1367)	-0,0120(0,1599)	-0,0742(0,1570)
BAR-2(16) Cat	-0,7002(0,1030)	0,2685(0,1588)	**0,2882(0,1309)**
BAR-2(27) Cat	-0,7638(0,1891)	0,3870(0,2021)	0,1394(0,2069)
ECA Cat	-0,4687(0,1295)	0,0245(0,1614)	-0,0610(0,1477)
ANGO Cat	-0,6280(0,1220)	0,2592(0,1490)	0,0639(0,1450)
ENOS Cat	-0,4879(0,0704)	-0,2525(0,2293)	-0,0039(0,1089)
p22 PHOX	-0,5021(0,0797)	0,0676(0,1749)	-0,0228(0,1108)
GNB3	-0,4505(0,0933)	-0,1746(0,1489)	-0,0404(0,1199)

pondo a restrição de efeito linear do número de alelos no genótipo, o modelo fica com apenas um parâmetro e mais fácil de interpretar. Tem-se na Tabela 4.11 a estimativa do *Odds Ratio* e seu respectivo valor p, considerando o modelo multiplicativo, observa-se que, para os marcadores BAR-2(16) Cat, ANGO Cat e p22PHOX há evidência de associação significativa em relação ao efeito linear de seus genótipos com hipertensão e para o marcador BAR-2(27) Cat com a obesidade.

Segue na Tabela 4.12 as estimativas dos parâmetros com seus respectivos erros padrão para o modelo logístico linear multiplicativo dado por:

$$\ln\left(\frac{p_1(x)}{1-p_1(x)}\right) = \alpha + \beta_2 x, \tag{4.5}$$

Tabela 4.11 *Estimativas do Odds Ratio no nível de genótipo sob o modelo multiplicativo.*

Marcador	$\hat{\Psi}$	Diabetes	valor p	Hipertensão	valor p	Obesidade	valor p
B1BK	$\hat{\Psi}_{AG/GG}$	0,9162	0,5170	0,9176	0,2650	1,0104	0,8900
NHPS2.5	$\hat{\Psi}_{GA/GG}$	0,5636	0,2680	1,1413	0,5590	0,8667	0,5320
BAR-2(16) Cat	$\hat{\Psi}_{AG/GG}$	1,0365	0,8000	0,8497	**0,0456**	1,1574	0,0615
BAR-2(27) Cat	$\hat{\Psi}_{GlnGlu/GluGlu}$	1,1750	0,2880	1,0636	0,4600	1,2432	**0,0086**
ECA Cat	$\hat{\Psi}_{DI/II}$	1,0628	0,6700	1,0307	0,7090	1,0245	0,7600
ANGO Cat	$\hat{\Psi}_{MT/MM}$	1,0883	0,5200	1,2078	**0,0112**	1,1507	0,0536
ENOS Cat	$\hat{\Psi}_{GT/GG}$	1,0888	0,5720	1,1649	0,0768	0,9398	0,4690
p22 PHOX	$\hat{\Psi}_{TC/CC}$	1,1087	0,4620	0,8058	**0,0069**	1,0152	0,8480
GNB3	$\hat{\Psi}_{CT/CC}$	0,9408	0,6400	1,1190	0,1320	0,9210	0,2620

onde x, por exemplo, para o marcador B1BK é:

$$x = \begin{cases} 2, & \text{se o indivíduo carrega o genótipo } AA \\ 1, & \text{se o indivíduo carrega o genótipo } AG \\ 0, & \text{caso contrário} \end{cases}$$

Tabela 4.12 *Estimativas dos parâmetros (erro padrão) no nível genotípico sob o modelo multiplicativo.*

	Diabetes		Hipertensão		Obesidade	
Marcador	Intercepto	$\hat{\beta}_2$	Intercepto	$\hat{\beta}_2$	Intercepto	$\hat{\beta}_2$
B1BK	-2,3502(0,1892)	-0,0875(0,1352)	-0,5588(0,1092)	-0,0859(0,0772)	-0,5204(0,1081)	0,0104(0,0757)
NHPS2.5	-2,4366(0,0956)	-0,5734(0,5173)	-0,6769(0,0548)	0,1322(0,2262)	-0,4955(0,0536)	-0,1431(0,2292)
BAR-2(16) Cat	-2,5069(0,1641)	0,0358(0,1411)	-0,6054(0,0915)	**-0,1629(0,0815)**	-0,6371(0,0911)	0,1462(0,0782)
BAR-2(27) Cat	-2,6963(0,2450)	0,1612(0,1517)	-0,7572(0,1318)	0,0617(0,0853)	-0,8201(0,1318)	**0,2177(0,0829)**
ECA Cat	-2,5227(0,1894)	0,0609(0,1427)	-0,7045(0,1063)	0,0303(0,0810)	-0,5221(0,1037)	0,0242(0,0791)
ANGO Cat	-2,5622(0,1845)	0,0846(0,1314)	-0,8917(0,1043)	**0,1888(0,0745)**	-0,6701(0,1011)	0,1404(0,0727)
ENOS Cat	-2,4890(0,1238)	0,0851(0,1504)	-0,7477(0,0704)	0,1527(0,0863)	-0,4726(0,0680)	-0,0621(0,0858)
p22 PHOX	-2,5325(0,1032)	0,1032(0,1402)	-0,3851(0,0741)	**-0,2160(0,0799)**	-0,5153(0,0750)	0,0151(0,0788)
GNB3	-2,4013(0,1487)	-0,0612(0,1321)	-0,7724(0,0870)	0,1124(0,0747)	-0,4324(0,0837)	-0,0824(0,0734)

Comparando o modelo multiplicativo em relação ao modelo saturado no nível genotípico, observa-se que os marcadores B1BK com hipertensão e BAR-2(16) Cat com obesidade não mostraram evidência de associação quando considera-se o modelo multiplicativo, indicando que o padrão de variação linear não se aplica a estes marcadores, possivelmente devido a um efeito de interação entre os alelos A e G. Enquanto que os marcadores BAR-2(16) Cat, ANGO Cat e p22 PHOX para hipertensão continuam significativos, indicando que um modelo mais simplificado é adequado. Considerando o método de Akaike

para comparar os dois modelos para os três marcadores com hipertensão, observa-se que o modelo multiplicativo é mais parcimonioso (Tabela 4.13).

Tabela 4.13 *Comparação dos modelos (AIC).*

Marcador	AIC	
Hipertensão	Genotípico	Genotípico (multiplicativo)
BAR-2(16) Cat	25,125	25,059
ANGO Cat	25,576	23,761
p22 PHOX	25,356	24,133

Fazem parte também deste conjunto de dados caracterizações fenotípicas como a etnia (ver Tabela B.3 na Seção B.1 do Apêndice B). Deste modo, se fez análises estratificando por etnia, para isto, primeiramente calculou-se a estatística de Cochran-Mantel-Haenszel (Tabela 4.14). Neste caso, as medidas de risco são combinadas por genótipo e por estrato. O único marcador com efeito significativo foi BAR-2(16) Cat para hipertensão, mantendo, para este caso, os resultados anteriores. Cuidados devem ser tomados nesta análise sobre a possibilidade de associação conflitante.

Tabela 4.14 *Teste de Cochran-Mantel-Haenszel controlando por etnia no nível genotípico.*

Marcador	Diabetes	valor p	Hipertensão	valor p	Obesidade	valor p
B1BK	3,8288	0,1474	5,4429	0,0658	0,8309	0,6600
BAR-2(16) Cat	0,1861	0,9112	6,2261	**0,0445**	5,2020	0,0742
BAR-2(27) Cat	1,1844	0,5531	0,9863	0,6104	4,6932	0,0957
ECA Cat	0,2293	0,8917	1,3028	0,5213	0,6036	0,7395
ANGO Cat	0,3892	0,8232	3,1852	0,2034	2,3673	0,3062
ENOS Cat	0,1909	0,9090	4,6539	0,0976	1,1297	0,5684
p22 PHOX	1,4523	0,4838	0,0557	0,9725	0,6271	0,7308
GNB3	0,8563	0,6517	0,2469	0,8839	4,1607	0,1249

Como visto na seção 3.2.3 alguns modelos logísticos podem ser considerados ao incluir no estudo possíveis variáveis de confundimento. Para uma análise mais detalhada considere a estratificação por etnia e o modelo Modelo 3 expressão (3.37), o qual foi avaliado para cada

marcador e Diabetes, Hipertensão e Obesidade. A etnia foi incluída no modelo adotando a categoria branco com referência.

Nas Tabelas 4.15, 4.16, 4.17, 4.18, 4.19, 4.20, 4.21, 4.22 estão as estimativas do modelo para cada marcador e para os três fenótipos. O único caso de efeito significativo de interação entre o risco relativo da doença e etnia foi para o marcador BAR-2(27) Cat e obesidade. Neste caso, as medidas de risco significativas foram

$$\begin{aligned} \hat{\Psi}^{B}_{GlnGlu|GluGlu} &= 3,7408 \Rightarrow \hat{\Psi}^{B}_{GluGlu|GlnGlu} = 0,2673, \\ \hat{\Psi}^{O}_{GlnGlu|GluGlu} &= 0,5935, \end{aligned} \tag{4.6}$$

para as etnias "brancos" e "outros", respectivamente. Assim, um modo de associação conflitante é identificado em que o risco de obesidade para o genótipo *GluGlu* é aproximadamente 27% menor que para o genótipo *GlnGlu* considerando a etnia "branco" e o risco de obesidade para o genótipo *GlnGlu* é 59% menor que para o genótipo *GluGlu* no caso de "outro". Para os casos em que não há efeito significativo de interação, os resultados das análises sem a inclusão da variável etnia foram considerados. Exceto para o marcador $B1BK$, para as demais situações houve efeito significativo de raça, indicando que o padrão de ocorrência da doença depende da raça. Tal resultado é esperado em se tratando de população com alta taxa de miscigenação como a brasileira. Dado que raça é um fator de heterogeneidade genética importante, a validação dos marcadores identificados como fatores de risco nesta população merece alguns cuidados, já que trata-se de populações com algum grau de *background* genético diferente.

Tabela 4.15 *Estimativas do modelo para o marcador B1BK no nível genotípico estratificando por etnia.*

	Diabetes		Hipertensão		Obesidade	
Parâmetro	Estimativa(EP)	valor p	Estimativa(EP)	valor p	Estimativa(EP)	valor p
α	-2,4363(0,3554)	**<0,0001**	-0,7584(0,2054)	**0,0002**	-0,6665(0,2010)	**0,0009**
β_2	-0,0517(0,2397)	0,8290	-0,0809(0,1385)	0,5593	-0,0283(0,1348)	0,8338
γ_2	0,2617(0,6414)	0,6830	0,4566(0,3684)	0,2152	0,3454(0,3664)	0,3459
γ_3	0,2424(0,4352)	0,5780	0,2552(0,2555)	0,3179	0,1020(0,2519)	0,6855
γ_4	-1,6127(1,3748)	0,2410	-0,4396(0,5521)	0,4258	0,3076(0,4867)	0,5274
δ_2	-0,4768(0,5576)	0,3920	0,3127(0,2778)	0,2604	0,1496(0,2759)	0,5877
δ_3	-0,0837(0,3031)	0,7820	0,0135(0,1765)	0,9388	0,1277(0,1728)	0,4598
δ_4	1,0388(0,7999)	0,1940	0,1816(0,3708)	0,6243	0,0891(0,3295)	0,7868

Tabela 4.16 *Estimativas do modelo para o marcador BAR-2(16) Cat no nível genotípico estratificando por etnia.*

	Diabetes		Hipertensão		Obesidade	
Parâmetro	Estimativa(EP)	valor p	Estimativa(EP)	valor p	Estimativa(EP)	valor p
α	-2,4874(0,2686)	**<0,0001**	-0,9317(0,1582)	**<0,0001**	-0,8746(0,1547)	**<0,0001**
β_2	-0,0077(0,2470)	0,9750	0,0012(0,1453)	0,9934	0,2365(0,1385)	0,0877
γ_2	-1,2056(0,9250)	0,1920	0,5938(0,3492)	0,0891	0,4436(0,3496)	0,2045
γ_3	0,1144(0,3523)	0,7450	0,5686(0,2054)	**0,0056**	0,3499(0,2029)	0,0846
γ_4	0,3584(0,6842)	0,6000	-0,0326(0,4395)	0,9410	0,6359(0,3960)	0,1083
δ_2	0,8521(0,6439)	0,1860	0,1320(0,3022)	0,6623	0,0446(0,3002)	0,8818
δ_3	0,0210(0,3124)	0,9470	-0,3311(0,1854)	0,0741	-0,1702(0,1770)	0,3360
δ_4	-0,7811(0,7234)	0,2800	-0,2089(0,3742)	0,5767	-0,2861(0,3258)	0,3798

