**Estudos de simetria
na associação genética
usando dados de trios**

Maria Jacqueline Batista

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

Programa: Estatística

Orientadora: Profa. Dra. Júlia Maria Pavan Soler

Durante o desenvolvimento deste trabalho a autora recebeu apoio financeiro da
CAPES e FAPESP processo N° 06/53612-0

São Paulo, dezembro de 2011

Estudos de simetria na associação genética usando dados de trios

Esta tese contém as correções e alterações sugeridas pela Comissão Julgadora durante a defesa realizada por Maria Jacqueline Batista em 02/12/2011.

O original encontra-se disponível no Instituto de Matemática e Estatística da Universidade de São Paulo.

Comissão Julgadora:

- Profa. Júlia Maria Pavan Soler (Orientadora) - IME/USP
- Prof. Carlos Alberto de Bragança Pereira - IME/USP
- Prof. Dalton Francisco de Andrade - INE/CTC/UFSC
- Profa. Clarice Garcia Borges Demétrio - ESALQ/USP
- Profa. Suely Ruiz Giolo - DEST/UFPR

“Aleluia!

Louvai, ó servos do Senhor, louvai o nome do Senhor.

Bendito seja o nome do Senhor, agora e para sempre.

Desde o nascer ao pôr-do-sol, seja louvado o nome do Senhor.

*O Senhor é excelso sobre todos os povos,
sua glória ultrapassa a altura dos céus.”*

Salmos 112, 1-4

*“Jesus olhou para eles e disse:
aos homens isto é impossível,
mas a Deus tudo é possível.”*

Mateus 19,26

*“O coração do homem
dispõe o seu caminho,
mas é o Senhor que
dirige seus passos.”*

Provérbios 16,9

*“Louvai o Senhor,
porque Ele é bom,
cantai à glória de Seu nome,
porque Ele é amável.”*

Salmos 134,3

“Este é o dia que o Senhor fez: seja para nós dia de alegria e de felicidade.

Senhor, dai-nos a salvação; dai-nos a prosperidade, ó Senhor!

Bendito seja o que vem em nome do Senhor!

Da casa do Senhor nós vos bendizemos.

O Senhor é nosso Deus, ele fez brilhar sobre nós a sua luz.

Organizai uma festa com profusão de coroas. E cheguem até os ângulos do altar.

Sois o meu Deus, venho agradecer-vos. Venho glorificar-vos, sois o meu Deus.

Dai graças ao Senhor porque ele é bom, eterna é sua misericórdia.”

Salmos 117, 24-29

*A Deus,
minha filha Ana Yasmin, meu querido marido Juvêncio,
meus pais: Otacílio e Socorro, meu irmão Júlio,
e minha avozinha querida Ana M. Batista (In memoriam).*

Agradecimentos

Agradeço,

A Deus Todo Poderoso, pela saúde e oportunidade. À Mãe Imaculada pelas graças alcançadas.

À minha família, alicerce de tudo: minha mãe Socorro, meu pai Otacílio e meu irmão Júlio. À minha avozinha: Ana, por toda dedicação e amor (muitas saudades). Amo vocês.

Ao meu amado marido Juvêncio, por tudo: paciência, amor, carinho, conselhos... E por neste doutorado ter nascido nossa filha Ana Yasmin, amo muito vocês dois! Meu nego, saiba que a nossa família é a maior alegria da minha vida. Porque família é tudo.

A D. Gracilene, pela força, e por ficar com minha filha enquanto eu tinha que viajar para resolver as pendências desta tese.

À minha orientadora, profa. Júlia Maria, sou muito agradecida a ela, não somente por ter me guiado neste tema e dado energia na orientação deste trabalho, mas também por ser uma amiga em todos os momentos desta trajetória, foi muito bom conhecê-la e aprender com ela.

Aos professores do IME-USP, em especial, Julio Singer, Elisabeti Kira e Antônio Carlos e os do DEMA-UFC, em especial Ana Maria, Maurício Mota, André Shiguemoto, Júlio Barros, Rosa Mota, Sílvia Freitas, João Welliandre e Ronald Nojosa (agradeço em especial as palavras de incentivo e implementação computacional que foi de GRANDE ajuda, agradeço também a sua esposa Francilene pelo apoio), e também as meninas da secretaria, Margeri e Luisa.

Aos meus amigos do IME-USP, em especial, a Michelli e Horácio, Tatiana e Alessandro, Patrícia e Raydonal, Lane e Marcelo, Luz Marina, Rafael, Michel, Caio, Alexandre, Gleiciane, Tatiana, Gérman e Lizandra.

À Núbia que me ajudou em todas as fases deste trabalho, com palavras, hospedagem, programas computacionais, ou seja, ela faz parte deste trabalho, muito obrigada!

Ao Laboratório de Genética e Cardiologia Molecular do Instituto do Coração da Faculdade de Medicina da Universidade de São Paulo (InCor-USP), pelos dados reais, em especial aos Drs. José Eduardo Krieger e Alexandre C. Pereira.

À banca examinadora, prof. Carlos Alberto de Bragança Pereira, prof. Dalton Francisco de Andrade, profa. Clarice Garcia Borges Demétrio e em especial a profa. Suely Giolo, pelo apoio e INCENTIVO no decorrer deste trabalho.

À FAPESP e CAPES pelo auxílio financeiro.

Não dá para citar todos os nomes que merecem agradecimentos, pois são muitas pessoas, mas saibam que todas estão no meu coração ♡.

Resumo

Estudos de simetria na associação genética usando dados de trios

O grande desafio da Epidemiologia Genética, atualmente, é identificar, em um espaço de variáveis preditoras de alta dimensão e esparsas, fatores de risco genéticos para doenças complexas. Um delineamento amostral útil nestes estudos é coletar dados de trios, que são pequenos núcleos familiares (pai e mãe, livres da doença, e filho afetado) e, em cada indivíduo, obter dados do genótipo de marcadores moleculares, sendo a plataforma de marcadores do tipo SNPs (do inglês, *Single Nucleotide Polymorphism*), com cerca de 1 milhão de variáveis preditoras genéticas, a mais adotada. Neste trabalho é proposto um procedimento em múltiplos estágios para identificar SNPs associados com a doença em dados de trios. A primeira etapa do procedimento é baseada em uma série de análises unilocos (para cada variável preditora), usando um teste de simetria em tabelas de contingência 2×2 (conhecido, em Genética, como teste TDT, do inglês, *Transmission Disequilibrium Test*). Em um segundo estágio da análise, os resultados destes testes são usados para construir uma estatística de somas acumuladas padronizadas (CUSUM) que permite a seleção de conjuntos de SNPs (isto é, conjuntos de variáveis preditoras), possivelmente associados com a doença. Como um terceiro passo da análise, nas regiões selecionadas no passo dois, são realizadas análises de simetria via testes exatos considerando tabelas 2×2 e 4×4 (pares de SNPs). A formulação do TDT em termos de testes de simetria é uma inovação na área de Genética e facilita a extensão do caso uniloco para o multilocos. A contribuição deste trabalho reside ainda na formulação exata do teste que é útil em situações de amostras pequenas que ocorrem com frequência em dados de trios. Neste caso inferências parciais foram realizadas a partir de decomposições apropriadas da função de verossimilhança. A modelagem do problema em termos do modelo logístico permitiu concluir que não é necessário corrigir a associação para o efeito de covariáveis avaliadas nos pais. O procedimento é implementado usando recursos dos aplicativos PLINK e R. A aplicação é realizada utilizando dados de 71 trios da população brasileira, em que os indivíduos caso (filhos) foram definidos em termos da ocorrência de uma cardiopatia e, em cada um dos 213 indivíduos, estão disponíveis dados genéticos de uma plataforma de SNPs.

Palavras-chave: Mapeamento genético, Análise multilocos, Estudos de simetria, Dados de trios, Teste TDT.

Abstract

Symmetry studies in the genetic association using data from trios

Currently, the great challenge of Genetic Epidemiology is to identify, in a high dimensional and sparse space of predictor variables, genetic risk factors for complex diseases. A useful sampling design in these studies is to collect data from trios, which are small nuclear families (father and mother, free from disease, and affected child), and obtain genotypic information from each individual. The molecular markers platform most commonly used for this purpose is of SNPs (Single Nucleotide Polymorphisms), with about 1 million genetic predictor variables. This work proposes a multi-stage procedure to identify SNPs associated with disease using data from trios. The first step of the procedure is based on a series of single locus analysis (for each predictor variable) using a test for symmetry in 2×2 contingency tables (known in genetics as TDT (Transmission Disequilibrium Test)). In a second stage of the analysis, the results of these tests are used to construct a standard statistic of the cumulative sums (CUSUM), which allows the selection of sets of adjacent SNPs (ie, sets of predictor variables), possibly associated with the disease. As a third step of the analysis, in the regions selected in step two, are performed an extended analysis of symmetry considering 4×4 contingency tables. The TDT formulation in terms of symmetry tests is an innovation in the genetics area and facilitates the extension of the single locus analysis to the multiloci case. The contribution of this work lies in the exact formulation of the symmetry test for square contingency tables that is useful in situations of small sample sizes that often occur in data from trios. In this case, partial inferences were performed from appropriate decompositions of the likelihood function. The structural modeling of the problem in terms of logistic model allowed us to conclude that there is no need to adjust the association for data from parents, but only for the effect of covariates evaluated in each parental haplotype. The procedure is implemented using resources of the R statistical environment and Plink. The application is performed using real data from 71 trios of the Southeast Brazilian population, in which affected child was defined in terms of the occurrence of one congenital heart disease, and in each of the 213 individuals, genomic data were collected using Affymetrix SNP 6.0 platform.

Keywords: Genetic mapping, Multiloci analysis, Symmetry studies, Data trios, Test TDT.

Índice

Agradecimentos	vii
Resumo	ix
Abstract	x
Lista de Tabelas	xiii
Lista de Figuras	xiv
1 Introdução	1
1.1 O contexto genético	1
1.2 Motivação	6
1.3 Proposta de trabalho	8
2 Delineamento com Trios em Genômica	10
2.1 Estrutura de delineamentos com trios	11
2.2 Teste de desequilíbrio de transmissão (TDT)	12
2.2.1 Risco relativo do haplótipo no nível genotípico	12
2.2.2 Risco relativo do haplótipo no nível cromossômico	14
2.3 Marcadores moleculares - SNPs	15
3 Teste TDT - Um Estudo de Simetria	19
3.1 Caso Uniloco	19
3.1.1 Teste de McNemar	20
3.1.2 Teste exato - Tabelas 2×2	22
3.1.3 Modelo logístico - Tabelas 2×2	27

3.2	Caso Multiloco - Análise Intervalar	29
3.2.1	TDT generalizado	29
3.2.2	Teste exato - Tabelas 4×4	34
3.2.3	Modelo logístico - Tabelas 4×4	37
4	Seleção de Regiões Candidatas	40
4.1	Métodos de seleção de regiões candidatas	40
4.1.1	Método de alto escore	41
4.1.2	Método CUSUM	43
4.1.3	Procedimento multiestágios	45
5	Aplicação	47
6	Considerações Finais	61
A	Genética - Conceitos Básicos e Revisão	65
A.1	Equilíbrio de Hardy-Weinberg	66
A.2	Análise de ligação	67
A.3	Desequilíbrio de ligação	68
B	Demonstrações de Algumas Expressões	71
C	Rotinas Computacionais	77
	Referências Bibliográficas	94

Lista de Tabelas

1.1	Ilustração dos estudos com Trios e SNPs.	7
2.1	Transmissão de alelos - nível genotípico.	13
2.2	Transmissão de alelos - nível genotípico (amostra pareada).	14
2.3	Transmissão de alelos - nível haplotípico.	15
2.4	Transmissão de alelos - nível haplotípico (amostra pareada).	15
3.1	Transmissão de alelos - amostra pareada.	20
3.2	Transmissão de haplótipos - Amostra pareada.	31
5.1	Ilustração dos estudos com trios.	48
5.2	Número de SNPs por cromossomo (dados de trios).	49
5.3	Número de SNPs analisados (dados de trios).	49
5.4	Valores p do TDT e teste exato em tabelas 2×2	56
5.5	Valores p do teste exato e TDT generalizado em tabelas 4×4	60
A.1	Relação entre probabilidades dos alelos e genótipos.	67

Lista de Figuras

1.1	Amostra de trios.	7
2.1	Amostra de n trios.	13
2.2	Ilustração de SNPs.	17
3.1	Ilustração de mapeamento intervalar considerando dois locos.	30
3.2	Composição dos alelos em dois locos.	31
3.3	Composição das subtabelas considerando dois locos.	32
5.1	Estatística do TDT para os 22 cromossomos.	50
5.2	CUSUM para o cromossomo 1.	51
5.3	CUSUM para os cromossomos 1, 2, 3, 4, 5 e 10.	52
5.4	CUSUM para os cromossomos 17, 18, 19 e 22.	53
5.5	CUSUM para os cromossomos 7, 12 e 14.	53
5.6	CUSUM para os cromossomos 8 e 15.	53
5.7	CUSUM para o cromossomo 16.	54
5.8	CUSUM para os cromossomos 9, 11, 13 e 21.	54
5.9	CUSUM para os cromossomos 6 e 20.	55
5.10	Valores-p dos testes TDT e exato em tabelas 2×2	57
5.11	Tabelas uniloco e para pares de locos considerando os SNPs 96 e 97.	59
A.1	Ilustração do Equilíbrio de Hardy-Weinberg.	66
A.2	Composição dos alelos em dois locos genéticos.	69
A.3	Decréscimo do desequilíbrio de ligação por gerações.	70

Capítulo 1

Introdução

1.1 O contexto genético

A Epidemiologia Genética tem contribuído com os estudos que visam identificar ou mapear genes associados a fatores que causam doenças, ou seja, estudos cuja finalidade é investigar a existência de associação entre um fator de risco genético e uma doença.

Com o acelerado avanço das pesquisas em Biologia Molecular dos últimos anos, novas tecnologias de mapeamento de genes têm sido desenvolvidas (Altshuler et al. 2008; Ziegler et al. 2008). Uma das principais contribuições à análise genética tem sido a possibilidade de amostrar o genoma humano, e de várias outras espécies, por meio de mapas de marcadores moleculares, cada vez maiores e mais densos, permitindo uma cobertura amostral mais representativa do genoma (Devlin et al. 2003; Conti and Gauderman, 2004; Duncan et al. 2005; Millstein et al. 2006; Huang et al. 2007).

Neste contexto, os principais mapas de marcadores moleculares são os do tipo microsatélites e SNPs (do inglês, *Single Nucleotide Polymorphism*). Microsatélites, também conhecidos como SSR (do inglês, *Simple Sequence Repeats*), são denominados de repetições de sequências simples (Pritchard and Feldman, 1996). Estes locos variam no número de repetições de uma simples sequência de DNA, sendo comumente

utilizados na análise de populações naturais, por serem altamente polimórficos (multialélicos, isto é, com muitas classes de resposta genotípica possível), além de serem úteis para entender a estrutura genética de uma população (Slatkin, 1995). Com o avanço da tecnologia, surgiram grandes plataformas do tipo SNPs que por sua vez são polimorfismos de um único nucleotídeo que ocorrem na população e são marcadores moleculares nos estudos genômicos. Cada SNP tem quatro possíveis alelos, da forma A, C, G e T, na prática são apresentados com dois alelos, por exemplo, como CT. Em particular os SNPs foram introduzidos a partir do International HapMap Project (2003).

Nos estudos de associação entre fatores de risco genéticos e doença (em geral, categorizada de forma dicotômica), existem diferentes alternativas de coleta de dados, como os delineamentos observacionais (transversal, prospectivo e retrospectivo) que podem ou não incorporar informação familiar (Ott, 1991; Duncan et al. 2005). Os estudos caso-controle (retrospectivos), são os mais comumente adotados e sua análise estatística é feita, em geral, por meio de modelos de regressão logística (Clayton, 2003). No caso uniloco, isto é, a análise de associação genética considerando um único marcador, Batista (2006) apresenta diferentes aplicações deste modelo, enfatizando as vantagens e limitações das análises genotípicas e cromossômicas. A utilização de análises no nível cromossômico, dado que os dados foram coletados no nível de indivíduos (genótipos), envolve o diagnóstico criterioso de hipóteses que são assumidas na re-estruturação dos dados, as quais correspondem a população infinita, casamentos aleatórios e genes em EHW- Equilíbrio de Hardy-Weinberg (Sasieni, 1997; Gianola and Sorensen, 2002).

Nos delineamentos caso-controle, 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 formado por indivíduos com e sem uma determinada doença de interesse como diabetes, hipertensão, entre outras. Uma das preocupações neste tipo de delineamento é com a obtenção dos dados amostrais, pois a partir de como estes são obtidos pode existir um efeito de confundimento gerado pelo desbalanceamento na constituição genômica dos indivíduos caso e controle, conduzindo a falsas conclusões resultantes de testes de hipóteses de interesse. Várias propostas para contornar este tipo de efeito existem na literatura. Batista et al. (2008) e Giolo et al. (2011) comparam modelos logísticos com e sem covariáveis genéticas e concluem que para dados caso-controle são necessários tais ajustamentos para controlar o efeito de estratificação. Outras alternativas

consideram delineamentos experimentais mais robustos para este efeito, por exemplo, situações em que 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. Neste sentido, o delineamento com trios (pequeno núcleo familiar: pai e mãe, livres da doença, e filho afetado) representa uma alternativa de controle do efeito de confundimento. Neste tipo de delineamento, em geral, o filho tem uma doença rara que se manifesta na infância, sendo que os pais destas não apresentam a doença em nenhuma fase da vida.