Tabela 4.17 *Estimativas do modelo para o marcador BAR-2(27) Cat no nível genotípico estratificando por etnia.*

	Diabetes		Hipertensão		Obesidade	
Parâmetro	Estimativa(EP)	valor p	Estimativa(EP)	valor p	Estimativa(EP)	valor p
α	-3,1123(0,4219)	**<0,0001**	-0,9115(0,2081)	**<0,0001**	-1,0767(0,2102)	**<0,0001**
β_2	0,4335(0,2637)	0,1000	0,0347(0,1405)	0,8047	0,2785(0,1392)	**0,0455**
γ_2	0,8393(1,0280)	0,4140	1,2600(0,5591)	**0,0242**	0,6499(0,5637)	0,2489
γ_3	0,7876(0,5339)	0,1400	0,2934(0,2841)	0,3019	0,3752(0,2855)	0,1889
γ_4	-0,4628(1,5182)	0,7600	-0,4510(0,7544)	0,5500	1,5736(0,6541)	**0,0161**
δ_2	-0,6758(0,6304)	0,2840	-0,2952(0,3362)	0,3800	-0,1345(0,3372)	0,6900
δ_3	-0,4570(0,3327)	0,1700	-0,0165(0,1844)	0,9289	-0,1030(0,1829)	0,5735
δ_4	0,1894(0,8826)	0,8300	0,1597(0,4670)	0,7324	-0,8003(0,4144)	**0,0535**

Tabela 4.18 *Estimativas do modelo para o marcador ECA Cat no nível genotípico estratificando por etnia.*

	Diabetes		Hipertensão		Obesidade	
Parâmetro	Estimativa(EP)	valor p	Estimativa(EP)	valor p	Estimativa(EP)	valor p
α	-2,4955(0,3491)	**<0,0001**	-0,9644(0,2037)	**<0,0001**	-0,8154(0,1974)	**<0,0001**
β_2	-0,0063(0,2585)	0,9806	0,0770(0,1496)	0,6069	0,0948(0,1449)	0,5129
γ_2	0,7619(0,6129)	0,2138	1,3678(0,3859)	**0,0004**	1,0529(0,3801)	**0,0056**
γ_3	-0,1160(0,4352)	0,7898	0,2495(0,2498)	0,3179	0,2898(0,2426)	0,2323
γ_4	-0,3953(0,9490)	0,6770	-0,8070(0,5739)	0,1597	0,3917(0,4683)	0,4029
δ_2	-1,2068(0,6853)	0,0782	-0,5564(0,3230)	0,0850	-0,5075(0,3198)	0,1125
δ_3	0,2329(0,3204)	0,4674	0,0310(0,1856)	0,8672	-0,0083(0,1804)	0,9632
δ_4	0,2648(0,7161)	0,7115	0,5841(0,4268)	0,1711	0,0631(0,3688)	0,8642

Tabela 4.19 *Estimativas do modelo para o marcador ANGO Cat no nível genotípico estratificando por etnia.*

	Diabetes		Hipertensão		Obesidade	
Parâmetro	Estimativa(EP)	valor p	Estimativa(EP)	valor p	Estimativa(EP)	valor p
α	-2,4632(0,1871)	**<0,0001**	-0,8521(0,1095)	**<0,0001**	-0,7929(0,1082)	**<0,0001**
β_2	-0,0808(0,1886)	0,6690	-0,0251(0,1057)	0,8125	0,1571(0,1002)	0,1169
γ_2	0,4483(0,4732)	0,3430	1,0488(0,3020)	**0,0005**	0,3546(0,3066)	0,2474
γ_3	-0,0285(0,2535)	0,9110	0,1589(0,1458)	0,2755	0,3581(0,1426)	**0,0120**
γ_4	-0,7949(0,7445)	0,2860	-0,5112(0,3552)	0,1501	0,2622(0,3019)	0,3850
δ_2	-0,6750(0,4599)	0,1420	-0,2007(0,2128)	0,3455	0,0480(0,2114)	0,8204
δ_3	0,2370(0,2277)	0,2980	0,1513(0,1299)	0,2441	-0,1642(0,1249)	0,1887
δ_4	0,7193(0,4863)	0,1390	0,4269(0,2692)	0,1127	0,1994(0,2459)	0,4173

Tabela 4.20 *Estimativas do modelo para o marcador ENOS Cat no nível genotípico estratificando por etnia.*

	Diabetes		Hipertensão		Obesidade	
Parâmetro	Estimativa(EP)	valor p	Estimativa(EP)	valor p	Estimativa(EP)	valor p
α	-2,4849(0,1689)	**<0,0001**	-0,8897(0,0986)	**<0,0001**	-0,6962(0,0950)	**<0,0001**
β_2	-0,1226(0,3097)	0,6923	0,1286(0,1580)	0,4155	-0,0462(0,1628)	0,7768
γ_2	-0,2877(0,4246)	0,4981	0,7720(0,2085)	**0,0002**	0,5446(0,2070)	**0,0085**
γ_3	0,1100(0,2135)	0,6064	0,2884(0,1245)	**0,0206**	0,2664(0,1207)	**0,0273**
γ_4	-0,0408(0,4566)	0,9288	-0,2254(0,2759)	0,4139	0,4730(0,2430)	0,0515
δ_2	1,5089(0,7961)	0,0581	6,7133(154,56)	0,9654	-6,6611(154,56)	0,9656
δ_3	0,2549(0,4126)	0,5367	-0,0416(0,2322)	0,8577	-0,0659(0,2390)	0,7827
δ_4	-6,3976(297,08)	0,9828	0,4290(0,4563)	0,3472	-0,1889(0,4759)	0,6915

Tabela 4.21 *Estimativas do modelo para o marcador p22PHOX no nível genotípico estratificando por etnia.*

	Diabetes		Hipertensão		Obesidade	
Parâmetro	Estimativa(EP)	valor p	Estimativa(EP)	valor p	Estimativa(EP)	valor p
α	-2,4911(0,2279)	**<0,0001**	-0,8288(0,1313)	**<0,0001**	-0,8207(0,1302)	**<0,0001**
β_2	-0,0181(0,2425)	0,9410	-0,0651(0,1402)	0,6421	0,1635(0,1344)	0,2240
γ_2	-0,3610(0,6541)	0,5810	0,6137(0,3144)	0,0510	0,5928(0,3141)	0,0592
γ_3	0,0516(0,2944)	0,8610	0,1768(0,1694)	0,2966	0,4164(0,1667)	**0,0125**
γ_4	-0,7222(0,7950)	0,3640	0,3386(0,3659)	0,3547	0,8157(0,3474)	**0,0189**
δ_2	0,2774(0,6165)	0,6530	0,2803(0,3183)	0,3786	-0,1042(0,3155)	0,7411
δ_3	0,1307(0,3084)	0,6720	0,1521(0,1798)	0,3977	-0,2168(0,1744)	0,2137
δ_4	0,7419(0,6917)	0,2830	-0,7792(0,4444)	0,0796	-0,5182(0,3688)	0,1600

Tabela 4.22 *Estimativas do modelo para o marcador GNB3 Cat no nível genotípico estratificando por etnia.*

	Diabetes		Hipertensão		Obesidade	
Parâmetro	Estimativa(EP)	valor p	Estimativa(EP)	valor p	Estimativa(EP)	valor p
α	-2,5650(0,1807)	**<0,0001**	-0,9073(0,1023)	**<0,0001**	-0,6964(0,0983)	**<0,0001**
β_2	0,1979(0,2073)	0,3400	0,1071(0,1305)	0,4121	-0,0275(0,1309)	0,8335
γ_2	-0,2973(0,5449)	0,5850	0,8532(0,2541)	**0,0008**	0,8046(0,2527)	**0,0015**
γ_3	0,2495(0,2302)	0,2780	0,3288(0,1329)	**0,0134**	0,3171(0,1286)	**0,0137**
γ_4	0,1226(0,4624)	0,7910	-0,3199(0,2942)	0,2768	0,3462(0,2543)	0,1734
δ_2	0,0457(0,4212)	0,9140	-0,1217(0,2268)	0,5914	-0,3732(0,2327)	0,1087
δ_3	-0,2856(0,2596)	0,2710	-0,1285(0,1573)	0,4140	-0,1054(0,1575)	0,5034
δ_4	-6,7598(219,41)	0,9750	0,4154(0,3574)	0,2451	0,2938(0,3501)	0,4013

Fazendo um paralelo do modelo multiplicativo da análise no nível genotípico (equação 4.5) com o modelo incluindo a variável etnia (sem efeito de interação), ver Tabela 4.23, nota-se que o efeito do genótipo para os marcadores BAR(2)-16 Cat com hipertensão e BAR(2)-27 Cat com obesidade continuaram significativo, enquanto que para os marcadores ANGO Cat e p22 PHOX com hipertensão ao incluir a variável etnia o efeito do genótipo deixou de ser significativo. Este resultado deve ser interpretado com cuidado, pois pode-se argumentar que o fator de risco pode não estar relacionado com o genótipo do marcador, e que o sinal significativo na análise marginal (sem o ajuste por raça) é um falso-positivo. Contudo, considerando que o fator raça envolve um efeito sobre risco da doença que pode levar em conta influências culturais mais também a influência do

background genético, tem-se que a inclusão de raça no modelo pode capturar o efeito do verdadeiro loco (o marcador sob análise) regulador da doença e, neste caso, o ajuste por raça pode conduzir a falso-negativo.

Tabela 4.23 *Estimativas dos parâmetros do modelo multiplicativo genotípico sem e com a variável etnia (sem interação).*

	Marcador		$\hat{\alpha}$	$\hat{\beta}_2$	$\hat{\gamma}_1$	$\hat{\gamma}_2$	$\hat{\gamma}_3$
Diabetes	B1BK	sem	**-2,350**	-0,088			
		com	**-2,389**	-0,088	-0,171	0,139	-0,091
	BAR(2)-16 Cat	sem	**-2,507**	0,036			
		com	**-2,510**	0,019	-0,269	0,131	-0,269
	BAR(2)-27 Cat	sem	**-2,696**	0,161			
		com	**-2,705**	0,148	-0,181	0,121	-0,111
	ECA Cat	sem	**-2,523**	0,061			
		com	**-2,585**	0,068	-0,101	0,159	-0,091
	ANGO Cat	sem	**-2,562**	0,085			
		com	**-2,534**	0,051	-0,175	0,132	-0,109
	ENOS Cat	sem	**-2,489**	0,085			
		com	**-2,513**	0,043	-0,136	0,148	-0,096
	p22 PHOX	sem	**-2,533**	0,103			
		com	**-2,576**	0,106	-0,147	0,141	-0,104
	GNB3	sem	**-2,401**	-0,061			
		com	**-2,502**	0,009	-0,153	0,143	-0,091
Hipertensão	B1BK	sem	**-0,559**	-0,086			
		com	**-0,825**	-0,030	**0,790**	**0,277**	-0,204
	BAR(2)-16 Cat	sem	**-0,605**	**-0,163**			
		com	**-0,790**	**-0,171**	**0,723**	**0,284**	-0,208
	BAR(2)-27 Cat	sem	**-0,757**	0,062			
		com	**-0,878**	0,009	**0,798**	**0,273**	-0,204
	ECA Cat	sem	**-0,705**	0,030			
		com	**-0,959**	0,073	**0,840**	**0,284**	-0,142
	ANGO Cat	sem	**-0,892**	**0,189**			
		com	**-0,901**	0,063	**0,762**	**0,258**	-0,171
	ENOS Cat	sem	**-0,748**	0,153			
		com	**-0,898**	0,171	**0,824**	**0,288**	-0,151
	p22 PHOX	sem	**-0,385**	**-0,216**			
		com	**-0,875**	0,005	**0,822**	**0,280**	-0,152
	GNB3	sem	**-0,772**	0,112			
		com	**-0,885**	0,033	**0,794**	**0,281**	-0,191
Obesidade	B1BK	sem	**-0,520**	0,010			
		com	**-0,781**	0,058	**0,525**	**0,268**	0,427
	BAR(2)-16 Cat	sem	**-0,637**	0,146			
		com	**-0,784**	0,132	**0,487**	0,197	0,365
	BAR(2)-27 Cat	sem	**-0,820**	**0,218**			
		com	**-0,941**	**0,178**	**0,460**	**0,236**	0,392
	ECA Cat	sem	**-0,522**	0,024			
		com	**-0,764**	0,052	**0,523**	**0,277**	0,454
	ANGO Cat	sem	**-0,670**	0,140			
		com	**-0,750**	0,083	**0,456**	**0,245**	0,433
	ENOS Cat	sem	**-0,473**	-0,062			
		com	**-0,684**	-0,120	**0,492**	**0,255**	0,446
	p22 PHOX	sem	**-0,515**	0,015			
		com	**-0,719**	0,015	**0,526**	**0,269**	0,455
	GNB3	sem	**-0,432**	-0,082			
		com	**0,670**	-0,123	**0,569**	**0,286**	0,420

4.2.2 Análise no Nível Cromossômico

Para cada marcador, levando-se em consideração dados cromossômicos, foram feitas tabelas de contingência avaliando os grupos de *caso* e *controle*, para diabetes, hipertensão e obesidade (ver Tabela B.2 Seção B.1 do Apêndice B). Observa-se que o número total da amostra $2n = 3154$ varia, pois existem dados faltantes e como já foi mencionado, a análise foi conduzida com dados completos.

Para estudar a associação entre doença e marcador foram calculados os *Odds Ratios*, como descrito na Subseção 3.2.2 do Capítulo 3. As estimativas ($\hat{\Psi}$), com os respectivos níveis descritivos (valor p) para cada marcador em relação a cada doença, são mostrados na Tabela 4.24. Adotando-se um nível de significância de 5%, observa-se que para o marcador BAR-2(16) Cat tem-se um *Odds Ratio* significativo, indicando que o risco de hipertensão para indivíduos que carregam o marcador na categoria alélica A é 84% menor que para o grupo G. Para o marcador BAR-2(27) Cat o risco de obesidade no grupo Gln é aproximadamente 25% maior que o risco no grupo Glu. Considerando o marcador ANGO Cat e hipertensão, indivíduos que carregam o alelo T neste marcador têm o risco de hipertensão 22% maior que o risco no grupo que carrega o alelo M. Da mesma maneira, para este marcador, o risco de obesidade para o grupo T é 16% maior que o risco no grupo M. Estas foram as únicas medidas de risco significativas encontradas.

Tabela 4.24 *Estimativas dos Odds Ratios no nível cromossômico.*

	Diabetes		Hipertensão		Obesidade	
Marcador	$\hat{\Psi}$	valor p	$\hat{\Psi}$	valor p	$\hat{\Psi}$	valor p
B1BK	0,9152	0,5150	0,9168	0,2630	1,0106	0,8900
NHPS2.5	0,5648	0,2690	1,1406	0,5600	0,8673	0,5330
BAR-2(16) Cat	1,0381	0,7950	0,8440	**0,0411**	1,1646	0,0560
BAR-2(27) Cat	1,1773	0,2850	1,0646	0,4570	1,2463	**0,0081**
ECA Cat	1,0556	0,6880	1,0272	0,7250	1,0217	0,7730
ANGO Cat	1,0946	0,5060	1,2245	**0,0085**	1,1621	**0,0456**
ENOS Cat	1,0643	0,6780	1,1606	0,0801	0,9413	0,4750
p22 PHOX	1,1059	0,4670	0,9985	0,9850	1,0148	0,8500
GNB3	0,9391	0,6390	1,1228	0,1260	0,9187	0,2550

Segue na Tabela 4.25 as estimativas dos parâmetros e correspondentes erros padrão do modelo de regressão logística:

$$\ln\left(\frac{p_1(x)}{1-p_1(x)}\right) = \alpha + \beta_2 x, \tag{4.7}$$

onde $p_1(x)$ representa a probabilidade de um indivíduo do grupo alélico x pertencer ao grupo *caso*.

A seguinte codificação foi atribuída aos dados genéticos:

Para o marcador B1BK:

$$x = \begin{cases} 1, & \text{se o indivíduo carrega o alelo } G \\ 0, & \text{caso contrário} \end{cases}$$

Para o marcador ANGO Cat:

$$x = \begin{cases} 1, & \text{se o indivíduo carrega o alelo } M \\ 0, & \text{caso contrário} \end{cases}$$

E assim por diante para os demais marcadores.

Tabela 4.25 *Estimativas dos parâmetros (erro padrão) do modelo no nível cromossômico.*

	Diabetes		Hipertensão		Obesidade	
Marcador	Intercepto	$\hat{\beta}_2$	Intercepto	$\hat{\beta}_2$	Intercepto	$\hat{\beta}_2$
B1BK	-2,4035(0,1060)	-0,0887(0,1360)	-0,6116(0,0610)	-0,0869(0,0776)	-0,5140(0,0603)	0,0106(0,0763)
NHPS2.5	-2,4491(0,0670)	-0,5713(0,5164)	-0,6731(0,0382)	0,1315(0,2256)	-0,4995(0,0374)	-0,1424(0,2286)
BAR-2(16) Cat	-2,4915(0,0975)	0,0374(0,1443)	-0,6742(0,0546)	**-0,1696(0,0830)**	-0,5736(0,0540)	0,1523(0,0797)
BAR-2(27) Cat	-2,5797(0,1317)	0,1632(0,1525)	-0,7135(0,0714)	0,0626(0,0841)	-0,6644(0,0711)	**0,2202(0,0832)**
ECA Cat	-2,4841(0,1021)	0,0541(0,1345)	-0,6854(0,0575)	0,0269(0,0763)	-0,5069(0,0561)	0,0215(0,0746)
ANGO Cat	-2,5154(0,1056)	0,0904(0,1358)	-0,7878(0,0597)	**0,2025(0,0770)**	-0,5931(0,0580)	**0,1502(0,0752)**
ENOS Cat	-2,4779(0,0776)	0,0623(0,1501)	-0,7067(0,0440)	0,1490(0,0851)	-0,4889(0,0427)	-0,0605(0,0847)
p22 PHOX	-2,4955(0,0830)	0,1006(0,1384)	-0,6679(0,0464)	-0,0015(0,0795)	-0,5101(0,0455)	0,0147(0,0778)
GNB3	-2,4275(0,0885)	-0,0628(0,1340)	-0,7229(0,0514)	0,1159(0,0758)	-0,4681(0,0496)	-0,0848(0,0744)

Assumindo um modelo multiplicativo tem-se, por exemplo, para o marcador BAR-2(16) Cat com hipertensão $\hat{\Psi}_{AG|GG} = \hat{\Psi}_{A|G} = 0{,}8440$ e $\hat{\Psi}_{AA|GG} = \hat{\Psi}^2_{A|G} = 0{,}7123$. Para o marcador BAR-2(27) Cat com obesidade $\hat{\Psi}_{GlnGlu|GluGlu} = \hat{\Psi}_{Gln|Glu} = 1{,}2463$ e $\hat{\Psi}_{GlnGln|GluGlu} = \hat{\Psi}^2_{Gln|Glu} = 1{,}5533$. Da mesma maneira para o marcador ANGO Cat com hipertensão $\hat{\Psi}_{MT|MM} = \hat{\Psi}_{T|M} = 1{,}2245$ e $\hat{\Psi}_{TT|MM} = \hat{\Psi}^2_{T|M} = 1{,}4994$ e para obesidade $\hat{\Psi}_{MT|MM} = \hat{\Psi}_{T|M} = 1{,}1621$ e $\hat{\Psi}_{TT|MM} = \hat{\Psi}^2_{T|M} = 1{,}3505$.

Observou-se na subseção anterior que comparando os modelos genotípicos (saturado e multiplicativo) alguns marcadores não mostraram evidência de associação, outros continuaram significativos. Da mesma maneira, quando compara-se os três modelos (incluindo o cromossômico), os marcadores BAR-2(16) Cat e ANGO Cat com hipertensão continuam significativos. Nota-se que o marcador ANGO Cat com obesidade mostrou evidência significativa de associação para a análise no nível cromossômico, o que não foi verificado nos outros modelos possivelmente, devido à ocorrência de algum efeito de interação entre os alelos na composição do genótipo correspondente.

Fazendo a análise estratificando por etnia, calculou-se as estimativas de Mantel-Haenszel e seus respectivos valores p, para cada marcador e cada doença (ver Tabela 4.26). Neste caso, os únicos marcadores com efeito significativo foram BAR-2(16) Cat e ENOS Cat para hipertensão, e os marcadores BAR-2(27) Cat e GNB3 para obesidade. Nota-se que ao combinar as medidas de associação por etnia tem-se diferentes evidências de associação entre as doenças e os marcadores.

Tabela 4.26 *Estimativas de Mantel-Haenszel controlando por etnia no nível cromossômico.*

	Diabetes		Hipertensão		Obesidade	
Marcador	$\hat{\Phi}_{MH}$	valor p	$\hat{\Phi}_{MH}$	valor p	$\hat{\Phi}_{MH}$	valor p
B1BK	0,9156	0,5662	0,9698	0,7271	1,0599	0,4749
NHPS2.5	0,5633	0,3574	1,2246	0,4391	0,9142	0,7800
BAR-2(16) Cat	1,0200	0,9480	0,8378	**0,0386**	1,1473	0,0949
BAR-2(27) Cat	1,1593	0,3696	1,0091	0,9504	1,1957	**0,0376**
ECA Cat	1,0622	0,7045	1,0667	0,4270	1,0476	0,5629
ANGO Cat	1,0921	0,5680	1,1449	0,0933	1,1208	0,1465
ENOS Cat	1,0663	0,7270	1,1996	**0,0402**	0,9583	0,6496
p22 PHOX	1,1082	0,5025	1,0050	0,9826	1,0143	0,8870
GNB3	0,9333	0,6594	1,0256	0,7746	0,8595	**0,0497**

Como visto na subseção 3.2.3 é possível considerar modelos logísticos para a análise no nível genotípico e cromossômico ao incluir no estudo possíveis variáveis de confundimento. Nas Tabelas 4.27, 4.28, 4.29, 4.30, 4.31, 4.32, 4.33, 4.34, 4.35 estão as estimativas dos parâmetros do modelo para cada marcador considerando diabetes, hipertensão e obesidade

no nível cromossômico. Somente para o marcador $GNB3$ com obesidade foi encontrado efeito significativo de interação entre o risco relativo da doença com raça. Neste caso, as únicas medidas de risco significativas para as etnias "negro" e "branco", respectivamente foram:

$$\begin{aligned} \hat{\Psi}^{B}_{T|C} &= 2,0322 \Rightarrow \hat{\Psi}^{B}_{C|T} = 0,4921, \\ \hat{\Psi}^{N}_{T|C} &= 0,5290, \end{aligned} \tag{4.8}$$

em que, o risco de obesidade para indivíduos "brancos" que carregam uma cópia do alelo C é 49% menor que aqueles que carregam uma cópia do alelo T e para indivíduos "negros" que carregam uma cópia do alelo T é aproximadamente 53% menor que aqueles que carregam uma cópia do alelo C. Observa-se que todos os 9 marcadores mostraram efeito significativo de raça, indicando que o padrão de ocorrência da doença depende da raça. Como comentado anteriormente, raça é um fator de heterogeneidade genética e, assim, a validação dos marcadores identificados nesta população merece um certo cuidado.