Spielman et al. (1993) introduziu para os dados de trios um teste conhecido como TDT (do inglês, *Transmission Disequilibrium Test*), neste caso a doença avaliada é um tipo de diabetes. A proposta deste teste é averiguar associação entre locos de marcadores moleculares e genes que influenciam a suscetibilidade de doença. Esta análise de associação genética considera um único marcador molecular por vez e equivale à análise de associação em tabelas de contingência 2×2 para dados pareados. O teste TDT neste caso nada mais é do que o bem conhecido teste de McNemar (ver, por exemplo, Agresti, 2002 e Paulino e Singer, 2006). Com o avanço nos estudos de marcadores, grandes plataformas de SNPs foram criadas e o desafio que se impôs foi o de encontrar regiões de SNPs associados com a doença a partir da avaliação de um espaço de variáveis preditoras de alta dimensão (cerca de 1 milhão de SNPs) e esparsos (muitos SNPs de efeito nulo).

Os primeiros trabalhos consideram a aplicação do TDT repetida e independentemente milhares de vezes, o que resultou em problemas de múltiplos testes, isto é, na necessidade de correções nos níveis descritivos (valor p) das estatísticas de associação. Lazzeroni and Lange (1998) propõem uma correção do tipo Bonferroni para valores p correspondentes às análises de uma sequência de SNPs adjacentes, associados entre si e associados com a doença. Sabe-se também que o efeito individual do SNP é pequeno (Spielman et al., 1993, Horvath and Baur, 2000), devido ao pouco desequilíbrio de ligação¹ entre cada SNP e genes associados com a doença, o que pode não acontecer quando mais de um SNP são avaliados na associação com a doença. Desta maneira, análises uniloco que consideram o efeito de cada SNP por

¹ Desequilíbrio de ligação: associação probabilística entre locos. Em genética é implicitamente usada nos estudos de associação entre loco genético (“gene”) e doença. Quando os locos estão em associação é mais fácil identificar uma região genômica candidata com base na amostragem de um único loco da região.

vez ignoram a informação adicional que a dependência entre estes marcadores pode trazer para o estudo de associação de genes com a doença (Conti and Gauderman, 2004).

Estes estudos de associação genética realizados em larga escala genômica são conhecidos como GWAS (do inglês, *Genome-Wide Associations*), isto é, estudos que pesquisam muitos locos do genoma simultaneamente, e têm ganhado grande atenção de epidemiologistas nos últimos anos principalmente devido ao acelerado crescimento e disponibilidade de mapas densos de SNPs (Altshuler et al. 2008). Apesar do sucesso que estes estudos têm trazido na identificação de genes (Duncan et al. 2005; Amos et al. 2008), ainda existem muitos desafios analíticos e de interpretação funcional/causal dos resultados a serem pesquisados. Neste cenário, os problemas que mais caracterizam a análise de dados genômicos são: (i) efeitos de confundimento devido à estratificação genética da população, (ii) efeito individual do SNP pequeno, (iii) alta dimensionalidade do espaço das variáveis preditoras (SNPs) a serem avaliadas e, (iv) tamanho amostral pequeno para a ocorrência das classes genotípicas de SNPs na amostra. Soluções para cada item destes serão tratadas neste trabalho. Assim, o uso de delineamentos com trios resolve o problema (i), sendo, em particular, útil para doenças genéticas que se manifestam precocemente em filhos (afetados) e pais livres da doença. Estudos que englobam dados de trios em genética e aplicam o teste TDT para avaliar a associação de marcadores moleculares do tipo SNP, podem ser vistos, por exemplo, em Bergen et al. (2003) e Sykes et al. (2009).

Relativamente aos problemas (ii) e (iii), a literatura tem buscado alternativas de redução de dimensionalidade do espaço de variáveis preditoras, como análises multilocos que simultaneamente testam um conjunto de locos por vez (Ziegler et al. 2008). Aschard et al. (2007) e Guedj et al. (2006), considerando estudos de associação caso-controle propõem a pré-seleção de “regiões genômicas”, isto é, de uma sequência de SNPs com algum sinal significativo, com base em uma estatística de somas acumuladas de qui-quadrados. A combinação de variáveis preditoras feita por estes autores leva em conta a medida de associação que cada uma tem com a doença, explorando o espaço todo das variáveis (SNPs individuais), independentemente de estarem correlacionadas ou não. Contudo, a maioria dos métodos de redução de dimensionalidade em genética realizam inicialmente um estudo do padrão de “correlação” entre os locos de SNPs, por meio de testes de desequilíbrio de ligação entre pares de locos. Isto equivale a estudar a dependência/associação entre variáveis preditoras (neste caso, categóricas e dispostas em uma sequência) e,

nas regiões ou janelas de dependência, realizar uma análise multilocos que considera o efeito das variáveis da região simultaneamente. Clayton and Jones (1999) e Zhao et al. (2000) considerando estudos com trios realizam inicialmente a pré-seleção de SNPs em associação e nas regiões ou janelas de dependência aplicam uma extensão do teste TDT baseado na hipótese de homogeneidade das marginais em tabelas de contingência quadradas.

As dificuldades na formulação do teste TDT para múltiplos locos são como combinar os dados de cada loco e como explorar a informação dos dados combinados que, em geral, ficam dispostos no formato de tabelas de contingência quadradas. A combinação dos dados dos genótipos de vários locos, em geral, é feita por definir os haplótipos correspondentes. Esta definição pode envolver ambigüidades que têm sido resolvidas por se adotar uma específica configuração (*cis* ou *trans*, ver por exemplo, Conti and Gauderman, 2004) ou por estimar a configuração ótima construindo uma verossimilhança que considera todas as combinações possíveis (Clayton and Jones, 1999; Zhao et al., 2000; Mاتيoli, 2002). A análise das tabelas de contingência geradas de dados multilocos tem sido feita por meio de testes qui-quadrado clássicos ou testes de homogeneidade das marginais que podem ser restritivos demais para testar modelos genéticos de interesse.

Neste contexto, os problemas (ii) e (iii) serão abordados neste trabalho primeiramente por reduzir o conjunto de locos cromossômicos a ser pesquisado usando uma estatística de somas acumuladas de qui-quadrados adaptada de Aschard et al. (2007) e Guedj et al. (2006) para o caso de dados de trios. Nas regiões de efeito de associação significativa uma análise multilocos é realizada considerando pares de SNPs por vez e o teste de hipóteses de simetria. Finalmente o problema (iv) tem recebido pouca atenção da literatura e, neste sentido, é apresentada uma formulação de testes exatos úteis para o estudo de hipóteses de simetria em dados com tamanhos amostrais pequenos.

1.2 Motivação

A população brasileira tem um histórico de grande miscigenação, o qual, certamente, é uma das preocupações para grupos de pesquisa nacionais interessados em realizar estudos genéticos. Esta miscigenação conta com aproximadamente 20 gerações de casamentos entre três grupos ascendentes (nativos, negros e europeus) tornando a população brasileira uma das mais heterogêneas do mundo (Pena et al., 2011). Esta estrutura genética da população pode interferir nos padrões de associação entre os locos genômicos (causam desequilíbrio de ligação entre locos), o que leva à inexistência de associação ou a “baixa” associação entre os marcadores (SNPs) e doenças. Também, a coleta de dados é dificultada, pois há muita chance de confundimento nos estudos do tipo caso-controle.

Iniciativas nestes tópicos de pesquisa estão sendo consideradas no Laboratório de Genética e Cardiologia Molecular do Instituto do Coração da Faculdade de Medicina da Universidade de São Paulo (InCor-USP), como é o caso dos dados do projeto TRIOS DO BRASIL, cujas amostras são baseadas em trios (Figura 1.1). Neste caso, tem-se 71 trios, ou seja, 213 indivíduos, amostrados do banco de dados do InCor, cujo filho(a) apresenta uma cardiopatia congênita da qual os pais estão livres. Vale ressaltar, que este tipo de delineamento é bastante útil quando o indivíduo afetado tem uma doença rara e precoce, o que viabiliza ter “facilidade” na coleta dos dados dos pais. Em todos estes indivíduos foram obtidas amostras de sangue, das quais o DNA foi extraído e avaliado por meio da plataforma de SNPs da *Affymetrics 6.0*², que consiste de 1 milhão de marcadores moleculares distribuídos ao longo do genoma. O objetivo do estudo é identificar SNPs ou regiões de SNPs associados com a síndrome cardíaca. Estes dados serão usados neste trabalho como motivação e para ilustrarem as análises propostas.

Veja, como ilustração na Tabela 1.1, as variáveis que compõem a estrutura familiar dos trios que são: TRIO- número do trio, ID- identificação do indivíduo, FA- pai e MO- mãe. Por exemplo, no trio 189, a mãe do indivíduo 130 é 129 e o pai é o 131. As variáveis fenotípicas são: SEX- sexo do indivíduo (1- masculino, 2- feminino) e AFFECT (1- não afetado, 2- afetado, por uma doença). As variáveis genotípicas são os genótipos dos marcadores do tipo SNP, em que, nesta referida tabela, têm-se 900.000 SNPs, com os seus respectivos genótipos, por exemplo, para o SNP2, os indivíduos podem ter os genótipos: TT, CT

² Tecnologia da empresa *Affymetrics*, a qual comercializa sistemas, em que um dos objetivos é melhor compreender o papel que os genes têm em doenças.

e CC. Estes marcadores são as variáveis preditoras e a resposta é o indivíduo ter ou não a doença, tendo a seguinte codificação (0, 1 e 2), em que, 0: indivíduos homozigotos para o alelo de maior frequência no SNP, 1: correspondendo a indivíduos heterozigotos e 2: correspondendo a indivíduos homozigotos para o alelo de menor frequência no SNP.

Tabela 1.1 *Ilustração dos estudos com Trios e SNPs.*

TRIOS	ID	FA	MO	SEX	AFFECT	SNP1	SNP2	...	SNP900.000
189	131	0	0	1	1	TT	TT	...	AA
189	129	0	0	2	1	GT	CT	...	AG
189	130	131	129	1	2	GT	CT	...	AA
191	262	0	0	1	1	GT	TT	...	AA
191	261	0	0	2	1	GG	CT	...	AA
191	263	262	261	1	2	GG	TT	...	AA
192	374	0	0	1	1	TT	CT	...	AG
192	373	0	0	2	1	GT	CC	...	GG
192	372	374	373	2	2	TT	CC	...	GG
193	421	0	0	1	1	GT	TT	...	GG
193	420	0	0	2	1	TT	CT	...	AG
193	419	421	420	2	2	TT	TT	...	GG
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
282	4097	0	0	1	1	TT	CC	...	AG
282	4096	0	0	2	1	GG	CT	...	AA
282	4095	4097	4096	2	2	GT	CC	...	AA

Figura 1.1 *Amostra de trios.*

1.3 Proposta de trabalho

Aliado a todas as dificuldades em termos da adaptação de uma linguagem estatística para a estruturação destes dados, desde a leitura dos mesmos até a formulação das correspondentes tabelas de contingência por eles geradas, e tendo em vista os problemas de análise caracterizados na seção 1.1, tem-se a motivação para o presente trabalho. Como objetivo geral propõe-se uma estratégia de análise de associação em múltiplos estágios na busca por regiões genômicas associadas com a doença, considerando dados de trios e plataformas de SNPs. O estudo de associação neste tipo de dados é um dos problemas alvo da genética epidemiológica na atualidade.

Primeiramente, temos os dados de trios em que os indivíduos têm a mesma oportunidade de estarem expostos ao fator de risco genético o que contorna o efeito de confundimento presente nestes estudos. Deste modo é realizada uma análise de associação uniloco, “SNP por SNP”, percorrendo todos os locos do mapa de marcadores. A estatística de associação usada neste caso é a TDT, em que tabelas de contingência 2×2 são construídas. Em particular, este teste é formulado como um teste de simetria para facilitar sua extensão para o caso multilocos. Ainda, é feita a fatoração da função de verossimilhança envolvida na modelagem dos dados, sendo obtido um teste exato de associação que pode ser aplicado em SNPs com poucas informações disponíveis. Isto evita que tais SNPs sejam desprezados da análise.

Com base nesta primeira fase da análise, o perfil da estatística de associação uniloco é avaliada por uma estatística de somas acumuladas padronizadas: CUSUM (do inglês, *Cumulative Sum*), conhecida como Carta de Controle de Somas Acumuladas. De encontro ao problema do efeito individual do SNP ser pequeno e para a redução do número de testes envolvidos, esta estatística permite capturar a associação genética presente em regiões ou janelas de SNPs cujos efeitos individuais combinados tornam-se significantes. O procedimento CUSUM foi proposto por Page (1954) para detectar falhas e monitorar a variabilidade de um determinado processo (Yi et al. 2006; Montgomery, 2008; Correa et al. 2009). No problema genômico tratado neste trabalho o CUSUM é utilizado para monitorar a variabilidade das estatísticas de associação uniloco, o que pode ser útil na detecção de regiões candidatas.

Nas regiões selecionadas pelo CUSUM, procede-se com as análises multilocos ou, mais especificamente, para pares de locos, o que corresponde a análises de tabelas 4×4 . Além do teste TDT generalizado, neste caso, são formulados os testes de simetria contextualizados para o problema genético. A versão de

testes exatos é apresentada a qual engloba uma fatoração conveniente do modelo multinomial. No caso das tabelas 4×4 propõe-se também uma forma mais útil de uso dos graus de liberdade envolvidos neste tipo de análise.

Tanto para o caso uniloco (tabelas 2×2) como para o intervalar (tabelas 4×4), como alternativa de análise são introduzidos os modelos logísticos. A aplicação do procedimento é realizada utilizando os dados de trios do InCor descritos anteriormente.

Este trabalho está organizado da seguinte forma: no Capítulo 2 é descrito o delineamento com trios, descrevendo a estrutura dos dados, citando exemplos de doenças modeladas por esta construção. O teste TDT, também, é introduzido neste capítulo além de uma seção referente aos marcadores SNPs. O TDT visto como um teste de simetria é abordado no Capítulo 3, junto com as propostas dos testes exatos e o modelo logístico em tabelas 2×2 (caso uniloco) e também 4×4 (caso intervalar). Os métodos de seleção de regiões candidatas incluindo o CUSUM e a descrição do procedimento proposto são descritos no Capítulo 4. Com o intuito de explorar um conjunto de dados genéticos para verificar se há evidência de SNPs (locos genéticos) ou blocos de SNPs associados à doença, no Capítulo 5, é apresentada a descrição do banco de dados reais e os resultados das análises ilustrando a metodologia proposta. A implementação do procedimento proposto é realizada com recursos computacionais dos aplicativos PLINK (Purcell et al. 2005) que é um aplicativo muito utilizado em Genética, e do R (<http://www.r-project.org>). O Capítulo 6 traz a discussão dos resultados obtidos com as análises dos dados, conclusões e contribuições desta tese, sendo também apresentadas algumas direções para pesquisas futuras. No Apêndice A, são destacados alguns conceitos em genética que são de interesse para o desenvolvimento deste trabalho. No Apêndice B, são apresentadas as demonstrações de algumas expressões chave, como, por exemplo, a fatoração da distribuição multinomial utilizada na formulação dos testes exatos. Finalmente, no Apêndice C, são apresentadas algumas rotinas computacionais implementadas no aplicativo R para a execução das análises envolvidas e as regiões candidatas selecionadas pelo método CUSUM para os 22 cromossomos considerando os dados reais analisados.

Capítulo 2

Delineamento com Trios em Genômica

Nos delineamentos do tipo trios (pai e mãe, livres da doença, e filho afetado), em cada indivíduo obtêm-se os dados do genótipo de marcadores moleculares, por exemplo, SNPs. Neste tipo de formulação, tem-se que o indivíduo afetado é jovem, o que possibilita ter as informações dos pais, e a doença a ser analisada é, em geral, rara. Algumas doenças têm sido modeladas, usando-se esta estrutura de dados, por exemplo, alguns tipos de diabetes (Spielman et al., 1993), doença arterial coronariana (Watkins, 2004), doença de inflamação no intestino (Pender et al., 2004). Para a proposta de análise de trios e plataformas SNPs, algumas doenças têm sido consideradas, por exemplo, anorexia nervosa (Bergen et al., 2003) e autismo (Sykes et al., 2009).

Em estudos de associação genética os delineamentos caso-controle são os mais usados, contudo, uma das preocupações é se as amostras para os grupos de *caso* e *controle* são coletadas de populações genéticas homogêneas. Como é difícil garantir a obtenção de tais amostras, pois o genótipo do indivíduo não é observável diretamente fazendo com que a coleta deste tipo de dados seja vulnerável aos efeitos de confundimento devido à estratificação ou estrutura genética da população, os delineamentos com trios constituem uma alternativa poderosa. Neste tipo de estudo casos, os filhos, e controles, os pais, são

homogêneos na constituição genética geral, carregam a mesma ancestralidade e, portanto, conduzem a amostras de casos e controles balanceadas ou homogêneas.

Para a análise deste tipo de dados, neste capítulo, serão apresentadas as tabelas de contingência que podem ser construídas da leitura dos dados em questão. Além disso, introduziremos o teste de associação clássico usado na análise destes dados. Na Seção 2.3 apresentaremos os marcadores moleculares do tipo SNP.

2.1 Estrutura de delineamentos com trios

Considere um estudo de associação genética em que a doença sob estudo é rara e se manifesta na infância, sendo que os pais de tais crianças afetadas não apresentam a doença em nenhuma fase da vida, isto é, são considerados livres da doença. Situações deste tipo caracterizam várias cardiopatias, alguns tipos de diabetes, como comentado anteriormente.