Tabela 4.27 *Estimativas do modelo para o marcador B1BK no nível cromossômico estratificando por etnia.*

	Diabetes		Hipertensão		Obesidade	
Parâmetro	Estimativa(EP)	valor p	Estimativa(EP)	valor p	Estimativa(EP)	valor p
α	-2,4699(0,1968)	**<0,0001**	-0,8109(0,1136)	**<0,0001**	-0,6849(0,1110)	**<0,0001**
β_2	-0,0529(0,2424)	0,8270	-0,0825(0,1398)	0,5553	-0,0289(0,1362)	0,8322
γ_2	0,0904(0,3847)	0,8140	0,6103(0,2158)	**0,0047**	0,4167(0,2151)	0,0527
γ_3	0,1922(0,2410)	0,4250	0,2653(0,1410)	0,0599	0,1836(0,1388)	0,1857
γ_4	-0,9313(0,7446)	0,2110	-0,3312(0,3176)	0,2970	0,3595(0,2803)	0,1997
δ_2	-0,5465(0,5858)	0,3510	0,3477(0,2924)	0,2345	0,1679(0,2913)	0,5644
δ_3	-0,0776(0,3031)	0,7980	0,0177(0,1763)	0,9202	0,1247(0,1726)	0,4699
δ_4	1,1515(0,8273)	0,1640	0,1974(0,3928)	0,6153	0,0983(0,3489)	0,7781

Tabela 4.28 *Estimativas do modelo para o marcador NHPS no nível cromossômico estratificando por etnia.*

	Diabetes		Hipertensão		Obesidade	
Parâmetro	Estimativa(EP)	valor p	Estimativa(EP)	valor p	Estimativa(EP)	valor p
α	-2,5058(0,1170)	**<0,0001**	-0,8860(0,0677)	**<0,0001**	-0,7131(0,0655)	**<0,0001**
β_2	0,0209(0,6122)	0,9730	0,5231(0,3325)	0,1157	0,2431(0,3356)	0,4688
γ_2	-0,1288(0,2840)	0,6500	0,7939(0,1461)	**<0,0001**	0,5201(0,1456)	**0,0004**
γ_3	0,1633(0,1477)	0,2690	0,2928(0,0860)	0,0007	0,2818(0,0837)	**0,0008**
γ_4	-0,0656(0,3216)	0,8330	-0,0976(0,1856)	0,5991	0,4873(0,1681)	**0,0037**
δ_2	-11,9524(509,65)	0,9801	0,2622(1,2757)	0,8372	0,6431(1,2765)	0,6144
δ_3	-1,2893(1,1874)	0,2780	-0,5097(0,4741)	0,2824	-0,6227(0,4855)	0,1996
δ_4	-12,0156(394,78)	0,9760	-13,1056(239,44)	0,9564	-1,4036(1,1775)	0,2333

Tabela 4.29 *Estimativas do modelo para o marcador BAR-2(16) Cat no nível cromossômico estratificando por etnia.*

	Diabetes		Hipertensão		Obesidade	
Parâmetro	Estimativa(EP)	valor p	Estimativa(EP)	valor p	Estimativa(EP)	valor p
α	-2,4907(0,1587)	**<0,0001**	-0,9312(0,0933)	**<0,0001**	-0,7710(0,0904)	**<0,0001**
β_2	-0,0076(0,2451)	0,9750	0,0012(0,1440)	0,9934	0,2315(0,1366)	0,0901
γ_2	-0,8051(0,5331)	0,1310	0,6435(0,2125)	**0,0025**	0,4468(0,2117)	**0,0348**
γ_3	0,1236(0,2084)	0,5530	0,4241(0,1215)	**0,0005**	0,2761(0,1192)	**0,0205**
γ_4	0,0799(0,4253)	0,8510	-0,1209(0,2647)	0,6478	0,5108(0,2367)	**0,0309**
δ_2	0,9761(0,6755)	0,1480	0,1530(0,3191)	0,6317	0,0928(0,3159)	0,7690
δ_3	0,0215(0,3140)	0,9450	-0,3453(0,1855)	0,0627	-0,1620(0,1771)	0,3603
δ_4	-0,8651(0,7492)	0,2480	-0,2331(0,3912)	0,5512	-0,2871(0,3405)	0,3992

Tabela 4.30 *Estimativas do modelo para o marcador BAR-2(27) Cat no nível cromossômico estratificando por etnia.*

	Diabetes		Hipertensão		Obesidade	
Parâmetro	Estimativa(EP)	valor p	Estimativa(EP)	valor p	Estimativa(EP)	valor p
α	-2,8048(0,2247)	**<0,0001**	-0,8884(0,1138)	**<0,0001**	-0,8884(0,1138)	**<0,0001**
β_2	0,4280(0,2616)	0,1020	0,0345(0,1399)	0,8055	0,2754(0,1380)	**0,0459**
γ_2	0,3625(0,5677)	0,5230	1,0487(0,3057)	**0,0006**	0,5656(0,3083)	0,0666
γ_3	0,4630(0,2860)	0,1060	0,2835(0,1544)	0,0663	0,3160(0,1541)	**0,0402**
γ_4	-0,2156(0,7582)	0,7760	-0,3055(0,3785)	0,4195	0,9349(0,3256)	**0,0041**
δ_2	-0,6937(0,6550)	0,2900	-0,3187(0,3476)	0,3592	-0,1180(0,3492)	0,7354
δ_3	-0,4521(0,3325)	0,1740	-0,0158(0,1848)	0,9318	-0,0966(0,1826)	0,5967
δ_4	0,1159(0,8373)	0,8900	0,1319(0,4357)	0,7620	-0,7145(0,3799)	0,0600

Tabela 4.31 *Estimativas do modelo para o marcador ECA Cat no nível cromossômico estratificando por etnia.*

	Diabetes		Hipertensão		Obesidade	
Parâmetro	Estimativa(EP)	valor p	Estimativa(EP)	valor p	Estimativa(EP)	valor p
α	-2,5000(0,1811)	**<0,0001**	-0,0999(0,1055)	**<0,0001**	-0,7483(0,1022)	**<0,0001**
β_2	-0,0052(0,2342)	0,9824	0,0634(0,1357)	0,6407	0,0781(0,1315)	0,5527
γ_2	0,2664(0,3536)	0,4513	1,0542(0,2081)	**<0,0001**	0,7643(0,2060)	**0,0002**
γ_3	0,0275(0,2299)	0,9049	0,2595(0,1323)	**0,0498**	0,2741(0,1285)	**0,0330**
γ_4	-0,2214(0,4960)	0,6554	-0,4158(0,2925)	0,1553	0,4245(0,2473)	0,0860
δ_2	-1,0571(0,6373)	0,0972	-0,4747(0,2943)	0,1068	-0,4336(0,2928)	0,1386
δ_3	0,2174(0,2973)	0,4647	0,0377(0,1724)	0,8267	0,0030(0,1677)	0,9856
δ_4	0,2180(0,6500)	0,7373	0,4707(0,3781)	0,2132	0,0516(0,3338)	0,8772

Tabela 4.32 *Estimativas do modelo para o marcador ANGO Cat no nível cromossômico estratificando por etnia.*

	Diabetes		Hipertensão		Obesidade	
Parâmetro	Estimativa(EP)	valor p	Estimativa(EP)	valor p	Estimativa(EP)	valor p
α	-2,4039(0,1557)	**<0,0001**	-0,9150(0,0945)	**<0,0001**	-0,7652(0,0918)	**<0,0001**
β_2	-0,2117(0,2307)	0,3590	0,0981(0,1325)	0,4590	0,1211(0,1286)	0,3463
γ_2	0,1703(0,4569)	0,7090	1,0442(0,2715)	**0,0001**	0,3732(0,2747)	0,1743
γ_3	-0,1556(0,2220)	0,4830	0,2169(0,1281)	0,0904	0,2914(0,1243)	**0,0190**
γ_4	-0,7316(0,6098)	0,2300	-0,5064(0,3124)	0,1050	0,2544(0,2602)	0,3282
δ_2	-0,3879(0,5862)	0,5080	-0,3594(0,3231)	0,2661	0,1389(0,3250)	0,6692
δ_3	0,5171(0,3003)	0,0850	0,0697(0,1714)	0,6842	-0,0672(0,1667)	0,6870
δ_4	1,0118(0,7230)	0,1620	0,5405(0,3896)	0,1653	0,3112(0,3392)	0,3589

Tabela 4.33 *Estimativas do modelo para o marcador ENOS Cat no nível cromossômico estratificando por etnia.*

	Diabetes		Hipertensão		Obesidade	
Parâmetro	Estimativa(EP)	valor p	Estimativa(EP)	valor p	Estimativa(EP)	valor p
α	-2,4965(0,1378)	**<0,0001**	-0,8676(0,0798)	**<0,0001**	-0,6694(0,0769)	**<0,0001**
β_2	-0,0276(0,2494)	0,9120	0,0062(0,1430)	0,9652	-0,1139(0,1402)	0,4166
γ_2	-0,0902(0,3095)	0,7710	0,6871(0,1629)	**<0,0001**	0,5292(0,1613)	**0,0010**
γ_3	0,0841(0,1732)	0,6270	0,2077(0,1002)	**0,0382**	0,2072(0,0969)	**0,0325**
γ_4	-0,0212(0,3565)	0,9530	-0,1339(0,2106)	0,5250	0,5199(0,1895)	**0,0061**
δ_2	-0,3815(0,8148)	0,6400	0,6369(0,3694)	0,0847	-0,2085(0,3683)	0,5712
δ_3	0,2309(0,3212)	0,4720	0,2680(0,1874)	0,1526	0,2059(0,1847)	0,2650
δ_4	-0,3992(0,8346)	0,6320	-0,1034(0,4379)	0,8133	-0,3556(0,3995)	0,3733

Tabela 4.34 *Estimativas do modelo para o marcador p22 PHOX no nível cromossômico estratificando por etnia.*

	Diabetes		Hipertensão		Obesidade	
Parâmetro	Estimativa(EP)	valor p	Estimativa(EP)	valor p	Estimativa(EP)	valor p
α	-2,4970(0,1423)	**<0,0001**	-0,8499(0,0809)	**<0,0001**	-0,7662(0,0796)	**<0,0001**
β_2	-0,0184(0,2444)	0,9400	-0,0664(0,1415)	0,6389	0,1664(0,1355)	0,2193
γ_2	-0,2331(0,3715)	0,5300	0,7290(0,1830)	**<0,0001**	-0,5642(0,1829)	**0,0020**
γ_3	0,0980(0,1795)	0,5850	0,2286(0,1036)	**0,0273**	0,3432(0,1016)	**0,0007**
γ_4	-0,3747(0,4425)	0,3970	0,0615(0,2193)	0,7793	0,6231(0,2055)	**0,0024**
δ_2	0,2399(0,5767)	0,6770	0,2519(0,3000)	0,4012	-0,1152(0,2979)	0,6990
δ_3	0,1270(0,3078)	0,6800	0,1499(0,1794)	0,4032	-0,2177(0,1738)	0,2103
δ_4	0,6565(0,6483)	0,3110	-0,6791(0,4148)	0,1016	-0,4828(0,3496)	0,1673

Tabela 4.35 *Estimativas do modelo para o marcador GNB3 no nível cromossômico estratificando por etnia.*

	Diabetes		Hipertensão		Obesidade	
Parâmetro	Estimativa(EP)	valor p	Estimativa(EP)	valor p	Estimativa(EP)	valor p
α	-2,5024(0,1442)	**<0,0001**	-0,8931(0,0837)	**<0,0001**	-0,7105(0,0809)	**<0,0001**
β_2	0,0092(0,2385)	0,9690	0,0483(0,1379)	0,7265	0,0174(0,1338)	0,8966
γ_2	-0,5772(0,5313)	0,2770	0,8272(0,2259)	**0,0003**	0,9090(0,2257)	**<0,0001**
γ_3	0,2429(0,1898)	0,1960	0,3004(0,1114)	**0,0070**	0,3758(0,1079)	**0,0005**
γ_4	-0,1368(0,42170)	0,7430	-0,2182(0,2411)	0,3654	0,3446(0,2143)	0,1078
δ_2	0,6209(0,6394)	0,3320	-0,0477(0,2986)	0,1600	-0,6541(0,2995)	**0,0290**
δ_3	-0,2120(0,2992)	0,4780	-0,0396(0,1734)	0,8193	-0,2366(0,1691)	0,1617
δ_4	0,1122(0,6528)	0,8640	0,0617(0,3822)	0,8719	0,1990(0,3425)	0,5612

Fazendo um paralelo do modelo da análise no nível cromossômico (equação 4.7) com o modelo incluindo a variável etnia (sem efeito de interação), ver Tabela 4.36, nota-se que o efeito do genótipo para os marcadores BAR(2)-16 Cat com hipertensão e BAR(2)-27 Cat com obesidade continuaram significativo. Para o marcador ANGO Cat com hipertensão e obesidade o efeito do genótipo na análise marginal é significativo, mas ao incluir a variável etnia não observa-se este efeito. As mesmas ressalvas feitas na interpretação dos resultados da Tabela 4.23 valem também para este caso.