Para o estudo de doenças com esta natureza os delineamentos com trios são recomendados para a coleta de dados. A idéia geral 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, com o intuito de amostrar *casos* e *controles* da mesma população genética (o núcleo familiar trio) e avaliar o risco relativo de genes para a doença. Considere um marcador molecular sob estudo, possivelmente associado com a doença. Para este marcador, 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, têm-se as amostras de casos e de controles de uma mesma população genética, isto é, pareadas.

Considere a Figura 1.1. Para a construção das tabelas de contingência primeiro coletam-se os genótipos dos três indivíduos. Pode-se tratar os dados no nível de alelos/haplótipo (uma das partes do material genético carregado pelo indivíduo) ou no nível de genótipo (informação conjunta das duas partes genéticas carregadas pelo indivíduo), que são análises correspondentes a tamanhos amostrais n , $2n$. Estas abordagens são discutidas, por exemplo, em Sasieni (1997). Na construção das tabelas de dados de trios podem-se ter estudos pareados ou não pareados, como será apresentado, posteriormente, e são observados um conjunto de muitas variáveis preditoras genéticas, neste caso, os SNPs.

2.2 Teste de desequilíbrio de transmissão (TDT)

O teste de desequilíbrio de transmissão (TDT) é uma ferramenta comumente adotada para a análise de associação genética em delineamentos com trios. Como citado anteriormente, o TDT não é afetado pelo efeito de variáveis de confundimento, como no caso das estatísticas de associação em estudos caso-controle, devido a populações heterogêneas (miscigenadas, por exemplo) que podem induzir a evidências falso-positivas (Ewens and Spielman, 2003). Para o TDT os dados amostrais são considerados balanceados geneticamente, para casos e controles, isto é, possuem o mesmo “*background*” genético¹, pois neste teste as amostras são baseadas em trios, em que os pais correspondem à amostra *controle* e os filhos correspondem à amostra *caso*.

Na análise desses dados, algumas soluções são propostas como, por exemplo, as estatísticas de associação baseadas no Risco Relativo do Haplótipo no Nível Genotípico (*GHRR*), propostos por Rubinstein et al. (1981), e os sugeridos por Terwilliger and Ott (1992, 1994) e Spielman et al. (1993) baseados no Risco Relativo do Haplótipo no Nível Cromossômico (*HHRR*).

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

2.2.1 Risco relativo do haplótipo no nível genotípico

Avaliando o risco relativo do haplótipo no nível genotípico, denotado do inglês por *GHRR*, considere um marcador (SNP, por exemplo) 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*”. Como ilustração considere o trio 1 indicado na Figura 2.1. 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 *HH* transmitidos) e os dois alelos paternos que não foram transmitidos para o filho são considerados na amostra “*controle*” (alelos não transmitidos), neste caso, os alelos *G* e *J*.

Para a construção da Tabela 2.1, considere novamente a Figura 2.1. Observa-se que o trio 1 contribuirá

¹ *Background* genético: conjunto de muitos genes.

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 .

Figura 2.1 *Amostra de n trios.*

O teste de associação genética adotado, neste caso, é formulado como um teste qui-quadrado clássico de homogeneidade, definido em termos das frequências dos alelos transmitidos (caso) e não transmitidos (controle) (Tabela 2.1). A estatística do teste é dada por:

$$\chi^2 = \frac{2n(WZ - XY)^2}{(W + X)(W + Y)(X + Z)(Y + Z)}, \quad (2.1)$$

em que n representa o número total de trios. Sob a hipótese $H_0 : P(D|\text{transmitiu}H) = P(D|\text{não transmitiu}H)$, ou seja, sob a hipótese de não existência de associação entre o fator de risco genético e a doença D , a estatística (2.1), segue assintoticamente uma distribuição qui-quadrado com 1 grau de liberdade, $\chi^2_{(1)}$.

Tabela 2.1 *Transmissão de alelos - nível genotípico.*

	H	$\bar{H}$	Total
Transmitido	W	X	$W+X$
Não Transmitido	Y	Z	$Y+Z$
	$W+Y$	$X+Z$	$2n$

Seguindo essa abordagem genotípica mas, alternativamente, considerando dados pareados (Tabela 2.2), cada trio contribuirá com uma única observação e será classificado em termos de genótipos transmitidos e não transmitidos. Sob essa leitura dos dados, pela Figura 2.1 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 .

Tabela 2.2 *Transmissão de alelos - nível genotípico (amostra pareada).*

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

Neste caso, a estatística do teste de associação é conhecida na área de Genética como estatística TDT (na análise genotípica, com tamanho amostral n) e é dada por:

$$\text{TDT} = \frac{(B - C)^2}{(B + C)}, \quad (2.2)$$

em que (2.2), sob a hipótese de não associação, segue assintoticamente uma distribuição qui-quadrado com 1 grau de liberdade. A estatística (2.2) é a tradicional estatística do teste de McNemar (veja, por exemplo, Sham, 1998; Agresti, 2002; Paulino e Singer, 2006).

Note que, as duas abordagens (2.1) e (2.2) são genotípicas com as construções das tabelas baseadas em tamanhos de amostras diferentes $2n$ (dados não pareados) e n (dados pareados), respectivamente.

2.2.2 Risco relativo do haplótipo no nível cromossômico

Avaliando agora o risco relativo do haplótipo no nível cromossômico, denotado do inglês por $HHRR$, considera-se para a Tabela 2.3 um total amostral de $4n$ (Terwilliger and Ott, 1992; Lange, 1997), ou seja, os alelos transmitidos e não transmitidos de cada um dos pais (supostamente independentes) fornecem quatro observações por família.

Considerando novamente a Figura 2.1 e a notação disposta na Tabela 2.3, o trio 1 contribuirá com duas observações na casela w e duas na casela z , as quatro observações do trio 2 serão adicionadas 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 mesmos referidos anteriormente para a Tabela 2.1.

Para este caso, mas sob uma formulação de dados pareados, na definição do teste de associação TDT considere n_{ij} definido como o número de trios em que os pais transmitem o alelo i e não transmitem o alelo j . Neste caso, os dados podem ser dispostos esquematicamente no formato da Tabela 2.4.

Tabela 2.3 *Transmissão de alelos - nível haplotípico.*

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

Tabela 2.4 *Transmissão de alelos - nível haplotípico (amostra pareada).*

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

Para a Figura 2.1 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} .

Comparando estas diferentes leituras dos dados de trios para serem dispostos em tabelas de contingência, Terwilliger and Ott (1992) mostram que a abordagem haplotípica $HHRR$ é mais poderosa que a genotípica $GHRR$ e Terwilliger and Ott (1994), adicionalmente, indicam que a análise considerando dados pareados tem maior poder que a análise considerando amostras independentes e tem a vantagem de usar o teste de McNemar que não precisa assumir EHW - Equilíbrio de Hardy Weinberg. A abordagem do teste TDT pareado (haplotípico) é mais viável e poderosa e será a considerada neste trabalho.

Spielman et al. (1993) propuseram analisar os dados dispostos no formato da Tabela 2.4 via a estatística TDT que corresponde ao teste de McNemar, como descrito anteriormente.

2.3 Marcadores moleculares - SNPs

Antes dos avanços alcançados em biotecnologia, os estudos de associação genética consideravam plataformas ou mapas de marcadores moleculares compostos de algumas dezenas de nucleotídeos chamados mi-

microsatélites. Este tipo de marcador consiste na identificação de regiões do DNA onde ocorre a repetição de pequenos conjuntos de bases em longas sequências (Pritchard and Feldman, 1996). O tamanho físico dos microsatélites permite que a amostragem do genoma seja feita por grandes pedaços. Este tipo de marcador é reconhecidamente útil para análises que envolvem dados de grandes famílias ou pedigrees. Com o avanço das técnicas de sequenciamento, foram identificadas regiões do genoma onde longas sequências diferem entre os indivíduos em apenas um nucleotídeo. O nome dado a estas regiões ou a este tipo de marcador do genoma é SNP (do inglês, *Single Nucleotide Polymorphism*), ou polimorfismo de um único nucleotídeo. Em particular, este tipo de mapa ou plataforma genômica foi introduzida e disponibilizada pelo International HapMap Project (2003), um consórcio entre grandes centros de pesquisa que se uniram para finalidade de descrever os padrões comuns de variação genética humana. Este projeto é um recurso fundamental para os pesquisadores em mapeamento de genes que buscam encontrar variantes genéticas que afetam a saúde pública, ver por exemplo,

http://en.wikipedia.org/wiki/International_HapMap_Project.

Outros estudos em que a abordagem de SNPs é usada podem ser vistos em Ambrosius et al. (2004), Aulchenko et al. (2007), Batista et al. (2008), McCarthy et al. (2008), Allen and Satten (2009), Yang et al. (2010) e Nielsen et al. (2011).

Os SNPs são polimorfismos (variações) de um único nucleotídeo que ocorrem na população e são utilizados como marcadores em estudos genômicos que informam sobre a localização de genes, em geral, posicionados na sua vizinhança, que estão associados com a doença de interesse. São considerados muito pouco polimórficos, isto é, assumem somente três classes genotípicas, por exemplo, para um certo loco de SNP, tem-se as possíveis categorias *AA*, *Aa* e *aa*, que podem ser a seguinte codificação 0 para indivíduos homozigotos para o alelo de maior frequência na população, 1 correspondendo a indivíduos heterozigotos e 2 correspondendo a indivíduos homozigotos para alelos de menor frequência, respectivamente.

Acredita-se que existam em torno de 10 a 13 milhões desses marcadores ou mudanças no genoma humano. Nestes marcadores, o que acontece é uma diferença na sequência de DNA nas bases A (adenina), T (timina), C (citosina) e G (guanina), por exemplo, na sequência *AAGCCTA* pode-se ter a sequência alterada para *AAGCTTA*. Como ilustração deste tipo de marcador segue na Figura 2.2 duas moléculas de DNA, em que a primeira difere da segunda em um único nucleotídeo (polimorfismo *C/T*). Esta Figura

está disponível no site:

http://en.wikipedia.org/wiki/Single-nucleotide_polymorphism.

Figura 2.2 *Ilustração de SNPs.*

Da análise de mapas moleculares do tipo SNP, é conhecido (Horvath and Baur, 2000) que o efeito individual do loco de SNP no controle da doença é, em geral, pequeno, devido ao baixo nível de associação ou desequilíbrio de ligação entre cada loco de SNP e os genes causais da doença. Outros problemas que ocorrem ao analisar SNPs individualmente (conhecida como, análise uniloco) é que múltiplos testes, em geral, são realizados simultaneamente, o que aumenta a ocorrência de associações falso-positivas. Logo, uma estratégia de análise destes dados tem sido mensurar o efeito de “regiões de SNPs” ou “janelas de SNPs”, definidas pela combinação de locos adjacentes, como o procedimento proposto por Guedj et al. (2006). Uma outra estratégia que tem sido explorada na análise de tais mapas é capturar a estrutura de haplótipos de SNPs, como explorado por Conti and Gauderman (2004), por exemplo.

Com o objetivo de comparação, Papachristou and Lin (2006) estudaram os marcadores microsatélites e os SNPs sob vários aspectos, como a sua utilidade na análise de ligação², cuja finalidade é a localização do gene da doença. Lembrando que microsatélites são marcadores que consistem de variações em grandes pedaços do DNA, menos densos no genoma, usados em dados de famílias estendidas e os SNPs, por sua

² Na análise de ligação estuda-se os eventos de recombinação entre dois locos cromossômicos, sejam eles genes, marcadores moleculares, aberrações cromossômicas, etc.

vez, são variações em uma única base, bastante densos no genoma e, preferencialmente usados em estudos observacionais ou que envolvem núcleos familiares pequenos, como trios. De maneira geral, os autores concluem que o uso de mapas com SNPs são mais precisos do que o uso de microsátélites para finalidade de mapeamento de genes com amostras de indivíduos independentes (sem grau de parentesco) ou que pertencem a núcleos familiares muito pequenos, como os dados de trios.

Capítulo 3

Teste TDT - Um Estudo de Simetria

A análise de dados de trios em Epidemiologia Genética atualmente, em geral, é realizada em larga escala genômica no sentido de ser necessário avaliar o efeito de um número muito grande de locos genômicos (fatores de risco) sobre a doença. Tais locos fazem parte de um mapa de marcadores moleculares que correspondem a uma amostra do genoma, como foi introduzido na Seção 2.2. A análise dessas variáveis genômicas pode ser feita via procedimentos uniloco ou multilocos (uma alternativa sendo o biloco ou intervalar), as quais serão tratadas neste Capítulo. Em ambos os contextos de análise dos dados de trios, testes TDT baseados na estatística qui-quadrado de McNemar são apresentados e formulados como testes de simetria. Neste Capítulo, apresentamos também alternativas exatas de construção de testes de associação para a análise de dados de trios, as quais são úteis para situações de tamanhos amostrais pequenos, o que é comum na genotipagem de SNPs. Para as situações multilocos, a estatística de associação é decomposta em componentes ortogonais, o que permite testar associações específicas de maior interesse ao estudo de fatores de risco genético e doença. Finalmente, sob as duas abordagens, uniloco e multilocos, um modelo de regressão logístico é apresentado em estudos de associação em dados de trios.

3.1 Caso Uniloco

A análise uniloco consiste em avaliar individualmente cada loco (variável preditora genômica) de um mapa de marcadores por vez, ou seja, realizar testes de associação independentemente para cada loco

em busca de um possível loco candidato a fator de risco para a doença. A seguir, é apresentado o teste¹ (assintótico) de associação comumente utilizado e é introduzido sua formulação como um teste exato.

3.1.1 Teste de McNemar

Considerando as versões do TDT vistas no Capítulo anterior, a proposta mais viável e mais utilizada (o modelo mais aceito) é a formulação de dados pareados como descrito na Tabela 3.1. Neste caso, o mesmo indivíduo é avaliado nas duas situações de Transmitido (T) e Não-transmitido (NT) e a resposta a ser avaliada em cada situação é a ocorrência dos alelos, digamos A e a , em que a unidade de pareamento é o indivíduo.

Como comentado anteriormente, para a formulação do TDT considere n_{ij} definido como o número de trios em que os pais transmitem o alelo i e não transmitem o alelo j e p_{ij} a probabilidade do alelo i ser transmitido e o alelo j ser não transmitido e π_{ij} a probabilidade de pais transmitirem o alelo i e não transmitirem o alelo j .

Tabela 3.1 *Transmissão de alelos - amostra pareada.*

Transmitidos	Não Transmitidos		Total
	A	a	
A	n_{11}	n_{12}	$n_{1.}$
a	n_{21}	n_{22}	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	$2n$

Spielman et al. (1993) propuseram analisar os dados dispostos no formato da Tabela 3.1 via a estatística TDT. Note que, em uma tabela de contingência 2×2 com dados pareados, o teste qui-quadrado de Pearson correspondente ao teste da hipótese $H_S : \pi_{12} = \pi_{21}$ se reduz ao conhecido teste de McNemar (ver, por exemplo, Paulino e Singer, 2006).

Considerando os dados da Tabela 3.1, o teste TDT é dado pela estatística:

$$\text{TDT} = \frac{(n_{12} - n_{21})^2}{(n_{12} + n_{21})}, \quad (3.1)$$

que sob a hipótese H_S de não associação, segue uma distribuição assintótica qui-quadrado com 1 grau de

¹ O teste de McNemar será abordado novamente pelo fato deste ser usado na proposta deste trabalho.

liberdade. Já a estatística de Pearson é dada por:

$$Q^2 = \sum_{i=1}^I \sum_{j=1}^J \frac{(O_{ij} - E_{ij})^2}{E_{ij}}, \quad (3.2)$$

com $i = 1, 2$ representando o índice de linha e $j = 1, 2$ representando o índice de coluna em tabelas 2×2 , O_{ij} a frequência observada da categoria ij e E_{ij} a correspondente frequência esperada sob a hipótese $H_S : \pi_{12} = \pi_{21}$. Logo, tem-se que a estatística de Pearson é dada por:

$$\begin{aligned} Q^2 &= \frac{(n_{11} - n_{11})^2}{n_{11}} + \frac{(n_{21} - \frac{n_{21}+n_{12}}{2})^2}{\frac{n_{21}+n_{12}}{2}} + \frac{(n_{12} - \frac{n_{21}+n_{12}}{2})^2}{\frac{n_{21}+n_{12}}{2}} + \frac{(n_{22} - n_{22})^2}{n_{22}} = \\ &= \frac{(\frac{n_{21}}{2} - \frac{n_{12}}{2})}{\frac{n_{21}+n_{12}}{2}} + \frac{(\frac{n_{12}}{2} - \frac{n_{21}}{2})}{\frac{n_{21}+n_{12}}{2}} = \frac{1}{2} \frac{(n_{21} - n_{12})^2}{n_{21} + n_{12}} + \frac{1}{2} \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}} = \frac{(n_{12} - n_{21})^2}{(n_{12} + n_{21})}, \end{aligned}$$

que coincide com a estatística TDT. Assim a estatística TDT usualmente adotada na área da genética é a estatística de McNemar para um teste de simetria em tabelas 2×2 com amostras pareadas, e é equivalentemente a estatística de Pearson. Ainda, no caso de tabelas 2×2 , estes testes equivalem a testar a hipótese de homogeneidade das marginais, isto é, $H_H: \pi_{1.} = \pi_{.1}$ (Paulino e Singer, 2006). Clayton and Jones (1999) e Zhao et al. (2000) reconhecem o teste TDT como um teste de homogeneidade das marginais, o qual é estendido para casos mais gerais de tabelas $r \times r$. A estatística TDT (McNemar) será aplicada neste trabalho independentemente considerando os dados de cada marcador, procedimento este denominado análise uniloco.

3.1.2 Teste exato - Tabelas 2×2

Não é recomendado usar a estatística TDT quando se tem frequências esperadas das caselas n_{12} e n_{21} menores ou iguais a 5, o que é frequente em dados de trios, que envolvem um número relativamente pequeno de trios e devido a dificuldades na genotipagem dos SNPs. Há ainda o problema de que para muitos SNPs pode-se ter trios não informativos. Como exemplo de dados de trios não informativos, suponha que tem-se pai e mãe homozigotos, com genótipo, AA e AA , respectivamente, o filho (afetado) com certeza será AA , ou seja, o material que é transmitido é igual ao que não é transmitido dos pais para o filho, ou seja, são trios que contribuem com as frequências das caselas da diagonal principal n_{11} e n_{22} da tabela, que são estatísticas ancilares. Para situações deste tipo testes exatos tornam-se uma ferramenta muito necessária.

Primeiramente, considere π_{ij} como a probabilidade de pais transmitirem o alelo i e não transmitirem o alelo j . Então, no caso de uma tabela 2×2 com dados pareados como na Tabela 3.1, a hipótese H_S : $\pi_{12} = \pi_{21}$, corresponde ao bem conhecido “teste de simetria” (Agresti, 2002; Paulino e Singer, 2006).

Considere uma tabela de contingência I^2 , em que I é o número de níveis de ambas as variáveis que definem as linhas e colunas, gerada por um modelo multinomial $M_{I^2-1}(n_{..}, \boldsymbol{\pi})$, em que $\boldsymbol{\pi} = (\pi_{ij})_{i \leq 1, j \leq I}$ é o vetor de parâmetros satisfazendo $\boldsymbol{\pi}^\top \mathbf{1} = \sum_{i,j} \pi_{ij} = 1$, e $n_{..}$ é o vetor de frequências observadas com $\sum_{ij} n_{ij} = n_{..}$.

Para dados dispostos como no formato da Tabela 3.1, no caso $I=2$, a distribuição conjunta (multinomial) é dada por:

$$P(\mathbf{n}|n_{..}, \boldsymbol{\pi}) = \frac{n!}{n_{11}!n_{12}!n_{21}!n_{22}!} \pi_{11}^{n_{11}} \pi_{12}^{n_{12}} \pi_{21}^{n_{21}} \pi_{22}^{n_{22}}, \quad (3.3)$$

em que $\mathbf{n} = (n_{11}, n_{12}, n_{21}, n_{22})^\top$, $\boldsymbol{\pi} = (\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22})^\top$ e $\mathbf{n} \sim \text{Mult}(n_{..}, \boldsymbol{\pi})$.

Note que o valor $n_{..}$ representa $2n$ observações, no caso de delineamentos com trios (Tabela 3.1), pois ao fazer o procedimento de leitura dos dados duplica-se o tamanho amostral de trios, pois cada pai e mãe contribuem com uma informação na tabela. Esta formulação mais geral de tabelas quadradas I^2 em delineamentos com trios é adequada para se estender o procedimento e estudar a associação devida a multilocos, como veremos na próxima seção. A vantagem desta formulação é que existem modelos

apropriados disponíveis na literatura e que podem ser explorados para esse tipo de análise, como os modelos de simetria ou os modelos de homogeneidade das marginais.

Considerando a Tabela 3.1, sob a validade do modelo Multinomial para a descrição das frequências observadas, a hipótese de simetria ou de não associação, isto é, $H_S : \pi_{12} = \pi_{21}$, para $I = 2$ pode ser formulada como um teste de contrastes

$$H_S : \mathbf{C}^\top \boldsymbol{\pi} = 0, \quad (3.4)$$

em que, $\mathbf{C}^\top = (0 \ -1 \ 1 \ 0)$ e $\boldsymbol{\pi} = (\pi_{11} \ \pi_{12} \ \pi_{21} \ \pi_{22})^\top$.

Com o objetivo de obter testes específicos sobre os parâmetros de interesse (π_{12} e π_{21}) independentemente dos demais parâmetros envolvidos, passa-se a considerar a fatoração da verossimilhança obtida do modelo multinomial (Tabela 3.1) por meio de condicionamento em uma marginal. A idéia é reduzir o modelo completo descartando as parcelas que dependem somente dos parâmetros de perturbação. Aqui consideraremos o método da redução da função de verossimilhança de forma análoga à utilizada para obter inferências parciais, cuja origem pode ser traçada nos trabalhos de Fisher (ver, por exemplo, Basu, 1975, 1977, 1979; Pereira, 1980). Para detalhes, com respeito a este e outros métodos de redução de modelos na presença de parâmetros de perturbação, veja, por exemplo, Farias et al. (2009).

Para o problema de associação, tal fatoração é facilitada pelas “boas” propriedades da distribuição multinomial (a distribuição marginal da soma de componentes multinomiais é também multinomial; a distribuição condicional de um subconjunto de componentes multinomiais, dado o vetor observado da soma destes componentes, é também multinomial). Este resultado, bastante conhecido, é útil quando temos interesse na distribuição dos totais marginais de uma tabela de contingência, sendo a distribuição dos componentes internos multinomial. Assim, estamos em condições de fatorar a distribuição multinomial (3.3) em partições que também possuem distribuição multinomial.

Primeiramente considere a verossimilhança completa,

$$L(\boldsymbol{\pi}) = P(\mathbf{n}|n_{..}, \boldsymbol{\pi}) = P(n_{11}, n_{12}, n_{21}|n_{..}, \boldsymbol{\pi}). \quad (3.5)$$

Agora, seja a seguinte fatoração da função de verossimilhança completa (3.5) nas variáveis n_{11} , n_{12} e $n_{12} + n_{21}$:

$$L(\boldsymbol{\pi}) = P(\mathbf{n}|n_{..}, \boldsymbol{\pi}) = P(n_{11}, n_{12}, n_{12} + n_{21}|n_{..}, \boldsymbol{\pi})$$

$$\begin{aligned}
&= P(n_{12} + n_{21}|n_{..}, \boldsymbol{\pi})P(n_{11}, n_{12}|n_{..}, n_{12} + n_{21}, \boldsymbol{\pi}) \\
&= P(n^*|n_{..}, \boldsymbol{\pi})P(n_{11}, n_{12}|n_{..}, n^*, \boldsymbol{\pi}) = P(n^*|n_{..}, \boldsymbol{\pi})P(n_{11}|n^*, \boldsymbol{\pi})P(n_{12}|n_{..}, n^*, n_{11}, \boldsymbol{\pi}) \\
&= P(n_{12} + n_{21}|n_{..}, \boldsymbol{\pi})P(n_{11}|(n_{12} + n_{21}), \boldsymbol{\pi})P(n_{12}|n_{..}, n_{12} + n_{21}, n_{11}, \boldsymbol{\pi})
\end{aligned}$$

Em que $n^* = n_{12} + n_{21}$, seja,

$$L_1(\boldsymbol{\pi}_{12} + \boldsymbol{\pi}_{21}) = P(n_{12} + n_{21}|n_{..}, \boldsymbol{\pi}), \quad (3.6)$$

$$L_2(\boldsymbol{\pi}_{11}, \boldsymbol{\pi}_{12}, \boldsymbol{\pi}_{21}) = P(n_{11}|(n_{12} + n_{21}), \boldsymbol{\pi}), \quad (3.7)$$

$$L_3(\boldsymbol{\pi}_{12}, \boldsymbol{\pi}_{21}) = P(n_{12}|n_{..}, n_{12} + n_{21}, n_{11}, \boldsymbol{\pi}). \quad (3.8)$$

Calculando as distribuições dos respectivos termos em (3.6), (3.7) e (3.8), separadamente, tem-se:

$n_{12} + n_{21}|n_{..}, \pi \sim \text{Bin}(n_{..} - (n_{12} + n_{21}), \pi_{12} + \pi_{21})$, referente a (3.6).

Dado que,

$$(n_{11}, n_{12} + n_{21}, n_{22}) \sim \text{Mult}(n_{..}, \pi_{11}, \pi_{12} + \pi_{21}, \pi_{22})$$

Tem-se,

$$P(n_{11}|(n_{12} + n_{21})) = P(n_{11} = x|(n_{12} + n_{21}) = y) = \frac{P(n_{11}=x, n_{12}+n_{21}=y)}{P(n_{12}+n_{21}=y)}$$

$$\begin{aligned}
&= \frac{\frac{n!}{x!y!(n-x-y)!} \pi_{11}^x (\pi_{12} + \pi_{21})^y (1 - \pi_{11} - \pi_{12} - \pi_{21})^{n-(x+y)}}{\frac{n!}{y!(n-y)!} (\pi_{12} + \pi_{21})^y (1 - \pi_{12} - \pi_{21})^{n-y}} \\
&= \frac{(n-y)!}{x!(n-x-y)!} \frac{\pi_{11}^x (1 - \pi_{11} - \pi_{12} - \pi_{21})^{n-x-y}}{(1 - \pi_{12} - \pi_{21})^{n-y}} \\
&= \frac{(n-y)!}{x!((n-y)-x)!} \frac{\pi_{11}^x (1 - \pi_{11} - \pi_{12} - \pi_{21})^{(n-y)-x}}{(1 - \pi_{12} - \pi_{21})^{(n-y)+x-x}} \\
&= \binom{n-y}{x} \left(\frac{\pi_{11}}{1 - \pi_{12} - \pi_{21}} \right)^x \left(1 - \frac{\pi_{11}}{1 - \pi_{12} - \pi_{21}} \right)^{(n-y)-x}.
\end{aligned}$$

Portanto, para o termo (3.7) tem-se

$$n_{11}|(n_{12} + n_{21}) \sim \text{Bin} \left(n_{..} - (n_{12} + n_{21}), \frac{\pi_{11}}{1 - \pi_{12} - \pi_{21}} \right).$$

Ainda, para o termo (3.8),

$$P(n_{12}|n_{12} + n_{21}, n_{11}) = P(n_{12} = x|n_{12} + n_{21} = y, n_{11} = z)$$

$$\begin{aligned} &= \frac{P(n_{12}=x, n_{12}+n_{21}=y, n_{11}=z)}{P(n_{12}+n_{21}=y, n_{11}=z)} = \frac{P(n_{12}=x, n_{12}=y-x, n_{11}=z)}{P(n_{12}+n_{21}=y, n_{11}=z)} \\ &= \frac{\frac{n!}{x!(y-x)!z!(n-x-y+x-z)!} \pi_{11}^z \pi_{12}^x \pi_{21}^{y-x} (1-\pi_{11}-\pi_{12}-\pi_{21})^{n-x-y+x-z}}{\frac{n!}{y!z!(n-y-z)!} (\pi_{12}+\pi_{21})^y \pi_{11}^z (1-\pi_{11}-\pi_{12}-\pi_{21})^{n-y-z}} \\ &= \frac{y!}{x!(y-x)!} \frac{\pi_{12}^x \pi_{21}^{y-x}}{(\pi_{12}+\pi_{21})^{y+x-x}} = \binom{y}{x} \left(\frac{\pi_{12}}{\pi_{12}+\pi_{21}} \right)^x \left(1 - \frac{\pi_{12}}{\pi_{12}+\pi_{21}} \right)^{y-x}. \end{aligned}$$

Implicando que,

$$n_{12}|(n_{12} + n_{21}), n_{11}, \pi \sim \text{Bin} \left(n_{12} + n_{21}, \frac{\pi_{12}}{\pi_{12} + \pi_{21}} \right).$$

Logo, a função de verossimilhança completa (3.5) fica convenientemente fatorada como:

$$\begin{array}{ccccc}
 P(n_{12}, n_{21}, n_{1.} | n_{..}, \boldsymbol{\pi}) & = & P(n_{11}, n_{12}, n_{21} + n_{21} | n_{..}, \boldsymbol{\pi}) & & \\
 & = & P(n_{12} + n_{21} | n_{..}, \boldsymbol{\pi}) & P(n_{11} | (n_{21} + n_{21}), \boldsymbol{\pi}) & P(n_{12} | (n_{21} + n_{21}), n_{11}, \boldsymbol{\pi}) \\
 & \downarrow & & \downarrow & \downarrow \\
 \text{Bin}(n_{..} - (n_{12} + n_{21}), \pi_{12} + \pi_{21}) & \text{Bin}\left(n_{..} - (n_{12} + n_{21}), \frac{\pi_{11}}{1 - \pi_{12} - \pi_{21}}\right) & \text{Bin}\left(n_{12} + n_{21}, \frac{\pi_{12}}{\pi_{12} + \pi_{21}}\right)
 \end{array}$$

Note que, o único termo da decomposição anterior que depende somente dos parâmetros de interesse π_{12} e π_{21} é L_3 , sendo que os demais termos dependem somente de parâmetros *nuisance*. Por conseguinte, pode-se utilizar a correspondente distribuição $\text{Bin}\left(n_{12} + n_{21}, \frac{\pi_{12}}{\pi_{12} + \pi_{21}}\right)$ como base para a construção de um teste exato da hipótese de interesse $H_S : \pi_{12} = \pi_{21}$. Perceba que sob H_S a distribuição condicional da variável n_{12} dado $n_{12} + n_{21}$ segue um modelo Binomial da forma $\text{Bin}\left(n_{12} + n_{21}, \frac{1}{2}\right)$. A redução do modelo completo, envolvendo todos os parâmetros do espaço paramétrico, para o modelo condicional obtido, envolvendo apenas os parâmetros de interesse, ocorreu sem qualquer perda de informação para as inferências sobre a hipótese de simetria. Deste modo, podemos utilizar a estatística que iremos denotar por TE para definir um teste exato (bicaudal) da hipótese de interesse, em que o nível descritivo é dado por,

$$\begin{aligned}
 p &= 2P(TE \geq n_{12} | H_S) \text{ se } n_{12} \geq \frac{(n_{12} + n_{21})}{2} \\
 &2P(TE < n_{12} | H_S) \text{ se } n_{12} < \frac{(n_{12} + n_{21})}{2},
 \end{aligned}$$

com $TE \sim \text{Bin}\left((n_{12} + n_{21}), \frac{1}{2}\right)$.

3.1.3 Modelo logístico - Tabelas 2×2

A análise de associação em dados de trios pode também ser tratada via modelos de regressão logística, os quais, em geral, são vantajosos no sentido de incluírem covariáveis de interesse, contudo como veremos no final desta Seção isto não se aplica para o modelo em questão.

Como destacado anteriormente, dentre as versões do TDT a mais utilizada em dados de trios é a formulação em dados pareados (Tabela 3.1). Neste contexto, considere a seguir a descrição do modelo logístico definido para dados pareados em tabelas 2×2 e contextualizada para o problema genético.

Seja $y_{ij}=1$ se o indivíduo i (mãe ou pai) carrega o alelo A e $y_{ij}=0$ se o indivíduo i (mãe ou pai) carrega o alelo a na condição j , sendo $j=1$ para alelo transmitido (T) e $j=2$ para alelo não transmitido (NT), com $i = 1, \dots, 2n$ com n o número de trios (para cada trio temos informação de dois indivíduos).

Na formulação do modelo logístico em questão, seja:

$$P(y_{ij} = 1) = \frac{\exp\{\mu_i + \beta x_{ij}\}}{1 + \exp\{\mu_i + \beta x_{ij}\}}, \quad (3.9)$$

em que, μ_i representa o efeito de indivíduo i , β é o parâmetro associado a transmissão e não transmissão dos alelos e $x_{ij} = \begin{cases} 1, & \text{se } j = 1 \\ 0, & \text{se } j = 2. \end{cases}$

Agora, escrevendo a probabilidade condicional de y_{ij} como a razão de probabilidades (ver, por exemplo, Stokes et al., 2000), tem-se:

$$\begin{aligned} P(y_{i1} = 1, y_{i2} = 0 | y_{i1} = 1, y_{i2} = 0 \text{ ou } y_{i1} = 0, y_{i2} = 1) = \\ = \frac{P(y_{i1} = 1)P(y_{i2} = 0)}{P(y_{i1} = 1)P(y_{i2} = 0) + P(y_{i1} = 0)P(y_{i2} = 1)}. \end{aligned} \quad (3.10)$$

Em que (3.10) corresponde à probabilidade do indivíduo transmitir o alelo A e não transmitir o alelo a dado que, além desta possibilidade, poderia ter ocorrido a transmissão do alelo a e a não transmissão do alelo A .