Tabela 4.36 *Estimativas dos parâmetros do modelo cromossômico sem e com a variável etnia (sem interação).*

	Marcador		$\hat{\alpha}$	$\hat{\beta}_2$	$\hat{\gamma}_1$	$\hat{\gamma}_2$	$\hat{\gamma}_3$
Diabetes	B1BK	sem	**-2,404**	-0,087			
		com	**-2,447**	-0,088	-0,157	0,144	-0,088
	BAR(2)-16 Cat	sem	**-2,492**	0,037			
		com	**-2,502**	0,020	-0,269	0,132	-0,268
	BAR(2)-27 Cat	sem	**-2,580**	0,163			
		com	**-2,606**	0,149	-0,162	0,133	-0,098
	ECA Cat	sem	**-2,484**	0,054			
		com	**-2,539**	0,060	-0,111	0,156	-0,096
	ANGO Cat	sem	**-2,515**	0,090			
		com	**-2,550**	0,088	-0,164	0,133	-0,105
	ENOS Cat	sem	**-2,478**	0,062			
		com	**-2,525**	0,064	-0,134	0,149	-0,092
	p22 PHOX	sem	**-2,496**	0,101			
		com	**-2,538**	0,103	-0,142	0,141	-0,102
	GNB3	sem	**-2,428**	-0,063			
		com	**-2,474**	0,069	-0,131	0,153	-0,089
Hipertensão	B1BK	sem	**-0,612**	-0,087			
		com	**-0,845**	-0,031	**0,795**	**0,279**	-0,203
	BAR(2)-16 Cat	sem	**-0,674**	**-0,170**			
		com	**-0,858**	**-0,178**	**0,718**	**0,276**	-0,208
	BAR(2)-27 Cat	sem	**-0,714**	0,063			
		com	**-0,872**	0,009	**0,799**	**0,274**	-0,203
	ECA Cat	sem	**-0,685**	0,027			
		com	**-0,911**	0,065	**0,830**	**0,281**	-0,147
	ANGO Cat	sem	**-0,788**	**0,203**			
		com	**-0,934**	0,135	**0,767**	**0,256**	-0,170
	ENOS Cat	sem	**-0,707**	0,149			
		com	**-0,923**	**0,182**	**0,825**	**0,287**	-0,143
	p22 PHOX	sem	**-0,668**	-0,002			
		com	**-0,874**	0,005	**0,822**	**0,279**	-0,152
	GNB3	sem	**-0,773**	0,116			
		com	**-0,885**	0,025	**0,803**	**0,284**	-0,193
Obesidade	B1BK	sem	**-0,514**	0,011			
		com	**-0,743**	0,058	**0,516**	**0,265**	**0,425**
	BAR(2)-16 Cat	sem	**-0,574**	0,015			
		com	**-0,730**	0,138	**0,490**	**0,203**	**0,374**
	BAR(2)-27 Cat	sem	**-0,664**	**0,220**			
		com	**-0,823**	**0,179**	**0,483**	**0,251**	0,408
	ECA Cat	sem	**-0,507**	0,022			
		com	**-0,729**	0,046	**0,556**	**0,274**	**0,450**
	ANGO Cat	sem	**-0,593**	**0,150**			
		com	**-0,762**	0,114	**0,479**	**0,250**	0,440
	ENOS Cat	sem	**-0,489**	-0,061			
		com	**-0,691**	-0,043	**0,501**	**0,261**	**0,447**
	p22 PHOX	sem	**-0,510**	0,015			
		com	**-0,714**	0,014	**0,527**	**0,269**	**0,455**
	GNB3	sem	**-0,468**	-0,085			
		com	**0,650**	**-0,123**	**0,548**	**0,283**	**0,428**

4.3 Trios de São Paulo

Para este banco de dados, os trios são de pacientes com cardiopatiopatias congênitas (mal-formações cardíacas no coração) que procuraram o serviço do INCOR/USP e os marcadores estudados estão todos no gene RALDH2, um gene candidato a conferir risco a essas mal-formações. Neste caso, os dados são compostos de 131 trios, com 5 marcadores sendo que, foram considerados na análise trios significativos com pelo menos um dos pais heterozigotos. Para os marcadores M1, M2, M3, M4, M5 foram utilizadas amostras de 57, 59, 69, 54 e 47 trios, respectivamente.

A análise foi feita de acordo com o descrito na Seção 3.3 do Capítulo 3 (Tabela 3.12) e utilizando recursos dos aplicativos MS-EXCEL e **R** (http://www.R-project.org).

Pelas Tabelas 4.37, 4.38, 4.39, 4.40 e 4.41, os valores das estatísticas TDT com os respectivos valores p, considerando a Tabela 3.12, isto é, baseado na estatística de McNemar, não são informativos, ao nível de significância de 5% e, portanto, não há evidência amostral de associação para nenhum dos marcadores sob análise.

Tabela 4.37 *Transmissão de alelos - Marcador M1.*

	Não Transmitidos		
Transmitidos	2	4	Total
2	31	42	55
4	38	21	59
Total	51	63	114

TDT= 0,2000 (valor p= 0,6547)

Tabela 4.38 *Transmissão de alelos - Marcador M2.*

	Não Transmitidos		
Transmitidos	3	4	Total
3	18	43	61
4	36	21	57
Total	54	64	118
TDT= 0,6203	(valor p= 0,4309)		

Tabela 4.39 *Transmissão de alelos - Marcador M3.*

	Não Transmitidos		
Transmitidos	1	3	Total
1	30	44	74
3	45	19	64
Total	75	63	138
TDT= 0,0112	(valor p= 0,9157)		

Tabela 4.40 *Transmissão de alelos - Marcador M4.*

	Não Transmitidos		
Transmitidos	1	3	Total
1	10	36	46
3	33	29	62
Total	43	65	108
TDT= 0,1304	(valor p= 0,7180)		

Tabela 4.41 *Transmissão de alelos - Marcador M5.*

	Não Transmitidos		
Transmitidos	2	4	Total
2	37	28	65
4	29	0	29
Total	66	28	94
TDT= 0,0175	(valor p= 0,8948)		

Capítulo 5

Considerações Finais

Tem crescido o interesse na literatura por estudos de associação como instrumento para o mapeamento genético de doenças. Neste trabalho foram considerados alguns conceitos fundamentais de genética, os quais foram introduzidos no contexto dos estudos de associação genética. Foram abordados alguns delineamentos amostrais para coleta de dados genéticos úteis para o estudo de associação entre fatores de risco genético e doença. Também foram discutidos métodos estatísticos de análise de dados específicos a cada situação, juntamente com algumas aplicações.

Um problema comum em estudos de associação é a ocorrência de falso-positivos devido à estrutura da população, como é o caso de populações estratificadas por etnia. Em dados genéticos este problema é intensificado pela limitação em caracterizar grupos *caso* e *controle* quanto à sua constituição genômica, o que tem conduzido à não reprodutibilidade dos estudos publicados na literatura.

No presente trabalho, para ilustrar o problema teórico analisou-se conjuntos de dados reais por meio do programa **R**. No intuito de verificar se o risco da doença está associado aos dados de marcadores, para a "População de Vitória" alguns resultados importantes foram notados. Considerando a análise no nível genotípico, por exemplo, o marcador BAR-2(16) Cat é significativo em relação ao genótipo AG com a hipertensão e obesidade. Na análise no nível cromossômico, por exemplo, para o marcador ANGO Cat e hipertensão, para indivíduos que carregam o alelo T neste marcador, o risco de hipertensão é 22% maior que o risco no grupo que carrega o alelo M.

Além destes resultados, para as análises no nível genotípico e cromossômico fez-se estratificação por etnia por meio da metodologia de Mantel-Haenszel e modelos logísticos.

Somente para obesidade os marcadores BAR-2(27) Cat e GNB3 mostraram efeito significativo. Além disso para esta população (Vitória) as diferentes raças são consideradas um fator de heterogeneidade genética importante. Como visto anteriormente, a validação dos marcadores identificados como fatores de risco merece alguns cuidados sobre possíveis resultados falso-positivos.

Nota-se, que dos marcadores sob análise nenhum mostrou evidência de associação para diabetes. Para os "Trios de São Paulo" não houve evidência amostral de associação para nenhum dos marcadores sob análise.

A seguir destaca-se alguns tópicos de futuras pesquisas a serem exploradas na área de Epidemiologia Genética:

- Avaliação simultânea de múltiplos locos como fatores de risco para doenças. Neste caso, se as análises forem feitas separadamente surgem problemas de correção nas estatísticas devido à realização de múltiplos testes, não necessariamente independentes. Além disso, um grande desafio é modelar o padrão de possíveis covariâncias entre os vários locos em um mapa sob estudo e o efeito dos "verdadeiros" genes distribuídos no mapa. A análise de múltiplos locos pode ser feita por meio de um único modelo logístico, porém para mais de 3 locos os efeitos de interação ficam difíceis de interpretar;

- Avaliação simultânea de múltiplas doenças, possivelmente sendo influenciadas pelo mesmo conjunto de genes comuns. Neste caso, surge a necessidade de generalizar as metodologias aplicadas para uma única doença em situações multivariadas;

- Utilização da Teoria da Resposta ao Item (TRI), a estudos caso-controle em que o efeito de muitos locos marcadores está sendo pesquisado. Vale ressaltar que esta metodologia tem sido aplicada na área de genética, por exemplo, Tavares et al. (2004);

- Aplicação de Meta-análise para combinar diferentes estudos epidemiológicos com dados genéticos. A utilidade de tais metodologias tem crescido imensamente já que não é esperado entender o complexo modelo genético de regulação de doenças com um único estudo observacional ou experimental;

- Desenvolver metodologias estatísticas que contemplem situações de dados incompletos.

Apêndice A

Expressão para o TDT

A.1 Capítulo 3- TDT

Referente à:

Seção 3.3

Subseção 3.3.3

A seguir são apresentadas as probabilidades correspondentes à Tabela 3.12 em função de parâmetros de estratificação, associação e fração de recombinação (Clayton, 2003).

$$
\begin{aligned}
p_{11} &= \left[\sum_j \alpha_j (p_j q_j^2 + \delta_j q_j) p_j\right] / \left[\sum_j \alpha_j p_j^2\right], \\
p_{12} &= \left[\sum_j \alpha_j [p_j q_j (1-q_j) + \delta_j (1-\theta-q_j)] p_j\right] / \left[\sum_j \alpha_j p_j^2\right], \\
p_{21} &= \left[\sum_j \alpha_j [p_j q_j (1-q_j) + \delta_j (\theta-q_j)] p_j\right] / \left[\sum_j \alpha_j p_j^2\right], \\
p_{22} &= \left[\sum_j \alpha_j [p_j (1-q_j)^2 + \delta_j (1-q_j)] p_j\right] / \left[\sum_j \alpha_j p_j^2\right],
\end{aligned}
\tag{A.1}
$$

onde α_j é a proporção da população no estrato j,

q_j e p_j são as probabilidades de H (alelo marcador) e o gene da doença, respectivamente no estrato j,

δ_j é o coeficiente de associação no estrato j e

θ é o parâmetro fração de recombinação entre o loco de marcador e da doença.

Apêndice B

Tabelas de Contingência

B.1 Capítulo 4- Conjunto de dados da População de Vitória

Tabelas referente ao conjunto de dados "População de Vitória"[1].

Análise no nível genotípico

Tabela B.1 *Tabelas de freqüência 2×3 referente ao marcador e cada fenótipo.*

	B1BK					B1BK					B1BK			
Diabetes	AA	AG	GG	Total	Hipertensão	AA	AG	GG	Total	Obesidade	AA	AG	GG	Total
Caso	50	49	24	123	Caso	210	231	92	533	Caso	235	265	87	587
Controle	564	673	200	1437	Controle	406	497	134	1037	Controle	380	456	139	975
Total	614	722	224	1560	Total	616	728	226	1570	Total	615	721	226	1562
	NHPS2.5					NHPS2.5					NHPS2.5			
Diabetes	AA	GA	GG	Total	Hipertensão	AA	GA	GG	Total	Obesidade	AA	GA	GG	Total
Caso	0	4	119	123	Caso	0	32	501	533	Caso	0	30	560	590
Controle	1	80	1361	1442	Controle	1	53	987	1041	Controle	1	55	920	976
Total	1	84	1480	1565	Total	1	85	1488	1574	Total	1	85	1480	1566
	BAR-2(16) Cat					BAR-2(16) Cat					BAR-2(16) Cat			
Diabetes	AA	AG	GG	Total	Hipertensão	AA	AG	GG	Total	Obesidade	AA	AG	GG	Total
Caso	24	48	33	105	Caso	87	192	157	436	Caso	113	255	141	509
Controle	263	591	393	1247	Controle	200	451	271	922	Controle	174	385	284	843
Total	287	639	426	1352	Total	287	643	428	1358	Total	287	640	425	1352

[1] Para o marcador BAR-2(27) Cat considere QQ=GlnGln, QE=GlnGlu e EE=GluGlu.

Continuação da Tabela B.1.

	BAR-2(27) Cat					BAR-2(27) Cat					BAR-2(27) Cat			
Diabetes	QQ	QE	EE	Total	Hipertensão	QQ	QE	EE	Total	Obesidade	QQ	QE	EE	Total
Caso	68	48	7	123	Caso	280	214	39	533	Caso	330	218	41	589
Controle	743	576	121	1440	Controle	535	414	91	1040	Controle	481	407	88	976
Total	811	624	128	1563	Total	815	628	130	1573	Total	811	625	129	1565

	Eca Cat					Eca Cat					Eca Cat			
Diabetes	DD	DI	II	Total	Hipertensão	DD	DI	II	Total	Obesidade	DD	DI	II	Total
Caso	38	66	19	123	Caso	155	293	81	529	Caso	177	315	97	589
Controle	415	783	232	1430	Controle	302	561	171	1034	Controle	276	535	155	966
Total	453	849	251	1553	Total	457	854	252	1563	Total	453	850	252	1555

	ANGO Cat					ANGO Cat					ANGO Cat			
Diabetes	TT	MT	MM	Total	Hipertensão	TT	MT	MM	Total	Obesidade	TT	MT	MM	Total
Caso	47	55	21	123	Caso	211	238	85	534	Caso	231	256	103	590
Controle	516	652	274	1442	Controle	358	469	214	1041	Controle	334	450	193	977
Total	563	707	295	1565	Total	569	707	299	1575	Total	565	706	296	1567

	ENOS Cat					ENOS Cat					ENOS Cat			
Diabetes	TT	GT	GG	Total	Hipertensão	TT	GT	GG	Total	Obesidade	TT	GT	GG	Total
Caso	8	52	65	123	Caso	39	218	277	534	Caso	31	233	326	590
Controle	88	563	791	1442	Controle	59	399	583	1041	Controle	65	381	531	977
Total	96	613	856	1565	Total	98	617	860	1575	Total	96	614	857	1567

	p22 PHOX					p22 PHOX					p22 PHOX			
Diabetes	TT	TC	CC	Total	Hipertensão	TT	TC	CC	Total	Obesidade	TT	TC	CC	Total
Caso	13	63	47	123	Caso	57	249	277	533	Caso	68	268	253	589
Controle	160	656	624	1440	Controle	117	475	448	1040	Controle	105	453	418	976
Total	173	719	671	1563	Total	174	724	675	1573	Total	173	721	671	1565

	GNB3					GNB3					GNB3			
Diabetes	TT	CT	CC	Total	Hipertensão	TT	CT	CC	Total	Obesidade	TT	CT	CC	Total
Caso	26	55	42	123	Caso	122	253	155	530	Caso	114	284	188	586
Controle	299	693	441	1433	Controle	207	498	331	1036	Controle	213	464	295	972
Total	325	748	483	1556	Total	329	751	486	1566	Total	327	748	483	1558

Análise no nível cromossômico

Tabela B.2 *Tabelas de freqüência 2×2 referente ao marcador e cada fenótipo.*

	B1BK				B1BK				B1BK		
Diabetes	A	G	Total	Hipertensão	A	G	Total	Obesidade	A	G	Total
Caso	149	97	246	Caso	651	415	1066	Caso	735	439	1174
Controle	1801	1073	2874	Controle	1309	765	2074	Controle	1216	734	1950
Total	1950	1170	3120	Total	1960	1180	3140	Total	1951	1173	3124

	NHPS2.5				NHPS2.5				NHPS2.5		
Diabetes	A	G	Total	Hipertensão	A	G	Total	Obesidade	A	G	Total
Caso	4	242	246	Caso	32	1034	1066	Caso	30	1150	1180
Controle	82	2802	2884	Controle	55	2027	2082	Controle	57	1895	1952
Total	86	3044	3130	Total	87	3061	3148	Total	87	3045	3132

	BAR-2(16) Cat				BAR-2(16) Cat				BAR-2(16) Cat		
Diabetes	A	G	Total	Hipertensão	A	G	Total	Obesidade	A	G	Total
Caso	96	114	210	Caso	366	506	872	Caso	481	537	1018
Controle	1117	1377	2494	Controle	851	993	1844	Controle	733	953	1686
Total	1213	1491	2704	Total	1217	1499	2716	Total	1214	1490	2704

	BAR-2(27) Cat				BAR-2(27) Cat				BAR-2(27) Cat		
Diabetes	Gln	Glu	Total	Hipertensão	Gln	Glu	Total	Obesidade	Gln	Glu	Total
Caso	184	62	246	Caso	774	292	1066	Caso	878	300	1178
Controle	2062	818	2880	Controle	1484	596	2080	Controle	1369	583	1952
Total	2246	880	3126	Total	2258	888	3146	Total	2247	883	3130

	Eca Cat				Eca Cat				Eca Cat		
Diabetes	D	I	Total	Hipertensão	D	I	Total	Obesidade	D	I	Total
Caso	142	104	246	Caso	603	455	1058	Caso	669	509	1178
Controle	1613	1247	2860	Controle	1165	903	2068	Controle	1087	845	1932
Total	1755	1351	3106	Total	1768	1358	3126	Total	1756	1354	3110

	ANGO Cat				ANGO Cat				ANGO Cat		
Diabetes	T	M	Total	Hipertensão	T	M	Total	Obesidade	T	M	Total
Caso	149	97	246	Caso	660	408	1068	Caso	718	462	1180
Controle	1684	1200	2884	Controle	1185	897	2082	Controle	1118	836	1954
Total	1833	1297	3130	Total	1845	1305	3150	Total	1836	1298	3134

Continuação da Tabela B.2.

	ENOS Cat				ENOS Cat				ENOS Cat		
Diabetes	T	G	Total	Hipertensão	T	G	Total	Obesidade	T	G	Total
Caso	66	180	246	Caso	296	772	1068	Caso	295	885	1180
Controle	739	2145	2884	Controle	517	1565	2082	Controle	511	1443	1954
Total	805	2325	3130	Total	813	2337	3150	Total	806	2328	3134
	p22 PHOX				p22 PHOX				p22 PHOX		
Diabetes	T	C	Total	Hipertensão	T	C	Total	Obesidade	T	C	Total
Caso	89	157	246	Caso	363	703	1066	Caso	404	774	1178
Controle	976	1904	2880	Controle	709	1371	2080	Controle	663	1289	1952
Total	1065	2061	3126	Total	1072	2074	3146	Total	1067	2063	3130
	GNB3				GNB3				GNB3		
Diabetes	T	C	Total	Hipertensão	T	C	Total	Obesidade	T	C	Total
Caso	107	139	246	Caso	497	563	1060	Caso	512	660	1172
Controle	1291	1575	2866	Controle	912	1160	2072	Controle	890	1054	1944
Total	1398	1714	3112	Total	1409	1723	3132	Total	1402	1714	3116

Estratificação da população

Tabela B.3 *Variável etnia.*

Etnia	Freqüência
Branco	548
Negro	122
Mulato	795
Outro	87

Apêndice C

Aplicação no nível genotípico

C.1 Capítulo 4- Aplicação

Referente ao:

Capítulo 4- Seção 4.2

Subseção 4.2.1

Análise no nível genotípico

Para a análise no nível genotípico, pode ser definido também os seguintes níveis genotípicos para cada marcador: B1BK (AG e GG), NHPS2.5 (GA e GG), BAR-2(16) Cat (AG e GG), BAR-2(27) Cat ($GlnGlu$ e $GluGlu$), Eca Cat (DI e II), ANGO Cat (MT e MM), ENOS Cat (GT e GG), p22 PHOX (TC e CC), GNB3 (CT e CC). A Tabela C.1, apresenta as estimativas dos *Odds Ratios*[1] obtidas e os correspondentes valores p para o estudo da associação. Observa-se que para o marcador BAR-2(27) Cat existe associação significativa em relação ao genótipo $GlnGlu$ com a obesidade. Nota-se também, que para o marcador ANGO Cat, há evidência de associação significativa para o genótipo MM e para o marcador p22PHOX para o genótipo CC, com a hipertensão.

Segue nas Tabelas C.2, C.3 e C.4 as estimativas dos parâmetros e correspondentes erros padrão considerando o modelo logístico dado por:

[1] Para o marcador NHPS2.5 não foi possível calcular as estimativas *Odds Ratios* devido a casela $n_{AA}^{(1)}$ da Tabela B.1 ser igual a zero.

Tabela C.1 *Estimativas dos Odds Ratios no nível genotípico.*

Marcador	$\hat{\Psi}$	Diabetes	valor p	Hipertensão	valor p	Obesidade	valor p
B1BK	$\hat{\Psi}_{AG/AA}$	1,2176	0,3460	1,1129	0,3586	1,0642	0,5834
	$\hat{\Psi}_{GG/AA}$	0,7388	0,2470	0,7534	0,0765	0,9881	0,9401
NHPS2.5	$\hat{\Psi}_{GA/AA}$	-	-	-	-	-	-
	$\hat{\Psi}_{GG/AA}$	-	-	-	-	-	-
BAR-2(16) Cat	$\hat{\Psi}_{AG/AA}$	1,1236	0,6550	1,0218	0,8891	0,9805	0,8922
	$\hat{\Psi}_{GG/AA}$	1,0868	0,7660	0,7509	0,0787	1,3081	0,0908
BAR-2(27) Cat	$\hat{\Psi}_{GlnGlu/GlnGln}$	1,0983	0,6330	1,0125	0,9117	1,2809	**0,0247**
	$\hat{\Psi}_{GluGlu/GlnGln}$	1,5820	0,2620	1,2212	0,3300	1,4725	0,0556
ECA Cat	$\hat{\Psi}_{DI/DD}$	1,0863	0,6970	0,9827	0,8870	1,0892	0,4752
	$\hat{\Psi}_{II/DD}$	1,1181	0,7030	1,0835	0,6310	1,0248	0,8795
ANGO Cat	$\hat{\Psi}_{MT/TT}$	1,0798	0,7110	1,1614	0,2038	1,2157	0,0921
	$\hat{\Psi}_{MM/TT}$	1,1884	0,5270	1,4839	**0,0108**	1,2959	0,0820
ENOS Cat	$\hat{\Psi}_{GT/TT}$	0,9808	0,9610	1,2098	0,3930	0,7799	0,2870
	$\hat{\Psi}_{GG/TT}$	1,1063	0,7960	1,3912	0,1310	0,7768	0,2710
p22 PHOX	$\hat{\Psi}_{TC/TT}$	0,8460	0,5980	0,9294	0,6830	1,0947	0,6025
	$\hat{\Psi}_{CC/TT}$	1,0787	0,8160	0,7879	**0,0469**	1,0700	0,6989
GNB3	$\hat{\Psi}_{CT/TT}$	1,0957	0,7120	1,1601	0,2810	0,8744	0,3320
	$\hat{\Psi}_{CC/TT}$	0,9130	0,7270	1,2586	0,1250	0,8398	0,2410

$$\ln\left(\frac{p_1(\mathbf{x})}{1-p_1(\mathbf{x})}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2, \tag{C.1}$$

em que para o marcador B1BK tem-se:

$$x_1 = \begin{cases} 1, & \text{se o indivíduo carrega o genótipo } AG \\ 0, & \text{caso contrário} \end{cases}$$

$$x_2 = \begin{cases} 1, & \text{se o indivíduo carrega o genótipo } GG \\ 0, & \text{caso contrário} \end{cases}$$

Para o marcador ANGO Cat tem-se:

$$x_1 = \begin{cases} 1, & \text{se o indivíduo carrega o genótipo } MT \\ 0, & \text{caso contrário} \end{cases}$$

$$x_2 = \begin{cases} 1, & \text{se o indivíduo carrega o genótipo } MM \\ 0, & \text{caso contrário} \end{cases}$$

E assim por diante para os demais marcadores.