Reescrevendo as probabilidades envolvidas em (3.10) em termos dos parâmetros do modelo (3.9), tem-se

$$P(y_{i1} = 1)P(y_{i2} = 0) = \frac{\exp\{\mu_i + \beta\}}{1 + \exp\{\mu_i + \beta\}} \frac{1}{1 + \exp\{\mu_i\}} e,$$

$$P(y_{i1} = 0)P(y_{i2} = 1) = \frac{1}{1 + \exp\{\mu_i + \beta\}} \frac{\exp\{\mu_i\}}{1 + \exp\{\mu_i\}}.$$

Assim,

$$P(y_{i1} = 1)P(y_{i2} = 0) + P(y_{i1} = 0)P(y_{i2} = 1) = \frac{\exp\{\mu_i + \beta\} + \exp\{\mu_i\}}{[1 + \exp\{\mu_i + \beta\}][1 + \exp\{\mu_i\}]}.$$

Logo, a razão de probabilidades (3.10), é dada por:

$$P(y_{i1} = 1, y_{i2} = 0 / y_{i1} = 1, y_{i2} = 0 \text{ ou } y_{i1} = 0, y_{i2} = 1) =$$

$$= \frac{\frac{\exp\{\mu_i + \beta\}}{1 + \exp\{\mu_i + \beta\}} \frac{1}{1 + \exp\{\mu_i\}}}{\frac{\exp\{\mu_i + \beta\} + \exp\{\mu_i\}}{[1 + \exp\{\mu_i + \beta\}][1 + \exp\{\mu_i\}]}} = \frac{\exp\{\mu_i + \beta\}}{\exp\{\mu_i + \beta\} + \exp\{\mu_i\}} = \frac{e^\beta}{1 + e^\beta}.$$

Note que, ao adotarmos esta probabilidade condicional estamos reduzindo o número de parâmetros a serem estimados, cuja expressão só depende de β , sendo os μ_i parâmetros de perturbação. Além disso, o modelo logístico formulado para dados de trios não é influenciado pela inclusão de covariáveis avaliadas nos pais nem mesmo nos filhos, exceto se estas covariáveis forem avaliadas dentro do indivíduo em cada situação cromossômica (cromossomo transmitido e cromossomo não transmitido).

Deste modo, o modelo de regressão logístico em dados de trios (análise pareada) pode ser formulado em termos da verossimilhança condicional, definida como:

$$L(\beta) \propto \prod_{i=1}^{n_{..}} \left\{ \frac{\exp\{\beta\}}{1 + \exp\{\beta\}} \right\}^{y_{i1}(1-y_{i2})} \left\{ \frac{1}{1 + \exp\{\beta\}} \right\}^{(1-y_{i1})y_{i2}}. \quad (3.11)$$

A hipótese de interesse a ser testada é $H_0 : \beta = 0$, que é equivalente a testar se a razão de probabilidades (3.10) é igual a 1/2. Fazendo analogia com a hipótese de simetria discutida na Seção anterior, $H_S : \pi_{12} = \pi_{21}$, tem-se que as duas abordagens, TE e regressão logística, envolvem a mesma redução do problema de inferência em tabelas 2×2 .

3.2 Caso Multiloco - Análise Intervalar

Nos estudos de associação em Genética, quando se trata de mapas densos de SNPs, uma possível busca por regiões candidatas é fazer uma análise multilocos, ou seja, avaliar dois ou mais locos de SNPs simultaneamente. Este tipo de análise é a motivação para muitas pesquisas atualmente, por este problema ainda não estar bem solucionado na literatura e ser uma alternativa que reduz o número de testes usados para verificar associação no genoma inteiro. Além disso, análises multilocos podem ser uma solução ao problema de falsos negativos devido ao efeito individual do SNP ser pequeno. Narain (2007) e Kao et al. (2002), o primeiro no contexto de dados de trios e o segundo no contexto de delineamentos com cruzamentos controlados, defendem que uma análise multilocos envolvendo apenas pares de locos é suficiente para reduzir a ocorrência de falsos negativos e dar poder para a análise de associação genética. Neste contexto, será abordada, a seguir, uma análise de associação multilocos intervalar. As formulações em termos de testes exatos e o modelo logístico em tabelas 4×4 decorrentes da avaliação simultânea dos genótipos para pares de SNPs são tratadas.

3.2.1 TDT generalizado

Considere que os SNPs dispostos em um mapa de marcadores moleculares estão arranjados sequencialmente de acordo com a sua posição no genoma (em pares de bases). Neste sentido, com o objetivo de fazer uma análise considerando a inclusão de efeitos de pares de SNPs adjacentes, uma formulação do teste de associação para dados de trios via modelos de simetria em tabelas 4×4 será considerada nessa seção. Na análise multilocos é avaliada a transmissão ou não, dos pais para um filho afetado, de alelos em dois ou mais locos consecutivos. Chamamos de haplótipo ao fragmento cromossômico correspondente ao arranjo dos alelos nestes locos.

As tabelas de contingência geradas dos dados de trios quando se considera um único loco são da ordem 2×2 (ocorrência da resposta A ou a nas condições transmitido ou não transmitido). Note que, para dois locos, podemos supor o loco 1 com os alelos A e a e o loco 2 com os alelos B e b . Neste caso, a tabela de contingência resultante é da ordem 4×4 (ocorrência de todas as respostas possíveis das respostas AB , Ab , aB ou ab nas condições transmitido ou não transmitido).

Narain (2007) discute que em vez de estudar um conjunto de marcadores em busca do gene da doença,

é suficiente considerar apenas um par de marcadores por vez, ou seja, avaliar se um pequeno intervalo de mapeamento está associado com a doença. Neste contexto, Narain (2007) sugere uma extensão à teoria do TDT para dois locos e propõe um teste de simetria baseado na estatística qui-quadrado, cujo poder é discutido em termos de parâmetros de não centralidade desta distribuição. O autor conclui que o teste intervalar tem maior poder que a análise uniloco.

O mapeamento intervalar supõe que o gene causal da doença (d) esteja entre os dois locos de marcadores genéticos (Figura 3.1) ou na sua vizinhança. O poder da análise intervalar depende do grau de associação do par de locos adjacentes com o loco da doença (quanto mais associados maior é o poder) e da prevalência da doença na população (quanto mais rara, maior é o poder).

Figura 3.1 *Ilustração de mapeamento intervalar considerando dois locos.*

Para a análise intervalar considere dois locos, por exemplo, 1 e 2 (ver Figura 3.2), com os genótipos (AA , Aa , aa) e (BB , Bb , bb), respectivamente. Para esses locos têm-se 4 possíveis haplótipos: AB , Ab , aB e ab . Como ilustração, a Tabela 3.2 apresenta uma disposição destes dados. Para este tipo de tabela 4×4 o teste de associação genético pode ser realizado por meio do teste da hipótese de simetria dada por $H_S : \pi_{ij} = \pi_{ji}$, em que, $1 \leq j \leq i \leq 4$ e π_{ij} é a probabilidade de o indivíduo (pai ou mãe de filho afetado) transmitir o haplótipo i e não transmitir o haplótipo j . A estatística para esse teste pode ser definida como a de McNemar generalizada:

$$\chi^2 = \sum \sum (n_{ij} - n_{ji})^2 / (n_{ij} + n_{ji}), \quad (3.12)$$

com $2 \leq i \leq 4$, $1 \leq j \leq (i-1)$. A estatística (3.12) segue uma distribuição assintótica qui-quadrado com 6 graus de liberdade. A Tabela 3.2 pode ser particionada em seis tabelas 2×2 formadas pelo condicionamento dos dados somente para uma determinada tabela escolhida (de interesse), ou seja, estamos considerando subtabelas e utilizando testes condicionais. A Figura 3.3 mostra tais subtabelas. Por exemplo, poderíamos estar interessados na estatística de um específico teste de associação ou na hipótese “condicional” quando

o marcador AB está associado ao gene da doença que é dada por,

$$\chi_1^2(AB/ab) = (n_{14} - n_{41})^2 / (n_{14} + n_{41}), \quad (3.13)$$

ou se o marcador Ab está associado ao gene da doença, cuja respectiva estatística do teste é:

$$\chi_1^2(Ab/aB) = (n_{23} - n_{32})^2 / (n_{23} + n_{32}). \quad (3.14)$$

Figura 3.2 Composição dos alelos em dois locos.

Tabela 3.2 Transmissão de haplótipos - Amostra pareada.

Transmitidos	Não Transmitidos				Total
	AB	Ab	aB	ab	
AB	n_{11}	n_{12}	n_{13}	n_{14}	$n_{1.}$
Ab	n_{21}	n_{22}	n_{23}	n_{24}	$n_{2.}$
aB	n_{31}	n_{32}	n_{33}	n_{34}	$n_{3.}$
ab	n_{41}	n_{42}	n_{43}	n_{44}	$n_{4.}$
Total	$n_{.1}$	$n_{.2}$	$n_{.3}$	$n_{.4}$	$n_{..}$

Deste modo, a estatística (3.12) fica da forma

$$\chi^2 = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}} + \frac{(n_{13} - n_{31})^2}{n_{13} + n_{31}} + \frac{(n_{14} - n_{41})^2}{n_{14} + n_{41}} + \frac{(n_{23} - n_{32})^2}{n_{23} + n_{32}} + \frac{(n_{24} - n_{42})^2}{n_{24} + n_{42}} + \frac{(n_{34} - n_{43})^2}{n_{34} + n_{43}}. \quad (3.15)$$

A estatística qui-quadrado de McNemar generalizada (3.15) é denominada na área da genética como estatística TDT generalizada. De maneira geral, tabelas quadradas do tipo I^2 são úteis para acomodar

Figura 3.3 *Composição das subtabelas considerando dois locos.*

análises de associação multilocos em que $I = 2^K$, com K igual ao número de locos incluídos na análise. Assim, a estatística TDT generalizada assintoticamente tem distribuição qui-quadrado com $I(I - 1)/2$ graus de liberdade.

Seja $\mathbf{n} = (n_{11}, n_{12}, n_{13}, n_{14}, n_{21}, n_{22}, n_{23}, n_{24}, n_{31}, n_{32}, n_{33}, n_{34}, n_{41}, n_{42}, n_{43}, n_{44})^\top$, o vetor de frequências, tal que $\sum_{ij} n_{ij} = n_{..}$ e considere o vetor de probabilidades correspondente $\boldsymbol{\pi} = (\pi_{11}, \pi_{12}, \pi_{13}, \pi_{14}, \pi_{21}, \pi_{22}, \pi_{23}, \pi_{24}, \pi_{31}, \pi_{32}, \pi_{33}, \pi_{34}, \pi_{41}, \pi_{42}, \pi_{43}, \pi_{44})^\top$, satisfazendo $\sum_{ij} \pi_{ij} = 1$.

Neste caso, a hipótese de simetria de interesse pode ser formulada também como um teste de contrastes:

$$H_S : \mathbf{C}^\top \boldsymbol{\pi} = \mathbf{0}, \quad (3.16)$$

em que,

$$\mathbf{C}^\top = \begin{pmatrix} 0 & 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 & 0 \end{pmatrix}.$$

As duas estatísticas (3.13 e 3.14) são as de maior importância em estudos de associação genética com dados de trios. Note que, este delineamento é útil para modelar doenças recessivas, em que os filhos

são afetados e os pais são livres da doença e tem-se como dados informativos aqueles trios em que os pais são heterozigotos. Assim, considerando subtabelas como vistas na Figura 3.3, os únicos componentes de interesse são quando o marcador AB ou Ab está associado ao gene da doença, caselas (n_{14}, n_{41}) e (n_{23}, n_{32}) , respectivamente.

3.2.2 Teste exato - Tabelas 4×4

Da mesma forma como visto em tabelas 2×2 , nesta Seção é desenvolvido o teste exato em tabelas 4×4 . O uso da estatística TDT generalizada não é recomendada quando se tem frequências esperadas nas caselas da Tabela 3.2 menores ou iguais a 5, o que é frequente em dados de trios, principalmente em casos multilocos. Desta maneira, considere a seguinte verossimilhança obtida a partir da distribuição multinomial imposta aos dados em tabelas 4×4 :

$$L(\boldsymbol{\pi}) = P(\mathbf{n}|\mathbf{n}_{..}, \boldsymbol{\pi}) = P(n_{11}, n_{12}, n_{13}, n_{14}, \dots, n_{44}|\mathbf{n}_{..}, \boldsymbol{\pi}). \quad (3.17)$$

A verossimilhança (3.17) pode ser decomposta da seguinte forma:

$$\begin{aligned} L(\boldsymbol{\pi}) &= P(\mathbf{n}|\mathbf{n}_{..}, \boldsymbol{\pi}) = P(n_{11}, n_{12}, n_{13}, n_{14}, \dots, n_{44}|\mathbf{n}_{..}, \boldsymbol{\pi}) \\ P(\mathbf{n}|\mathbf{n}_{..}, \boldsymbol{\pi}) &= P[(n_{12}, n_{21}), (n_{13}, n_{31}), (n_{14}, n_{41}), (n_{23}, n_{32}), (n_{24}, n_{42}), (n_{34}, n_{43}), n_{1.}, n_{2.}, n_{3.}|\mathbf{n}_{..}, \boldsymbol{\pi}] = \\ &= P(n_{11}, n_{22}, n_{33}, n_{12}, n_{13}, n_{14}, n_{23}, n_{24}, n_{34}, n_{12} + n_{21}, n_{13} + n_{31}, n_{14} + n_{41}, n_{23} + n_{32}, n_{24} + n_{42}, n_{34} + \\ &\quad n_{43}|\mathbf{n}_{..}, \boldsymbol{\pi}) \\ &= P(n_{12} + n_{21}|\mathbf{n}_{..}, \boldsymbol{\pi}) \\ &\quad P(n_{11}/(n_{12} + n_{21}), \mathbf{n}_{..}, \boldsymbol{\pi}) \\ &\quad P(n_{12}/n_{11}, n_{12} + n_{21}, \mathbf{n}_{..}, \boldsymbol{\pi}) \\ &\quad P(n_{13} + n_{31}|n_{11}, n_{12}, (n_{12} + n_{21}), \mathbf{n}_{..}, \boldsymbol{\pi}) \\ &\quad P(n_{13}|n_{11}, n_{12}, n_{12} + n_{21}, n_{13} + n_{31}, \mathbf{n}_{..}, \boldsymbol{\pi}) \\ &\quad P[(n_{14} + n_{41})|n_{11}, n_{12}, n_{13}, n_{12} + n_{21}, n_{13} + n_{31}, \mathbf{n}_{..}, \boldsymbol{\pi}] \\ &\quad P(n_{14}|n_{11}, n_{12}, n_{13}, n_{12} + n_{21}, n_{13} + n_{31}, n_{14} + n_{41}, \mathbf{n}_{..}, \boldsymbol{\pi}) \\ &\quad P(n_{22}|n_{11}, n_{12}, n_{13}, n_{14}, n_{12} + n_{21}, n_{13} + n_{31}, n_{14} + n_{41}, \mathbf{n}_{..}, \boldsymbol{\pi}) \\ &\quad P[(n_{23} + n_{32})|n_{11}, n_{12}, n_{13}, n_{14}, n_{22}, n_{12} + n_{21}, n_{13} + n_{31}, n_{14} + n_{41}, \mathbf{n}_{..}, \boldsymbol{\pi}] \\ &\quad P(n_{23}|n_{11}, n_{12}, n_{13}, n_{14}, n_{22}, n_{12} + n_{21}, n_{13} + n_{31}, n_{14} + n_{41}, n_{23} + n_{32}, \mathbf{n}_{..}, \boldsymbol{\pi}) \\ &\quad P(n_{24} + n_{42}|n_{11}, n_{12}, n_{13}, n_{14}, n_{22}, n_{23}, n_{12} + n_{21}, n_{13} + n_{31}, n_{14} + n_{41}, n_{23} + n_{32}, \mathbf{n}_{..}, \boldsymbol{\pi}) \\ &\quad P(n_{24}|n_{11}, n_{12}, n_{13}, n_{14}, n_{22}, n_{23}, n_{12} + n_{21}, n_{13} + n_{31}, n_{14} + n_{41}, n_{23} + n_{32}, n_{24} + n_{42}, \mathbf{n}_{..}, \boldsymbol{\pi}) \\ &\quad P(n_{33}|n_{11}, n_{12}, n_{13}, n_{14}, n_{22}, n_{23}, n_{24}, n_{12} + n_{21}, n_{13} + n_{31}, n_{14} + n_{41}, n_{23} + n_{32}, n_{24} + n_{42}, \mathbf{n}_{..}, \boldsymbol{\pi}) \\ &\quad P[(n_{34} + n_{43})|n_{11}, n_{12}, n_{13}, n_{14}, n_{22}, n_{23}, n_{24}, n_{33}, n_{12} + n_{21}, n_{13} + n_{31}, n_{14} + n_{41}, n_{23} + n_{32}, n_{24} + n_{42}, \mathbf{n}_{..}, \boldsymbol{\pi}] \\ &\quad P(n_{34}|n_{11}, n_{12}, n_{13}, n_{14}, n_{22}, n_{23}, n_{24}, n_{33}, n_{12} + n_{21}, n_{13} + n_{31}, n_{14} + n_{41}, n_{23} + n_{32}, n_{24} + n_{42}, n_{34} + n_{43}, \mathbf{n}_{..}, \boldsymbol{\pi}). \end{aligned}$$