Tabela C.2 *Estimativas dos parâmetros (erro padrão) do modelo para Diabetes no nível genotípico.*

	Diabetes		
Marcador	Intercepto	$\hat{\beta}_1$	$\hat{\beta}_2$
B1BK	-2,3172(0,3006)	0,1969(0,2090)	-0,3028(0,2616)
BAR-2(16) Cat	-2,5938(0,3176)	0,1165(0,2608)	0,0832(0,2798)
BAR-2(27) Cat	-2,9436(0,4356)	0,0937(0,1965)	0,4587(0,4089)
ECA Cat	-2,5851(0,3195)	0,0828(0,2125)	0,1116(0,2927)
ANGO Cat	-2,6454(0,3069)	0,0768(0,2072)	0,1727(0,2729)
ENOS Cat	-2,4795(0,4172)	-0,0194(0,3967)	0,1010(0,3912)
p22 PHOX	-2,4188(0,3513)	-0,1672(0,3171)	0,0758(0,3257)
GNB3	-2,4427(0,2958)	0,0914(0,2479)	-0,0910(0,2605)

Tabela C.3 *Estimativas dos parâmetros (erro padrão) do modelo para Hipertensão no nível genotípico.*

	Hipertensão		
Marcador	Intercepto	$\hat{\beta}_1$	$\hat{\beta}_2$
B1BK	-0,4830(0,1786)	0,1069(0,1165)	-0,2832(0,1599)
BAR-2(16) Cat	-0,5674(0,1843)	0,0216(0,1547)	-0,2865(0,1630)
BAR-2(27) Cat	-0,8597(0,2217)	0,0124(0,1119)	0,1998(0,2051)
ECA Cat	-0,7298(0,1821)	-0,0175(0,1223)	0,0802(0,1672)
ANGO Cat	-0,0730(0,1741)	0,1497(0,1178)	**0,3947(0,1548)**
ENOS Cat	-0,9347(0,2345)	0,1905(0,2229)	0,3302(0,2189)
p22 PHOX	-0,2892(0,1958)	-0,0733(0,1795)	**-0,3567(0,1795)**
GNB3	-0,9072(0,1687)	0,1485(0,1378)	0,2300(0,1500)

Tabela C.4 *Estimativas dos parâmetros (erro padrão) do modelo para Obesidade no nível genotípico.*

	Obesidade		
Marcador	Intercepto	$\hat{\beta}_1$	$\hat{\beta}_2$
B1BK	-0,5307(0,1776)	0,0622(0,1134)	-0,0120(0,1599)
BAR-2(16) Cat	-0,6805(0,1781)	-0,0197(0,1453)	0,2686(0,1588)
BAR-2(27) Cat	-1,0113(0,2189)	**0,2475(0,1102)**	0,3870(0,2021)
ECA Cat	-0,5542(0,1763)	0,0854(0,1197)	0,0245(0,1614)
ANGO Cat	-0,8233(0,1683)	0,1954(0,1160)	0,2592(0,1490)
ENOS Cat	-0,2392(0,2439)	-0,2486(0,2336)	-0,2525(0,2293)
p22 PHOX	-0,5925(0,1911)	0,0905(0,1737)	0,0676(0,1749)
GNB3	-0,3163(0,1669)	-0,1342(0,1384)	-0,1746(0,1489)

Apêndice D

Programas Computacionais

D.1 Capítulo 4- Programas Computacionais

Programa **R** (http://www.R-project.org)

Dados de Câncer de Mama (Clayton, 2003).

Análise no nível genotípico.

```
#referência ProPro
rm(list=ls(all=TRUE))
genotipo<-data.frame(x1=c(1,0,0),x2=c(0,1,0),n=c(145,619,608),y=c(89,369,342))
genotipo$Ymat<-cbind(genotipo$y,genotipo$n-genotipo$y)
ajuste<-glm(Ymat~x1+x2,family=binomial,data=genotipo)
summary(ajuste)
coef(ajuste)
exp(coef(ajuste))
```

Programa **R**

Dados de Câncer de Mama (Clayton, 2003).

Análise no nível genotípico sob o modelo multiplicativo.

```
#An. gen. mult.
rm(list=ls(all=TRUE))
genomultiplic<-data.frame(x=c(2,1,0),n=c(145,619,608),y=c(89,369,342))
genomultiplic$Ymat<-cbind(genomultiplic$y,genomultiplic$n-genomultiplic$y)
ajuste<-glm(Ymat~x,family=binomial, data=genomultiplic)
summary(ajuste)
```

Programa **R**

Dados de Câncer de Mama (Clayton, 2003).

Análise no nível cromossômico.

```
#An. crom.
rm(list=ls(all=TRUE))
alelo<-data.frame(x=c(1,0),n=c(909,1835),y=c(547,1053))
alelo$Ymat<-cbind(alelo$y,alelo$n-alelo$y)
ajuste<-glm(Ymat~x,family=binomial, data=alelo)
summary(ajuste)
coef(ajuste)
exp(coef(ajuste))
```

Programa **R**

Análise para o conjunto de dados "População de Vitória".

Exemplo para análise no nível genotípico - Marcador B1BK.

```
#Marcador B1BK- Diabetes
#referência GG
rm(list=ls(all=TRUE))
genotipo<-data.frame(x1=c(1,0,0),x2=c(0,1,0),n=c(614,722,224),y=c(50,49,24))
genotipo$Ymat<-cbind(genotipo$y,genotipo$n-genotipo$y)
ajuste<-glm(Ymat~x1+x2,family=binomial,data=genotipo)
summary(ajuste)
coef(ajuste)
exp(coef(ajuste))

#Marcador B1BK- Hipertensão
#referência GG
rm(list=ls(all=TRUE))
genotipo<-data.frame(x1=c(1,0,0),x2=c(0,1,0),n=c(616,728,226),y=c(210,231,92))
genotipo$Ymat<-cbind(genotipo$y,genotipo$n-genotipo$y)
ajuste<-glm(Ymat~x1+x2,family=binomial,data=genotipo)
summary(ajuste)
coef(ajuste)
exp(coef(ajuste))

#Marcador B1BK- Obesidade
#referência GG
rm(list=ls(all=TRUE))
genotipo<-data.frame(x1=c(1,0,0),x2=c(0,1,0),n=c(615,721,226),y=c(235,265,87))
genotipo$Ymat<-cbind(genotipo$y,genotipo$n-genotipo$y)
ajuste<-glm(Ymat~x1+x2,family=binomial,data=genotipo)
summary(ajuste)
coef(ajuste)
exp(coef(ajuste))
```

Programa **R**

Análise para o conjunto de dados "População de Vitória".

Exemplo para análise no nível genotípico sob o modelo multiplicativo - Marcador B1BK.

```
#Marcador B1BK- Diabetes
rm(list=ls(all=TRUE))
genomultiplic<-data.frame(x=c(2,1,0),n=c(614,722,224),y=c(50,49,24))
genomultiplic$Ymat<-cbind(genomultiplic$y,genomultiplic$n-genomultiplic$y)
ajuste<-glm(Ymat~x,family=binomial, data=genomultiplic)
summary(ajuste)

#Marcador B1BK- Hipertensão
rm(list=ls(all=TRUE))
genomultiplic<-data.frame(x=c(2,1,0),n=c(616,728,226),y=c(210,231,92))
genomultiplic$Ymat<-cbind(genomultiplic$y,genomultiplic$n-genomultiplic$y)
ajuste<-glm(Ymat~x,family=binomial, data=genomultiplic)
summary(ajuste)

#Marcador B1BK- Obesidade
rm(list=ls(all=TRUE))
genomultiplic<-data.frame(x=c(2,1,0),n=c(615,721,226),y=c(235,265,87))
genomultiplic$Ymat<-cbind(genomultiplic$y,genomultiplic$n-genomultiplic$y)
ajuste<-glm(Ymat~x,family=binomial, data=genomultiplic)
summary(ajuste)
```

Programa **R**

Exemplo para análise no nível genotípico - Marcador B1BK.

Estatística de Cochran-Mantel-Haenszel

```
#Marcador B1BK- Diabetes
rm(list=ls(all=TRUE))
dados<-scan('c:\\B1BK2_diabet.txt',list(B1BK=0,Diabetes=0,Etnia=0),
na.string='9999')
attach(dados)
dados
fnames<-list(Diabetes=c('1:Caso','0:Controle'))
Diabetes <-factor(Diabetes)
Diabetes<-C(Diabetes,treatment)
fnames<-list(B1BK=c('11', '21','22'))
B1BK<-factor(B1BK)
B1BK<-C(B1BK,treatment)
Etnia<-factor(Etnia)
library(stats)
mantelhaen.test(Diabetes,B1BK,Etnia)

#Marcador B1BK- Hipertensao
rm(list=ls(all=TRUE))
dados<-scan('c:\\B1BK2_hipert.txt',list(B1BK=0,Hipertensao=0,Etnia=0),
na.string='9999')
attach(dados)
dados
fnames<-list(Hipertensao=c('1:Caso','0:Controle'))
Hipertensao <-factor(Hipertensao)
Hipertensao<-C(Hipertensao,treatment)
fnames<-list(B1BK=c('11', '21','22'))
B1BK<-factor(B1BK)
B1BK<-C(B1BK,treatment)
Etnia<-factor(Etnia)
library(stats)
mantelhaen.test(Hipertensao,B1BK,Etnia)
```

```
#Marcador B1BK- Obesidade
rm(list=ls(all=TRUE))
dados<-scan('c:\\B1BK2_obesid.txt',list(B1BK=0,Obesidade=0,Etnia=0),
na.string='9999')
attach(dados)
dados
fnames<-list(Obesidade=c('1:Caso','0:Controle'))
Obesidade <-factor(Obesidade)
Obesidade<-C(Obesidade,treatment)
fnames<-list(B1BK=c('11', '21','22'))
B1BK<-factor(B1BK)
B1BK<-C(B1BK,treatment)
Etnia<-factor(Etnia)
library(stats)
mantelhaen.test(Obesidade,B1BK,Etnia)
```

Programa **R**

Análise para o conjunto de dados "População de Vitória".

Exemplo para análise no nível cromossômico - Marcador B1BK.

```
#Marcador B1BK- Diabetes
rm(list=ls(all=TRUE))
alelo<-data.frame(x=c(1,0),n=c(1950,1170),y=c(149,97))
alelo$Ymat<-cbind(alelo$y,alelo$n-alelo$y)
ajuste<-glm(Ymat~x,family=binomial, data=alelo)
summary(ajuste)
coef(ajuste)
exp(coef(ajuste))

#Marcador B1BK- Hipertensão
rm(list=ls(all=TRUE))
alelo<-data.frame(x=c(1,0),n=c(1960,1180),y=c(651,415))
alelo$Ymat<-cbind(alelo$y,alelo$n-alelo$y)
ajuste<-glm(Ymat~x,family=binomial, data=alelo)
summary(ajuste)
coef(ajuste)
exp(coef(ajuste))

#Marcador B1BK- Obesidade
rm(list=ls(all=TRUE))
alelo<-data.frame(x=c(1,0),n=c(1951,1173),y=c(735,439))
alelo$Ymat<-cbind(alelo$y,alelo$n-alelo$y)
ajuste<-glm(Ymat~x,family=binomial, data=alelo)
summary(ajuste)
coef(ajuste)
exp(coef(ajuste))
```

Programa **R**

Exemplo para análise no nível cromossômico - Marcador B1BK.