Considere a seguinte notação para os respectivos componentes da função de verossimi-lhança dada anteriormente em 3.17:

$$L(\boldsymbol{\pi}) = L_1(\pi_{12} + \pi_{21})$$

$$L_2(\pi_{11}, \pi_{12}, \pi_{21})$$

$$L_3(\pi_{12}, \pi_{21})$$

$$L_4(\pi_{11}, \pi_{12}, \pi_{21}, \pi_{13}, \pi_{31})$$

$$L_5(\pi_{13}, \pi_{31})$$

$$L_6(\pi_{11}, \pi_{12}, \pi_{21}, \pi_{13}, \pi_{31}, \pi_{14}, \pi_{41})$$

$$L_7(\pi_{14}, \pi_{41})$$

$$L_8(\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}, \pi_{13}, \pi_{31}, \pi_{14}, \pi_{41})$$

$$L_9(\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}, \pi_{13}, \pi_{31}, \pi_{14}, \pi_{41}, \pi_{23}, \pi_{32})$$

$$L_{10}(\pi_{23}, \pi_{32})$$

$$L_{11}(\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}, \pi_{13}, \pi_{31}, \pi_{14}, \pi_{41}, \pi_{23}, \pi_{32}, \pi_{24}, \pi_{42})$$

$$L_{12}(\pi_{24}, \pi_{42})L_{13}(\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}, \pi_{13}, \pi_{31}, \pi_{33}, \pi_{14}, \pi_{41}, \pi_{23}, \pi_{32}, \pi_{24}, \pi_{42})$$

$$L_{14}(\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}, \pi_{13}, \pi_{31}, \pi_{33}, \pi_{14}, \pi_{41}, \pi_{23}, \pi_{32}, \pi_{24}, \pi_{42}, \pi_{34}, \pi_{43})$$

$$L_{15}(\pi_{34}, \pi_{43}).$$

As correspondentes distribuições de probabilidades associadas a cada termo da decomposição de 3.17 estão demonstradas no Apêndice B. Vale ressaltar que sob esta partição, as 6 distribuições de interesse são as referentes aos termos $L_3(\pi_{12}, \pi_{21})$, $L_5(\pi_{13}, \pi_{31})$, $L_7(\pi_{14}, \pi_{41})$, $L_{10}(\pi_{23}, \pi_{32})$, $L_{12}(\pi_{24}, \pi_{42})$ e $L_{15}(\pi_{34}, \pi_{43})$ da fatoração de 3.17, que são as únicas que dependem dos parâmetros de interesse envolvidos na hipótese de simetria de interesse. As distribuições para os respectivos termos são:

$$L_3(\pi_{12}, \pi_{21}):$$

$$n_{12}|n_{12} + n_{21}, n_{..}, \boldsymbol{\pi} \sim \text{Bin}\left(n_{12} + n_{21}, \frac{\pi_{12}}{\pi_{12} + \pi_{21}}\right)$$

$$L_5(\pi_{13}, \pi_{31}):$$

$$n_{13}|n_{11}, n_{12}, n_{12} + n_{21}, n_{13} + n_{31}, n_{..}, \boldsymbol{\pi} \sim \text{Bin}\left(n_{13} + n_{31}, \frac{\pi_{13}}{\pi_{13} + \pi_{31}}\right)$$

$$L_7(\pi_{14}, \pi_{41}):$$

$$n_{14}|n_{11}, n_{12}, n_{13}, n_{12} + n_{21}, n_{13} + n_{31}, n_{14} + n_{41}, n_{..}, \boldsymbol{\pi} \sim \text{Bin}\left(n_{14} + n_{41}, \frac{\pi_{14}}{\pi_{14} + \pi_{41}}\right)$$

$$L_{10}(\pi_{23}, \pi_{32}):$$

$$n_{23}|n_{11}, n_{12}, n_{13}, n_{14}, n_{22}, n_{12} + n_{21}, n_{13} + n_{31}, n_{14} + n_{41}, n_{23} + n_{32}, n_{..}, \boldsymbol{\pi} \sim \text{Bin}\left(n_{23} + n_{32}, \frac{\pi_{23}}{\pi_{23} + \pi_{32}}\right)$$

$$L_{12}(\pi_{24}, \pi_{42}):$$

$$n_{24}|n_{11}, n_{12}, n_{13}, n_{14}, n_{22}, n_{23}, n_{12} + n_{21}, n_{13} + n_{31}, n_{14} + n_{41}, n_{23} + n_{32}, n_{24} + n_{42}, n_{..}, \boldsymbol{\pi} \sim \text{Bin}\left(n_{24} + n_{42}, \frac{\pi_{24}}{\pi_{24} + \pi_{42}}\right)$$

$$L_{15}(\pi_{34}, \pi_{43}):$$

$$n_{34}|n_{11}, n_{12}, n_{13}, n_{14}, n_{22}, n_{23}, n_{24}, n_{33}, n_{12} + n_{21}, n_{13} + n_{31}, n_{14} + n_{41}, n_{23} + n_{32}, n_{24} + n_{42}, n_{34} + n_{43}, n_{..}, \boldsymbol{\pi} \sim \text{Bin}\left(n_{34} + n_{43}, \frac{\pi_{34}}{\pi_{34} + \pi_{43}}\right)$$

Os componentes da função de verossimilhança $L_7(\pi_{14}, \pi_{41})$ e $L_{10}(\pi_{23}, \pi_{32})$ são os de maior interesse no estudo de associação genética, visto que na Tabela 3.2 é a diagonal que é informativa (componentes de interesse). Além disso, note que, por exemplo, considerando o termo L_7 sob a hipótese $H_{S7} : \pi_{14} = \pi_{41}$ tem-se $n_{14}|(n_{14} + n_{41}), n_{..}, \boldsymbol{\pi} \sim \text{Bin}\left(n_{14} + n_{41}, \frac{1}{2}\right)$, isto é, reduz-se à distribuição na qual um teste exato da hipótese de simetria H_{S7} pode ser construído. De forma análoga, pode-se obter o teste exato para a outra hipótese de interesse $H_{S10} : \pi_{23} = \pi_{32}$ (termo L_{10}), em que $n_{23}|(n_{23} + n_{32}), n_{..}, \boldsymbol{\pi} \sim \text{Bin}\left(n_{23} + n_{32}, \frac{1}{2}\right)$.

3.2.3 Modelo logístico - Tabelas 4×4

A seguir, descrevemos um modelo logístico para estudos de associação genética com dados pareados como na Tabela 3.2, que pode ser uma formulação útil na análise incluindo os efeitos de locos adjacentes, no nosso caso, a cada dois SNPs de um mapa de marcadores.

A forma funcional do modelo é dada por:

$$P\{y_{ij}\} = \frac{\exp\{\mu_i + \beta_{12}x_{1ij} + \beta_{13}x_{2ij} + \beta_{14}x_{3ij} + \beta_{23}x_{4ij} + \beta_{24}x_{5ij} + \beta_{34}x_{6ij}\}}{1 + \exp\{\mu_i + \beta_{12}x_{1ij} + \beta_{13}x_{2ij} + \beta_{14}x_{3ij} + \beta_{23}x_{4ij} + \beta_{24}x_{5ij} + \beta_{34}x_{6ij}\}}, \quad (3.18)$$

em que,

$$y_{ij} = \begin{cases} 0, & \text{se ocorre o haplótipo } ab \\ 1, & \text{se ocorre o haplótipo } aB \\ 2, & \text{se ocorre o haplótipo } Ab \\ 3, & \text{se ocorre o haplótipo } AB, \end{cases}$$

μ_i é o efeito de indivíduo, com o i -ésimo indivíduo ($i = 1, \dots, n$) na j -ésima condição, com $j=1$ para transmitido (T), $j=2$ para não transmitido (NT), os β 's são os parâmetros associados a transmissão e não transmissão dos haplótipos e, considerando

$$k = \{ab, aB, Ab, AB\},$$

$$x_{1ij} = \begin{cases} 1, & \text{se } j = 1, k = AB \\ 0, & \text{se } j = 2, k = Ab, \end{cases} \quad x_{2ij} = \begin{cases} 1, & \text{se } j = 1, k = AB \\ 0, & \text{se } j = 2, k = aB, \end{cases}$$

$$x_{3ij} = \begin{cases} 1, & \text{se } j = 1, k = AB \\ 0, & \text{se } j = 2, k = ab, \end{cases} \quad x_{4ij} = \begin{cases} 1, & \text{se } j = 1, k = Ab \\ 0, & \text{se } j = 2, k = aB, \end{cases}$$

$$x_{5ij} = \begin{cases} 1, & \text{se } j = 1, k = Ab \\ 0, & \text{se } j = 2, k = ab, \end{cases} \quad x_{6ij} = \begin{cases} 1, & \text{se } j = 1, k = aB \\ 0, & \text{se } j = 2, k = ab. \end{cases}$$

Escrevendo as probabilidades condicionais de y_{ij} como a razão de probabilidades (ver, por exemplo, Stokes et al. 2000), e fazendo analogia com o caso anterior avaliado em Tabelas 2×2 , têm-se os 6 casos de medidas de risco genético do haplótipo, como segue:

Caso 1:

$$\begin{aligned} & P(y_{i1} = 3, y_{i2} = 2 | y_{i1} = 3, y_{i2} = 2 \text{ ou } y_{i1} = 2, y_{i2} = 3) = \\ &= \frac{P(y_{i1} = 3)P(y_{i2} = 2)}{P(y_{i1} = 3)P(y_{i2} = 2) + P(y_{i1} = 2)P(y_{i2} = 3)} = \frac{e^{\beta_{12}}}{1 + e^{\beta_{12}}}. \end{aligned} \quad (3.19)$$

Caso 2:

$$\begin{aligned} & P(y_{i1} = 3, y_{i2} = 1 | y_{i1} = 3, y_{i2} = 1 \text{ ou } y_{i1} = 1, y_{i2} = 3) = \\ &= \frac{P(y_{i1} = 3)P(y_{i2} = 1)}{P(y_{i1} = 3)P(y_{i2} = 1) + P(y_{i1} = 1)P(y_{i2} = 3)} = \frac{e^{\beta_{13}}}{1 + e^{\beta_{13}}}. \end{aligned} \quad (3.20)$$

Caso 3:

$$\begin{aligned} & P(y_{i1} = 3, y_{i2} = 0 | y_{i1} = 3, y_{i2} = 0 \text{ ou } y_{i1} = 0, y_{i2} = 3) = \\ &= \frac{P(y_{i1} = 3)P(y_{i2} = 0)}{P(y_{i1} = 3)P(y_{i2} = 0) + P(y_{i1} = 0)P(y_{i2} = 3)} = \frac{e^{\beta_{14}}}{1 + e^{\beta_{14}}}. \end{aligned} \quad (3.21)$$

Caso 4:

$$\begin{aligned} & P(y_{i1} = 2, y_{i2} = 1 | y_{i1} = 2, y_{i2} = 1 \text{ ou } y_{i1} = 1, y_{i2} = 2) = \\ &= \frac{P(y_{i1} = 2)P(y_{i2} = 1)}{P(y_{i1} = 2)P(y_{i2} = 1) + P(y_{i1} = 1)P(y_{i2} = 2)} = \frac{e^{\beta_{23}}}{1 + e^{\beta_{23}}}. \end{aligned} \quad (3.22)$$

Caso 5:

$$\begin{aligned} & P(y_{i1} = 2, y_{i2} = 0 | y_{i1} = 2, y_{i2} = 0 \text{ ou } y_{i1} = 0, y_{i2} = 2) = \\ &= \frac{P(y_{i1} = 2)P(y_{i2} = 0)}{P(y_{i1} = 2)P(y_{i2} = 0) + P(y_{i1} = 0)P(y_{i2} = 2)} = \frac{e^{\beta_{24}}}{1 + e^{\beta_{24}}}. \end{aligned} \quad (3.23)$$

Caso 6:

$$\begin{aligned} & P(y_{i1} = 1, y_{i2} = 0 | y_{i1} = 1, y_{i2} = 0 \text{ ou } y_{i1} = 0, y_{i2} = 1) = \\ &= \frac{P(y_{i1} = 1)P(y_{i2} = 0)}{P(y_{i1} = 1)P(y_{i2} = 0) + P(y_{i1} = 0)P(y_{i2} = 1)} = \frac{e^{\beta_{34}}}{1 + e^{\beta_{34}}}. \end{aligned} \quad (3.24)$$

Como discutido nas seções anteriores os casos de maior interesse nesse tipo de estudo são os Casos 3 e 4, ou seja, se o genótipo AB , de um par de marcador, está associado ao gene da doença e se o genótipo Ab está associado com a doença, respectivamente.

As hipóteses de interesse a serem testadas nestes casos são: $H_{03} : \beta_{14} = 0$ e $H_{04} : \beta_{23} = 0$, que é o mesmo que testar se as razões de probabilidades (3.21) e (3.22) são iguais a $1/2$. Fazendo analogia com a hipótese de interesse a ser testada, por exemplo, como no Caso 3, tem-se interesse em testar $H_{C3} : \pi_{14} = \pi_{41}$. Ao fatorarmos a distribuição multinomial temos interesse em testar se $\left(\frac{\pi_{14}}{\pi_{14} + \pi_{41}} = \frac{1}{2}\right)$, ou seja, as formulações exatas e logística envolvem o mesmo procedimento de redução nas inferências realizadas. O mesmo ocorrendo para a outra hipótese de interesse (Caso 4) $H_{C4} : \pi_{23} = \pi_{32}$, que são os dois casos de maior interesse.

Diante destes resultados, vale ressaltar, que a possibilidade de diminuição do número de graus de liberdade a ser considerado nas estatísticas de teste na análise intervalar é atrativa, considerando a alta dimensão do problema e a ocorrência de muitos falsos negativos vindos da análise uniloco.

Capítulo 4

Seleção de Regiões Candidatas

Atualmente, os estudos que visam a busca por marcadores moleculares (regiões cromossômicas) associados com uma doença têm crescido muito. Neste capítulo, revisaremos alguns métodos de seleção de regiões candidatas em espaços de alta dimensão, como o proposto por Aschard et al. (2007) e, como proposta deste trabalho, para a identificação destas regiões do genoma é adotado o método CUSUM (do inglês, *Cumulative Sum*). Os passos do procedimento proposto são descritos neste capítulo, o qual abrange alternativas ao teste TDT.

4.1 Métodos de seleção de regiões candidatas

Alguns métodos têm sido propostos em estudos de associação para a detecção de regiões cromossômicas candidatas definidas por um subconjunto de marcadores dispostos de forma ordenada no genoma, como alternativa às análises que investigam locos individuais ao longo do genoma.

Com o objetivo de pesquisar o mapa de marcadores para encontrar regiões candidatas pode-se destacar duas classes de procedimentos: análises de associação entre os locos de marcadores e análises de associação entre os locos e a doença. Considerando os dados de marcadores como uma sequência de variáveis

aleatórias assumindo valores inteiros no intervalo $[0,1,2]$, no primeiro caso, em geral usando apenas dados de indivíduos sem a doença (controles), a estratégia é encontrar janelas de dependência entre os marcadores e, somente nas janelas identificadas, realizar análises de associação dos marcadores com a doença de interesse. Estas análises em geral consideram pares de marcadores adjacentes por vez para cobrir todo o genoma e usam testes de associação (conhecidos como testes de desequilíbrio de ligação) em tabelas 3×3 (Terwillinger and Ott, 1994; Weir, 1996). Vários aplicativos na área da Genética adotam tal estratégia, sendo o HaploView (Barrett et al. 2005), disponível no site www.broad.mit.edu/mpg/haploview e o PLINK (Purcell et al. 2005), disponível no site www.pngu.mgh.harvard.edu/~purcell/plink, os mais comuns. Uma alternativa a estas análises realizadas aos pares de marcadores é a aplicação de cadeias de Markov (Greenspan and Geiger, 2006). A segunda classe de procedimentos foca diretamente na doença sob estudo e visa encontrar regiões de marcadores adjacentes associados com ela, mais do que associados entre si independentemente da doença sob estudo. A primeira classe de procedimentos é mais genérica e as regiões selecionadas podem ser então associadas a qualquer doença ou variável resposta de interesse. Contudo, neste trabalho é adotada a segunda classe de estratégia exatamente por ser mais dirigida e específica à doença sob estudo. Passa-se a apresentar uma destas alternativas.

4.1.1 Método de alto escore

Os estudos denominados GWAS (do inglês, *Genome-Wide Associations*) têm sido muito abordados na literatura da área da Genética devido ao crescimento e à disponibilidade de mapas densos de marcadores moleculares. Com o objetivo de reduzir custos do estudo e minimizar o problema de múltiplos testes, Aschard et al. (2007) propõem um procedimento em dois passos, em que o primeiro consiste em selecionar regiões genômicas candidatas e o segundo em avaliar o nível de significância destas regiões por meio do FBAT-LC (do inglês, *family-based associations test- linear combination*). Para ilustração do procedimento, esses autores usaram os dados do GAW 15 (*Genetic Analysis Workshop 15*), especificamente o problema 3, o qual corresponde a dados simulados de indivíduos caso-controle, com os casos referentes a artrite reumatóide (AR).

O método de Aschard et al. (2007) consiste em identificar segmentos cromossômicos de alto escore,

como uma estratégia para selecionar regiões candidatas a partir de dados de um mapa denso de SNPs, e se baseia nos três procedimentos descritos a seguir:

1- Produzir uma sequência inicial de valores X corrigidos: os autores obtêm em cada marcador uma estatística de associação (X_i) correspondente à qui-quadrado de Pearson para tabelas de contingência em estudos caso-controle (adotando dados genotípicos). Realizam uma correção destas estatística pela subtração de um valor crítico X_c (em geral, recomendam um nível de significância não muito conservador, por exemplo, 5%). O perfil destas estatísticas corrigidas deve oscilar em torno do zero, havendo interesse nas tendências mais positivas e naquelas que se mantêm positiva, pois indicam a existência de algum possível efeito na região.

2- Identificar uma região de maior escore: uma abordagem para obter um escore local que consiste em comparar o valor de $\sum_a^b (X_i - X_c)$ para todas as janelas $[a, b]$, cobrindo todo o genoma e excluindo regiões que abragem diferentes cromossomos.

3- Identificar uma próxima região de maior escore: uso de um algoritmo que ao encontrar uma região de alto escore, remove ela da sequência, e aplica o algoritmo novamente até que não haja mais escores locais positivos na sequência. Por fim, o número de testes será reduzido de M marcadores para N regiões cromossômicas candidatas classificadas de acordo com os seus escores locais.