Estimativa de Mantel-Haenszel.

```
#Marcador B1BK- Diabetes
rm(list=ls(all=TRUE))
dados<-scan('c:\\B1BK_diabet.txt',list(B1BK=0,Diabetes=0,Etnia=0),
na.string='9999')
attach(dados)
dados
fnames<-list(Diabetes=c('1:Caso','0:Controle'))
Diabetes <-factor(Diabetes)
fnames<-list(B1BK=c('1', '2'))
B1BK<-factor(B1BK)
Etnia<-factor(Etnia)
library(stats)
mantelhaen.test(Diabetes,B1BK,Etnia)

#Marcador B1BK- Hipertensao
rm(list=ls(all=TRUE))
dados<-scan('c:\\B1BK_hipert.txt',list(B1BK=0,Hipertensao=0,Etnia=0),
na.string='9999')
attach(dados)
dados
fnames<-list(Hipertensao=c('1:Caso','0:Controle'))
Hipertensao <-factor(Hipertensao)
fnames<-list(B1BK=c('1', '2'))
B1BK<-factor(B1BK)
Etnia<-factor(Etnia)
library(stats)
mantelhaen.test(Hipertensao,B1BK,Etnia)

#Marcador B1BK- Obesidade
rm(list=ls(all=TRUE))
dados<-scan('c:\\B1BK_obesid.txt',list(B1BK=0,Obesidade=0,Etnia=0),
na.string='9999')
```

```
attach(dados)
dados
fnames<-list(Obesidade=c('1:Caso','0:Controle'))
Obesidade <-factor(Obesidade)
fnames<-list(B1BK=c('1', '2'))
B1BK<-factor(B1BK)
Etnia<-factor(Etnia)
library(stats)
mantelhaen.test(Obesidade,B1BK,Etnia)
```

Referências Bibliográficas

[1] Agresti, A. (2002). *Categorial data analysis, 2nd Ed.* New York: Wiley.

[2] Akaike, H. (1974). A new look at statistical model identification. *IEEE Transactions on Automatic Control* **AU-19**, 716-722.

[3] Andrade, M. & Pinheiro, H.P. (2002). *Métodos estatísticos aplicados em genética humana*. 15° SINAPE, São Paulo: ABE.

[4] Barcellos, L.F.; Klitz, W.; Field, L.L.; Tobias, R.; Bowcock, A.M.; Wilson, R.; Nelson, M.P.; Nagatomi, J. & Thomson, G. (1997). Association mapping od disease loci, by use of a pooled DNA genomic sreen. *Am. J. Hum. Genet.* **61**, 734-747.

[5] Bishop, Y.M.M., Fienberg, S.E. & Holand, D.W. (1975). *Discrete multivariate analysis: theory and practice*. Cambridge: Mite Press.

[6] Boehnke, M. & Langefeld, C.D. (1998). Genetic association mapping based on discordant sib pairs: the discordant alleles test. *Am. J. Hum. Genet.* **62**, 950-961.

[7] Breslow, N.E. & Day, N.E. (1980). *Statistical methods in cancer research. Vol. I: The analysis of case control studies*. Lyon: IARC Scientific Publications **32**, International Agency for Research on Cancer.

[8] Cardon, L.R. & Palmer, L.J. (2003). Population stratification and spurious allelic association. *Lancet* **361**, 598-604.

[9] Clayton, D. (2003). Population association. In *Handbook of statistical genetics, 2nd Ed.* New York: Wiley, 939-960.

[10] Collins, A. & Morton, N.E. (1998). Mapping a disease locus by allelic association. *Proc. Natl. Acad. Sci.* **95**, 1741-1745.

[11] Devlin, B., Risch, N. & Roeder, K. (1996). Disequilibrium mapping: composite likelihood for pairwise disequilibrium. *Genomics* **36**, 1-16.

[12] Duncan, C.T., Haile, R.W. & Duggan, D. (2005). Recent development in genome-wide association scans: a workshop summary and review. *Am. J. Hum. Genet.* **77**, 337-345.

[13] Dunning, A.M., Chiano, M., Smith, N.R., Dearden, J., Gore, M., Oakes, S., Wilson, C., Stratton, M., Peto,J., Easton, D., Clayton, D. & Ponder, B.A.J. (1997). Common BRCA1 variants and susceptibility to breast and ovarian cancer in the general population. *Human Molecular Genetics* **6**, 285-289.

[14] Ewens, W.J. & Spielman, R.S. (1995). The transmission/disequilibrium test: history, subdivision, and admixture. *Am. J. Hum. Genet.* **57**, 455-464.

[15] Ewens, W.J. & Spielman, R.S. (2003). The transmission/disequilibrium test. In *Handbook of statistical genetics, 2nd Ed.* New York: Wiley, 961-972.

[16] Falconer, D.S. & Mackay, T.F.C. (1996). *Introduction to quantitative genetics.* London: Prentice Hall.

[17] Farah, S.B. (1997). *DNA segredos e mistérios.* São Paulo: Sarvier.

[18] Felsenstein, J. (1979). A mathematically tractable family of genetics mapping functions with different amounts of interference. *Genetics* **91**, 769-775.

[19] Fleiss, J.L., Levin, B. & Paik, M.C. (2003). *Statistical methods for rates and proportions, 3rd ed.* New York: Wiley.

[20] Grizzle, J.E., Starmer, C.F. & Koch, G.G. (1969). Analysis of categorial data by linear models. *Biometrics* **25**, 489-504.

[21] Hall, D.B., Woolson, R.F., Clarke, W.R. & Jones, M.F. (2000). Cochran-Mantel-Haenszel techniques: aplications involving epidemiology survey data. In *Handbook of Statistics, Volume 18: Bio-environmental and Public Health Statistics* Eds. P.K. Sen and C.R. Rao. Amsterdam: North Holland, 483-500.

[22] Hill, W.G. & Robertson, A. (1968). Linkage disequilibrium in finite populations. *Theor. Appl. Genet.* **38**, 226-231.

[23] IBGE (2002). *Censo demográfico 2000*. Disponível em: <http://www.ibge.gov.br /home/presidencia/noticias/20122002censo.shtm>, acesso em janeiro 2006.

[24] Jorde, L.B. (1995). Linkage disequilibrium as a gene-mapping tool. *Am. J. Hum. Genet.* **56**, 11-14.

[25] Jorde, L.B. (2000). Linkage disequilibrium and the search for complex disease genes. *Genome Res.* **10**, 1435-1444.

[26] Kerstann, K.F., Feingold, E., Freeman, S.B., Bean, L.J.H., Pyatt, R., Tinker, S., Jewel, A.H., Capone, G. & Sherman, S.L. (2004). Linkage disequilibrium mapping in trisomic populations: analytical approaches and an application to congenital heart defects in Down Syndrome. *Genet. Epidemiol.* **27**, 240-251.

[27] Lander, E.S. & Schork, N.J. (1994). Genetic dissection of complex traits. *Science* **46**, 2037-2048.

[28] Landis, J.R., Heyman, E.R. & Koch, G.G. (1978). Average partial association in three-way contingency tables: a review and discussion of alternative tests. *International Statistical Review* **46**, 237-254.

[29] Lange, K. (1997). *Mathematical and statistical methods for genetic analysis*. New York: Springer.

[30] Lazzeroni, L.C. (1998). Linkage disequilibrium and gene mapping: an empirical least-squares approach. *Am. J. Hum. Genet.* **62**, 159-170.

[31] Lazzeroni, L.C. & Lange, K. (1998). A conditional inference framework for extending the transmission/disequilibrium test. *Hum. Hered.* **48**, 67-81.

[32] Leon, J.M. (2002). *Mapeamento genético em populações humanas via modelos de componentes de variância*. Dissertação de mestrado, São Paulo: IME/USP.

[33] Lewontin, R.C. (1964). The interation of selection and linkage. I General considerations, heterotic models. *Genetics* **49**, 49-67.

[34] Liu, B.H. (1998). *Statistical genomics: linkage, mapping and QTL analysis*. New York: CRC Press.

[35] Long, A.D. & Langley, C.H. (1999). The power of association sudies to detect the contribuition of candidate genetic loci to variation in complex traits. *Genome Res.* **9**, 720-731.

[36] Marchini, J., Cardon, L.R., Phillips, M.S. & Donnelly P. (2004). The effects of human population structure on large genetic association studies. *Nature Genetics* **35**, 512-517.

[37] Mather, K. (1951). *The measurement of linkage in hereditary*. London: Methuen.

[38] McCullagh, P. & Nelder, J.A. (1989). *Generalized linear models*. London: Chapman and Hall.

[39] Morton, N.E. (1955). Sequential tests for the detection of linkage. *Am. J. Hum. Genet.* **7**, 277-318.

[40] Morton, N.E. & Collins, A. (1998). Tests and estimates of allelic association in complex inheritance. *Proc. Natl. Acad. Sci.* USA **95**, 11389-11393.

[41] Newman, S.C. (2001). *Bioestatistical methods in epidemiology*. New York: Wiley.

[42] Ott, J. (1991). *Analysis of human genetic linkage*. London: Johns Hopkins.

[43] Paula, G.A. (2004). *Modelos de regressão com apoio computacional*. São Paulo: IME-USP.

[44] Paulino, C.D.M. & Singer, J.M. (2004). *Análise de dados categorizados*. Lisboa: Versão parcial preliminar.

[45] Prentice, R.L. & Pyke, R. (1972). Logistic disease incidence models and case control studies. *Biometrika* **66**, 403-411.

[46] Pritchard, J.K. & Donnelly, P. (2001). Case-control of association in strutured or admixed populations. *Theor. Popul. Biol.* **60**, 227-237.

[47] Pritchard, J.K. & Przeworski, M. (2001). Linkage disequilibrium in humans: models and data. *Am. J. Hum. Genet.* **69**, 1-14.

[48] R Development Core Team (2005). *R: A language and environment for statistical computing*. Austria: R Foundation for Statistical Computing, URL http://www.R-project.org.

[49] Rabinowitz, D. (1997). A transmission disequilibrium test for quantitative trait loci. *Hum. Hered.* **47**, 342-350.

[50] Rao, D.C. & Gu Chi (2001). False positives and false negatives in genome scans. In *Genetic dissection of complex traits*. London: Academic Press, 487-498.

[51] Schlesselman, J.J. (1982). *Case-control studies*. New York: Oxford.

[52] Schork. N.J.; Fallin, D.; Thiel, B.; Xu, X.; Broeckel, U; Jacob, H. & Cohen, D. (2001). The future of genetic case-control studies. In *Genetic dissection of complex traits*. London: Academic Press, 191-212.

[53] Schork. N.J. (2002). Power calculations for genetic association studies using estimated probability distributions. *Am. J. Hum. Genet.* **70**, 1480-1489.

[54] Schuster, I. & Cruz, C. D. (2004). *Estatística genômica*. Viçosa: UFV.

[55] Sham, P.C. & Curtis, D. (1995). And extended transmission/disequilibrium test (TDT) for multi-allele marker loci. *Ann. Hum. Genet.* **59**, 323-336.

[56] Sham, P.C. (1998). *Statistics in human genetics*. New York: Arnold.

[57] Spielman, R.S., McGinnis, R.E. & Ewens, W.J. (1993). Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am. J. Hum. Genet.* **52**, 506-516.

[58] Spielman, R.S., McGinnis, R.E. & Ewens, W.J. (1994). The transmission/disequilibrium test detects cosegregation and linkage. *Am. J. Hum. Genet.* **54**, 559-560.

[59] Spielman, R.S. & Ewens, W.J. (1998). A sibship test for linkage in the presense of association: the sib transmission/disequilibrium test. *Am. J. Hum. Genet.* **62**, 450-458.

[60] Sturt, E. (1976). A mapping function for human chomossomes. *Ann. Hum. Genet.* **40**, 147-163.

[61] Tavares, H.R., Andrade, D.F. & Pereira, C.A.B. (2004). Detection of determinant genes and diagnostic via item response theory. *Genetics and Molecular Biology* **27**, 679-685.

[62] Terwilliger, J.D. & Ott, J. (1992). A haplotype-based haplotype relative risk statistic. *Hum. Hered.* **42**, 337-346 .

[63] Terwilliger, J.D. & Ott, J. (1994). *Handbook of human genetic linkage*. London: Johns Hopkins.

[64] Weir, B.S. (1996). *Genetic data analysis II*. Sunderland: Sinauer Associates.

[65] Weir, B.S., Hill, W.G. & Cardon, L.R. (2004). Allelic association patterns for a dense SNP map. *Genet. Epidemiol.* **27**, 442-450.