4.1.2 Método CUSUM

Com o objetivo de propor neste trabalho uma abordagem de busca por regiões cromossômicas candidatas, a ferramenta adotada se baseia no método CUSUM (do inglês, *Cumulative Sum*), adaptada a proposta de Aschard et al. (2007), vista na seção anterior. Esta foi proposta por Page (1954) e é conhecida como Carta de Somas Acumuladas, bastante usada no controle de processos industriais.

Este método de controle estatístico do processo é indicado para o monitoramento de processos sujeitos a pequenas perturbações, de forma que a decisão sobre a situação do processo é baseada na informação acumulada de uma sequência de amostras (Yi et al. 2006; Montgomery, 2008; Correa et al. 2009). Essas cartas monitoram a média do processo e incorporam diretamente toda a informação da sequência de valores amostrais registrando as somas acumuladas desses valores relativamente a um valor *target* (μ_0), admitindo que as observações seguem uma distribuição normal de média (μ_0) e desvio padrão σ .

Uma das maneiras de se construírem as cartas CUSUM, considerando observações individuais, é a seguinte: uma tabela desta carta é feita acumulando derivações de μ_0 superiores ao valor *target* com uma estatística C^+ e de valores de μ_0 inferiores com uma outra estatística C^- . Essas estatísticas C^+ e C^- podem ser calculadas da seguinte forma:

$$\begin{aligned} C_i^+ &= \max[0, x_i - (\mu_0 + K) + C_{i-1}^+] \\ C_i^- &= \max[0, (\mu_0 - K) - x_i + C_{i-1}^-], \end{aligned} \quad (4.1)$$

em que x_i é o valor da i -ésima observação, K é chamado de valor de referência ou tolerância, e é aproximadamente a metade do valor que se tem interesse em detectar rapidamente, frequentemente calculado por $K = \frac{|\mu_1 - \mu_0|}{2}$, em que μ_1 é o valor da média fora de controle. Este fator K está diretamente relacionado com a magnitude da variação que desejamos detectar com o gráfico CUSUM, quanto menor este fator, menor será a faixa de variação que o gráfico será capaz de detectar e maior será a sensibilidade do gráfico, em geral usa-se $K = 1/2$. Os valores iniciais são $C_0^+ = C_0^- = 0$. Se C_i^+ e C_i^- excedem o intervalo de decisão H , considera-se que o processo está fora de controle. É razoável adotar que o parâmetro H seja um valor igual a cinco vezes o desvio padrão σ do processo e, para finalidade de análises exploratórias, a

hipótese de normalidade pode ser relaxada, ver, por exemplo, Montgomery (2008). O intervalo de decisão pode também ser obtido via procedimentos bootstrap. No nosso caso, a carta CUSUM será utilizada para encontrar blocos ou subconjuntos de SNPs adjacentes, sendo x_i o valor da i -ésima observação (isto é, do i -ésimo marcador) utilizando as estatísticas TDT que são obtidas via a análise uniloco, ou seja, cada SNP na sequência de realizações tem um valor desta estatística. O valor μ_0 é a média de todas as observações e μ_1 , como dito anteriormente, é o valor da média fora de controle. Os pontos “fora de controle” correspondem às regiões candidatas que serão de nosso interesse para uma análise posterior.

4.1.3 Procedimento multiestágios

Nesta seção é apresentada uma sequência de passos a ser adotada como estratégia de análise de grandes mapas de SNPs para finalidade de encontrar regiões candidatas a estarem associadas com a doença de interesse.

Passo 1: *Limpeza dos dados.*

Inicialmente, é rigorosamente recomendada uma “limpeza” dos dados de genótipos de SNPs que consiste na retirada daqueles cujas frequências não seguem o equilíbrio de Hardy-Weinberg. Em seguida, retirar aqueles cuja menor frequência alélica (denota por MAF, do inglês, Minor Allele Frequency) for inferior a 1%. Outras análises exploratórias devem ser realizadas como verificar marcadores com muitos dados faltantes ou, correspondentemente, retirar indivíduos com muitos dados faltantes. Todas estas precauções fazem parte do controle de qualidade dos dados. Finalmente, considerando as análises de associação que serão realizadas, o próximo passo é distinguir os SNPs com tamanho amostral pequeno, isto é, cujas frequências esperadas em alguma das caselas informativas (conforme ilustrado na Tabela 2.4) seja inferior a 5. Este procedimento deve distinguir entre os SNPs que deverão ser analisados via procedimentos de testes de associação exatos ou assintóticos.

Passo 2: *Análise uniloco*

Neste estágio é feita uma análise de associação uniloco baseada na teoria do teste TDT (McNemar), como apresentado na Seção 3.1.1, para as situações nas quais o procedimento assintótico se aplica. Os SNPs são avaliados individualmente, procedimento este feito no aplicativo PLINK (Purcell et al. 2005). Em alternativa ao teste TDT, em tabelas 2×2 , considerando os casos de tamanhos amostrais pequenos, é feita uma análise individual considerando a proposta de teste exato (Capítulo 3, seção 3.1.2). O modelo logístico pode também ser aplicado no caso assintótico como alternativa ao TDT e, apesar de exigir um esforço computacional maior, recomenda-se sua aplicação pela equivalência na derivação teórica com o caso exato. Os resultados destas análises podem ser dispostos em gráficos para a visualização dos perfis dos valores p ou das estatísticas de teste ao longo do genoma.

Passo 3: *Busca por regiões candidatas.*

A partir do gráfico com os perfis das estatísticas dos testes de associação, subconjuntos de SNPs são avaliados para cada cromossomo por meio da seleção de regiões de SNPs pelo método CUSUM, apresentado neste Capítulo na seção 4.1.2. Os valores das variáveis aplicadas a essas somas acumuladas podem ser definidas a partir das estatísticas do teste TDT ou do modelo logístico. Nesta análise são também identificados locos isolados (individuais) significantes para a associação.

Passo 4: *Análise de associação intervalar*

Tendo sido selecionadas regiões contendo subconjuntos de SNPs adjacentes associados com a doença, nestas regiões procede-se com a análise de associação intervalar (descrita no capítulo 3, na seção 3.2), percorrendo pares de SNPs e construindo as correspondentes tabelas 4×4 . Também neste caso, os intervalos que contêm caselas com frequências esperadas menores que 5 são selecionados para a aplicação dos procedimentos de testes exatos. Aos demais pares de SNPs são aplicados testes de associação assintóticos, TDT generalizado ou para específicos termos de interesse, ou ainda a versão logística, que exige mais esforço computacional mas guarda equivalência com a formulação teórica da solução exata.

Capítulo 5

Aplicação

No presente capítulo, tem-se a aplicação das propostas discutidas nos capítulos anteriores. Ressalta-se, que neste trabalho é apresentado um procedimento de análise em multiestágios para selecionar regiões genéticas candidatas, a partir da informação de grandes mapas de marcadores moleculares do tipo SNP e delineamentos com trios, sendo consideradas alternativas ao TDT. A Tabela 5.1, ilustra a entrada de dados que é comumente aceita na leitura feita pelos aplicativos da área de genética, por exemplo, o PLINK (Purcell et al. 2005).

As variáveis que compõem a estrutura familiar são: TRIO- número do trio, ID- identificação do indivíduo, FA- pai e MO- mãe. Por exemplo, no trio 189, a mãe do indivíduo 130 é 129 e o pai é o 131. As variáveis (fenotípicas) observadas são: SEX- sexo dos indivíduos (1- masculino, 2- feminino) e AFFECT (1- não afetado, 2- afetado, por uma doença). As variáveis genotípicas são avaliadas no genoma de todos os indivíduos considerando marcadores do tipo SNPs, os quais são considerados como fatores de risco para a doença sob estudo. Por exemplo, para o SNP2, os indivíduos podem ter os genótipos TT, CT e CC. Estes dados são codificados por meio do número de alelos (raros), possivelmente associados com

a doença, que cada indivíduo carrega. Assim, para o SNP2, se T é o alelo raro (de menor frequência), então, codificam-se as categorias genotípicas TT, CT e CC como 2, 1 e 0, respectivamente.

Tabela 5.1 *Ilustração dos estudos com trios.*

TRIOS	ID	FA	MO	SEX	AFFECT	SNP1	SNP2	...	SNP906.485
189	131	0	0	1	1	TT	TT	...	AA
189	129	0	0	2	1	GT	CT	...	AG
189	130	131	129	1	2	GT	CT	...	AA
191	262	0	0	1	1	GT	TT	...	AA
191	261	0	0	2	1	GG	CT	...	AA
191	263	262	261	1	2	GG	TT	...	AA
192	374	0	0	1	1	TT	CT	...	AG
192	373	0	0	2	1	GT	CC	...	GG
192	372	374	373	2	2	TT	CC	...	GG
193	421	0	0	1	1	GT	TT	...	GG
193	420	0	0	2	1	TT	CT	...	AG
193	419	421	420	2	2	TT	TT	...	GG
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
282	4097	0	0	1	1	TT	CC	...	AG
282	4096	0	0	2	1	GG	CT	...	AA
282	4095	4097	4096	2	2	GT	CC	...	AA

A estratégia de análise em multiestágios proposta no capítulo anterior é ilustrada neste capítulo considerando os dados reais denominados TRIOS DO BRASIL, disponibilizados pelo Laboratório de Genética e Cardiologia Molecular do InCor-USP. Este banco de dados é formado por 71 trios (213 indivíduos) avaliados nos 22 cromossomos autossomos (869.222 SNPs, ver Tabela 5.2), sendo que ao todo, no genoma, têm-se 906.485 SNPs genotipados da plataforma Affymetrics 6.0. Nestes trios, os filhos são afetados por uma cardiopatia congênita. A amostra corresponde a voluntários da população urbana de São Paulo cujos pais procuraram o serviço do InCor para tratamento de seus filhos e a família (trio) foi convidada a participar do estudo. Para estes dados segue o procedimento proposto no Capítulo 4.

Passo 1: *Limpeza dos dados.*

Inicialmente, fez-se uma “limpeza” dos dados que consistiu na retirada daqueles SNPs cujas frequências não seguiram o equilíbrio de Hardy-Weinberg. Em seguida, foram retirados aqueles SNPs cuja menor frequência alélica era inferior a 1%. Posteriormente, e para aplicação das metodologias assintóticas (TDT e modelo logístico), fez-se a retirada de SNPs cuja frequência nas caselas informativas (conforme ilustrado na Tabela 2.4) fosse inferior a 5 (estes SNPs foram retidos da análise para aplicação dos procedimentos de testes exatos). Restaram 673.006 SNPs (Tabela 5.3), em torno de 22% dos SNPs foram retirados, e manteve-se o número de trios.

Tabela 5.2 *Número de SNPs por cromossomo (dados de trios).*

Cromossomo	1	2	3	4	5	6	7	8	9	10	11
Nº de SNPs	71312	73936	60684	55995	56416	56271	47056	48608	41442	48195	44539
Cromossomo	12	13	14	15	16	17	18	19	20	21	22
Nº de SNPs	42555	34283	28065	26074	27716	20658	26529	11929	22843	12579	11537

Tabela 5.3 *Número de SNPs analisados (dados de trios).*

Cromossomo	1	2	3	4	5	6	7	8	9	10	11
Nº de SNPs	54351	56945	47415	42731	43738	43744	36865	37940	32368	37199	34451
Cromossomo	12	13	14	15	16	17	18	19	20	21	22
Nº de SNPs	32896	25793	21704	20361	21604	16183	20283	9558	17994	9906	8977

Passo 2: *Análise uniloco.*

Neste estágio foi realizada a análise de associação uniloco baseada na teoria do teste TDT (McNemar), como apresentado na Seção 3.1.1. Para esta finalidade foi utilizado o aplicativo PLINK (Purcell et al. 2005). Segue na Figura 5.1 os resultados dos testes TDT para os SNPs em cada cromossomo. Note que, por exemplo, para o cromossomo 5 observa-se pontos (SNPs) que merecem “atenção”, ou seja, que podem estar associados com a cardiopatia devido ao alto valor da estatística de teste.

Figura 5.1 *Estatística do TDT para os 22 cromossomos.*

Passo 3: *Busca por regiões candidatas.*

Neste estágio, utilizou-se o perfil das estatísticas do teste TDT ao longo do genoma para identificar regiões cromossômicas associadas à síndrome cardíaca dos dados em questão. Neste caso, a seleção de regiões de SNPs foi feita pelo método CUSUM, apresentado no Capítulo 4. Os valores das variáveis aplicadas a essas somas acumuladas foram as estatísticas do teste TDT. Esta análise foi realizada usando os recursos do aplicativo R *library(qcc)*.

Como temos muitas variáveis, este critério identificou um número muito alto de pontos fora de controle (Figura 5.2), isto é, possivelmente associados com a doença. Assim optou-se por mudar o critério de detecção para $3\mu_0$ e 6σ , pois esta foi a proposta que melhor discriminou as regiões genômicas, relativamente a outras constantes avaliadas multiplicando μ_0 e σ .

Como resultado da aplicação do CUSUM, nos cromossomos 1, 2, 3, 4, 5 e 10 (Figura 5.3) foram identificados blocos de SNPs adjacentes possivelmente associados com a cardiopatia. Por exemplo, para o cromossomo 1, três regiões foram encontradas, em torno das posições em BP (pares de bases): 73559057 (41 SNPs), 165828994 (17 SNPs) e 192908673 (6 SNPs). Assim, para este cromossomo foram identificadas três regiões candidatas, com 64 SNPs ao todo. Vale ressaltar que estes SNPs, na sua maioria, são significantes

Figura 5.2 *CUSUM para o cromossomo 1.*

ao nível de 0,05. Na mesma situação, encontram-se os cromossomos 17, 18, 19 e 22 (Figura 5.4). Por exemplo, no cromossomo 22, foram identificadas duas regiões candidatas e nota-se que apenas um SNP, em torno da posição 33568768 BP, foi detectado como loco candidato pelo método CUSUM. Ao avaliar as estatísticas TDT de outros SNPs em torno deste, percebe-se que mais 6 são significantes ao nível de 0,05. Na segunda região (45768799 BP) deste cromossomo, foram detectados 15 SNPs formando um bloco candidato.

Quando regiões são identificadas sob o procedimento CUSUM, nota-se que regiões próximas podem também mostrar algum sinal de associação, como acontece nos cromossomos 7, 12 e 14 (Figura 5.5). Por exemplo, considerando o cromossomo 7, no qual detectou-se uma região, há duas regiões que podem ser possíveis candidatas (próximas ao limite de tolerância), as quais estão definidas pelos SNPs nas posições 50590206 e 127946146. Isto também acontece para os cromossomos 12 e 14. Observa-se esta mesma situação para os cromossomos 8 e 15 (Figura 5.6). Isto é esperado acontecer, pois o verdadeiro gene associado com a doença pode estar localizado entre os SNPs de uma região selecionada e seu efeito pode ser refletido em toda uma vizinhança de locos de marcadores correlacionados com ele (em desequilíbrio de ligação).

Considerando o cromossomo 16 (Figura 5.7), nota-se que não foi possível, por este critério, detectar algum bloco de SNP com efeito significativo, o que também ocorre nos cromossomos 9, 11, 13 e 21 (Figura 5.8).

Figura 5.3 *CUSUM* para os cromossomos 1, 2, 3, 4, 5 e 10.

Fazendo uma análise dos cromossomos 6 e 20, não se observa nenhuma região candidata, mas com o intuito de flexibilizar o critério de busca por SNPs e de ilustrar que ao diminuir a constante δ que multiplica o desvio padrão no critério CUSUM, aumentam-se os pontos que ficaram fora das bandas de controle, fez-se uma mudança (diminuição) em δ . Nota-se, pela Figura 5.9, para o cromossomo 6, que 5 regiões passaram a ser detectadas, com 34 SNPs, mas em apenas uma região, em torno da posição 87176209 BP, os SNPs são significantes. Fazendo o mesmo para o cromossomo 20 o SNP detectado é significativo e também os que estão próximos a este.

Na sequência da análise destes dados escolhemos uma particular região dentre as identificadas pelo CUSUM, de maior interesse para os médicos que estudam o mapeamento de genes associados com a cardiopatia congênita em questão. Nesta região as análises exatas uniloco, quando apropriadas, foram

Figura 5.4 *CUSUM* para os cromossomos 17, 18, 19 e 22.

Figura 5.5 *CUSUM* para os cromossomos 7, 12 e 14.

Figura 5.6 *CUSUM* para os cromossomos 8 e 15.

Figura 5.7 *CUSUM para o cromossomo 16.*

Figura 5.8 *CUSUM para os cromossomos 9, 11, 13 e 21.*

realizadas para o caso de tabelas 2×2 , bem como procedeu-se com as análises de associação intervalares, no caso de tabelas 4×4 .

Passo 4: *Análise uniloco via o TDT e teste exato na região selecionada*

A região de maior interesse ao estudo de associação selecionada por especialistas do InCor nesta cardiopatia foi uma região do cromossomo 5, também selecionada pelo CUSUM. Nesta região foram evidenciados

Figura 5.9 CUSUM para os cromossomos 6 e 20.

23 SNPs em torno da posição 8703479 BP e, para análise, abriu-se uma janela contendo 100 SNPs cobrindo esta região.

O teste exato em tabelas 2×2 foi descrito no Capítulo 3 na Seção 3.1.2. Desta maneira segue na Figura 5.10 os resultados da aplicação do teste exato considerando todos os SNPs com tamanho amostral pequeno dentre os 100 selecionados. Estes totalizaram 53 SNPs sendo que para os restantes 47 o teste TDT (assintótico) foi aplicado. Na figura estão indicados os valores p correspondentes. O valor de corte mostrado na Figura 5.10 considera um nível de significância de 5%. Sob este critério, ao todo 28 SNPs são significantes.

Seguem, na Tabela 5.4, os valores dos respectivos valores p, considerando os testes TDT (SNPs com asterisco*) e exato. Note que, a partir do SNP 39 (rs12517710), que é onde inicia a região selecionada pelo CUSUM, tem-se um bloco de SNPs que pelos testes avaliados (TDT e exato) são significantes a 5%.

Passo 5: *Análise intervalar na região selecionada*

Após a análise uniloco, ou seja, em tabelas 2×2 , avaliou-se o teste exato e o TDT generalizado considerando a análise em pares de locos para a região de interesse do cromossomo 5. A construção das tabelas 4×4 neste caso não é direta e foi implementado um programa computacional no aplicativo R

Tabela 5.4 *Valores p do TDT e teste exato em tabelas 2×2 .*

	SNP	valor p		SNP	valor p
1	rs16880860	1,0000	51	rs200091*	0,3017
2	rs16880863*	0,2059	52	rs200090	0,6516
3	rs11134317*	0,7815	53	rs6882860	0,6636
4	rs6863400*	0,3538	54	rs200077	0,2221
5	rs6867425*	0,5862	55	rs999419	0,8238
6	rs6881821	1,0000	56	rs10512978	1,0000
7	rs11741809	0,6076	57	rs7341131	0,2500
8	rs1010177	0,7798	58	rs200056	0,2221
9	rs12517255*	1,0000	59	rs200055	0,7359
10	rs11737974	1,0000	60	rs200050	0,2221
11	rs11748838	1,0000	61	rs10060493	1,0000
12	rs4702561	0,3750	62	rs200044	0,3020
13	rs199169*	1,0000	63	rs200031	1,0000
14	rs6555557*	0,0173	64	rs200012	0,2221
15	rs1501320*	0,0173	65	rs200001*	0,1814
16	rs2106320	0,0001	66	rs199999*	0,1967
17	rs11741773*	0,1011	67	rs199998*	0,1213
18	rs1501340	0,7539	68	rs199997*	0,1967
19	rs1392961*	0,2673	69	rs6875808*	0,0168
20	rs2892491*	0,0219	70	rs16881139	0,1797
21	rs1392962*	0,0162	71	rs9313253	0,3750
22	rs1501342*	0,2482	72	rs156477*	0,0192
23	rs16880891	1,0000	73	rs156466	0,0117
24	rs16880893	1,0000	74	rs16881265	1,0000
25	rs4702563*	0,1824	75	rs156457*	0,0269
26	rs199175*	0,0357	76	rs156453	0,3750
27	rs4701815	0,7110	77	rs156450*	0,0961
28	rs1501346	0,0001	78	rs274643*	0,1764
29	rs1501347*	0,0002	79	rs13356951	0,3750
30	rs4702564	0,8555	80	rs16881298*	0,5151
31	rs10040826	0,0001	81	rs999428*	0,0067
32	rs6860594	0,2188	82	rs999427*	0,0222
33	rs7707778	0,1214	83	rs13157524*	0,5078
34	rs7732493	0,0352	84	rs2963394*	0,0578
35	rs7737474	0,0227	85	rs10491225	1,0000
36	rs16880982	0,6250	86	rs7716900	0,6250
37	rs200118	0,6900	87	rs7703050	0,0001
38	rs200116	0,6636	88	rs10491223*	0,2482
39	rs12517710*	0,0001	89	rs7717275*	0,2278
40	rs12187730	0,0001	90	rs3111121	0,0001
41	rs736970	0,0001	91	rs2938816*	0,2249
42	rs7704554	0,0188	92	rs7730627	0,6250
43	rs200114*	0,0009	93	rs2904935	1,0000
44	rs10076745*	0,0234	94	rs10042965	0,3750
45	rs10053640*	0,0820	95	rs3105426*	0,0833
46	rs199198	0,4531	96	rs992319*	0,5078
47	rs199196	1,0000	97	rs1505023*	0,0026
48	rs199195*	0,0016	98	rs10491222*	0,0742
49	rs200107*	0,0295	99	rs1158727*	0,0754
50	rs199194*	0,1451	100	rs995555*	0,0961

Figura 5.10 *Valores-p dos testes TDT e exato em tabelas 2×2 .*

para a leitura dos dados (disponível em <http://jacqueline.dema.ufc.br>). Todos os possíveis pares de locos adjacentes dentre os 100 SNPs da região foram considerados, totalizando 99 pares, nos quais procedeu-se com a análise intervalar. Na Tabela 5.5 os valores em “—” correspondem aos SNPs em que o teste de associação intervalar não foi realizado, devido à ocorrência de caselas nulas. Ainda, nesta construção consideramos a configuração *cis* nos casos de ambigüidades na identificação do material dos dois locos transmitidos dos pais para o filho (proposta esta adotada por vários autores, por exemplo, Narain, 2007).

Teste TDT generalizado e exato em tabelas 4×4

No Capítulo 3, Seção 3.2.1, foi descrito o TDT generalizado em tabelas 4×4 e na Seção 3.2.2 o teste exato. Analisando a região candidata sob estudo, seguem na Tabela 5.5, os valores p correspondentes às duas análises. Os valores p^1 e p^2 correspondem às configurações testadas em H_{S7} e H_{S10} (hipóteses vistas no Capítulo 3, seção 3.2.2), respectivamente. Os SNPs com valor p em asterisco(*) foram analisados pelo TDT generalizado. Essas configurações foram descritas quando consideramos a partição da tabela 4×4 em 6 subtabelas (Seção 3.2.2) e isolamos duas delas de maior interesse na análise. Note que a vantagem de aplicação do teste exato é grande e cresce com o aumento do número de locos envolvidos nas análises (multilocos). Neste caso, dentre os 99 pares de SNPs analisados para as duas hipóteses (correspondendo a 198 situações) somente 20 situações apresentaram tamanhos amostrais apropriados à análise assintótica.

Como resultado, 19 pares de SNPs na região do cromossomo 5 foram significantes para esta análise. Note que, fazendo comparação com as análises uniloco anteriores (CUSUM e os testes TDT e exato uniloco), a região a partir do SNP 39 (rs12517710) mostra-se como candidata a estar associada com o fator de risco para a doença. Observe também que uma região em torno do SNP 96 (rs992319) apresenta-se significativa, o que não aconteceu nas análises uniloco anteriores. Isto mostra, como esperado, a maior sensibilidade da análise intervalar em identificar locos candidatos comparados com a análise uniloco.

A Figura 5.11 apresenta as tabelas uniloco e a correspondente tabela para pares de locos considerando os SNPs 96 e 97. Os dados dos SNPs 96 e 97 mostram que as estimativas de risco, isto é, da probabilidade do indivíduo transmitir o alelo A e não transmitir o alelo a , dado que além desta possibilidade poderia ter ocorrido a transmissão do alelo a e a não transmissão do alelo A , denotada por $\pi_{12}/(\pi_{12} + \pi_{21})$, correspondem a 0,5438 ($p = 0,5966$) e 0,2727 ($p = 0,0037$), respectivamente. Isto indica que somente o segundo loco é significativo, sendo o alelo b do SNP 97 é um fator de risco para a cardiopatia. Considerando a análise intervalar destes SNPs e as estimativas de risco de interesse, tem-se que: (i) a estimativa da probabilidade de um pai (duplo heterozigoto em fase *cis*) transmitir o haplótipo AB e não transmitir o haplótipo ab , dado que além dessa possibilidade ele poderia ter transmitido ab e não transmitido AB , é 0,5714 ($p = 1,0000$) e, (ii) a estimativa da probabilidade de um pai (duplo heterozigoto em fase *trans*) transmitir o haplótipo Ab e não transmitir o haplótipo aB , dado que além dessa possibilidade ele poderia

ter transmitido aB e não transmitido Ab , é 0,9231 ($p = 0,0034$). Isto indica que para os SNPs 96 e 97 o haplótipo aB é de risco para a cardiopatia sob estudo. Deste modo, a partir do SNP 96 até o 99 notou-se uma região genômica candidata o que não ocorreu nas análises uniloco, mostrando a superioridade da análise intervalar em identificar padrões de risco genético associados com doenças de interesse.

SNP 96			SNP 97		
Transmitidos	Não transmitidos		Transmitidos	Não transmitidos	
	A	a		B	b
A	11	31	B	9	12
a	26	68	b	32	33

SNPs 96 e 97				
Transmitidos	Não transmitidos			
	AB	Ab	aB	ab
AB	0	1	1	4
Ab	0	10	12	14
aB	0	1	8	6
ab	3	22	17	37

Figura 5.11 Tabelas uniloco e para pares de locos considerando os SNPs 96 e 97.

Tabela 5.5 *Valores p do teste exato e TDT generalizado em tabelas 4×4 .*

SNPs	valor p^1	valor p^2	SNPs	valor p^1	valor p^2
1 e 2	—	—	51 e 52	0,4408	—
2 e 3	1,0000*	—	52 e 53	0,0625	1,0000
3 e 4	0,4561*	—	53 e 54	0,1250	0,6250
4 e 5	0,2188	0,1573*	54 e 55	0,5000	0,3750
5 e 6	1,0000	1,0000	55 e 56	—	—
6 e 7	0,5000	1,0000	56 e 57	—	—
7 e 8	0,8555	—	57 e 58	—	—
8 e 9	1,0000	0,1797	58 e 59	0,6076	—
9 e 10	1,0000	0,5000	59 e 60	0,6076	—
10 e 11	1,0000	—	60 e 61	0,2500	0,0625
11 e 12	1,0000	—	61 e 62	0,2500	0,2188
12 e 13	—	1,0000	62 e 63	—	—
13 e 14	0,5078	1,0000	63 e 64	—	—
14 e 15	0,0173*	—	64 e 65	0,1433	—
15 e 16	0,0009	1,0000	65 e 66	0,1655*	1,0000
16 e 17	—	0,1250	66 e 67	0,1521*	1,0000
17 e 18	1,0000	1,0000	67 e 68	0,1451*	1,0000
18 e 19	0,7266	1,0000	68 e 69	0,7456*	1,0000
19 e 20	1,0000	0,1460	69 e 70	1,0000	—
20 e 21	0,0219*	—	70 e 71	0,6250	—
21 e 22	1,0000	0,0923	71 e 72	0,6250	—
22 e 23	—	—	72 e 73	0,0703	—
23 e 24	—	—	73 e 74	1,0000	0,2500
24 e 25	—	1,0000	74 e 75	1,0000	1,0000
25 e 26	0,0213	1,0000	75 e 76	—	1,0000
26 e 27	0,0625	0,2891	76 e 77	—	—
27 e 28	1,0000	0,7539	77 e 78	0,1235*	—
28 e 29	0,0001	1,0000	78 e 79	0,5000	1,0000
29 e 30	1,0000	1,0000	79 e 80	0,5000	1,0000
30 e 31	1,0000	0,5078	80 e 81	1,0000	0,7539
31 e 32	1,0000	1,0000	81 e 82	0,0136	—
32 e 33	1,0000	1,0000	82 e 83	1,0000	1,0000
33 e 34	0,4531	—	83 e 84	1,0000	0,6825
34 e 35	0,5078	—	84 e 85	—	1,0000
35 e 36	1,0000	—	85 e 86	—	—
36 e 37	—	1,0000	86 e 87	1,0000	—
37 e 38	0,6636	—	87 e 88	—	0,0009
38 e 39	—	0,3438	88 e 89	0,3173*	—
39 e 40	0,0001	1,0000	89 e 90	1,0000	0,0018
40 e 41	0,0001*	—	90 e 91	1,0000	0,0009
41 e 42	0,0009	1,0000	91 e 92	—	1,0000
42 e 43	0,0026	1,0000	92 e 93	—	—
43 e 44	0,0001	0,1250	93 e 94	—	—
44 e 45	0,0001*	—	94 e 95	—	0,5000
45 e 46	1,0000	1,0000	95 e 96	0,0704*	1,0000
46 e 47	1,0000	—	96 e 97	1,0000	0,0034
47 e 48	1,0000	1,0000	97 e 98	1,0000	0,0225
48 e 49	0,0027*	1,0000	98 e 99	0,0117	0,3938*
49 e 50	0,8658*	0,2500	99 e 100	0,0960*	1,0000
50 e 51	0,1814*	—			

Capítulo 6

Considerações Finais

Tendo em vista o crescente interesse na literatura por estudos que englobam o mapeamento genético de doenças (por exemplo, Giolo et al. 2011, Ziegler et al. 2008, Conti and Gauderman, 2004) e os muitos pontos de pesquisa em aberto nesta área, no presente trabalho foram consideradas diferentes metodologias de análise estatística de dados provenientes de delineamentos com trios. Este tipo de delineamento corresponde a pequenos núcleos familiares (pai e mãe, livres da doença, e filho afetado), os quais representam uma classe importante de amostragem de dados familiares para finalidade de mapeamento de genes (Spielman et al. 1993; Watkins, 2004; Pender et al. 2004), em que são coletados, em cada indivíduo, dados do genótipo de marcadores moleculares além de outras variáveis que possam ser de interesse. Neste trabalho consideramos dados genéticos provenientes de plataformas de marcadores moleculares do tipo SNP (do inglês, *Single Nucleotide Polymorphism*), as quais vêm sendo utilizadas e recomendadas por vários autores (por exemplo, Yang et al., 2010; Amos et al., 2008; Weir et al., 2004) e centros de pesquisa (Projeto HapMap) na avaliação de fatores de risco genéticos para doenças.

Na introdução dos delineamentos com trios foram levantados alguns problemas que mais caracterizam

a análise de dados genômicos, quando o objetivo é o mapeamento de genes, isto é, o de estudar a associação de fatores de risco genético com uma doença. São eles: efeito de confundimento devido à estratificação genética da população, efeito muito pequeno de cada fator de risco (SNP) sobre a doença, alta dimensionalidade do espaço das variáveis preditoras (plataformas de SNPs) e tamanhos amostrais pequenos. Relativamente à estratificação genética da população os delineamentos com trios são úteis pois garantem a amostragem de indivíduos casos e controles que são homogêneos geneticamente. Além disso, na análise destes dados, dependendo da leitura que é feita do delineamento experimental, diferentes tabelas de contingência podem ser construídas, as quais foram apresentadas neste trabalho. Na análise de dados de trios, em geral, aplica-se o teste de associação conhecido na área da Genética como TDT (do inglês, *Transmission Disequilibrium Test*), introduzido por Spielman et al. (1993), para avaliar a associação de marcadores moleculares com a doença, sendo que, atualmente, os marcadores do tipo SNP são os mais utilizados (ver, por exemplo, Bergen et al., 2003; Sykes et al., 2009). A formalização do TDT foi considerada neste trabalho como um teste de simetria em tabelas de contingência, o que permitiu sua generalização para o caso de tamanhos amostrais pequenos e de análises intervalares (isto é, para pares de SNPs simultaneamente). Outros autores têm usado o TDT seguindo a alternativa de testes de homogeneidade em tabelas de contingência (Zhao et al., 2000), o que corresponde a um caso particular da análise de simetria. O problema da alta dimensionalidade do espaço das variáveis preditoras (SNPs) e de seu pequeno efeito individual sobre a doença foi abordado primeiramente usando uma estatística de somas acumuladas (CUSUM, Page, 1954) de qui-quadrados, adaptada de Millstein et al. (2006) e Guedj et al. (2006) para o caso de dados de trios. Com base nesta estatística, nas regiões de SNPs consecutivos com efeito de associação significativa, análises intervalares podem ser realizadas para encontrar as variáveis (SNPs) de maior efeito sobre a doença.

Como uma estratégia de análise de dados de trios e plataformas de SNPs, após a limpeza dos dados, um procedimento em múltiplos estágios foi proposto neste trabalho, no qual os resultados de testes de simetria em tabelas de contingência 2×2 são monitorados pelo método CUSUM, da área de controle de processos, que permite a identificação de sub-regiões do genoma em possível associação com a doença.

Estas sub-regiões são, então, exaustivamente pesquisadas por meio de análises de simetria para pares de SNPs consecutivos (cujos dados geram tabelas 4×4). Formulações teóricas do teste de simetria em tabelas 2×2 e 4×4 foram desenvolvidas e apresentadas, sob contextos assintóticos e exatos, visando a redução da função de verossimilhança em componentes de maior interesse genético. Ainda, explorando a estrutura dos dados de trios, um modelo logístico é proposto para a análise de simetria. Vale ressaltar, que a disposição dos dados de trios em tabelas de contingência 2×2 e 4×4 não é direta e pode envolver ambiguidades. Para esta finalidade existem pacotes computacionais disponíveis, como, por exemplo, o módulo `tdthap` do R (<http://www.r-project.org/>), ou o PLINK (Purcell et al. 2005), mas estes fornecem somente as caselas de frequência de interesse para o cálculo da estatística do teste. Deste modo, neste trabalho, foi desenvolvido um programa computacional para a obtenção das tabelas completas (disponível em <http://jacqueline.dema.ufc.br>).

Para ilustrar a aplicação dos procedimentos teóricos desenvolvidos analisou-se um conjunto de dados reais fornecido pelo Laboratório de Genética e Cardiologia Molecular do InCor/USP, denominado **Trios do Brasil**. Na implementação dos procedimentos foram usadas as facilidades dos programas PLINK e R. No intuito de identificar fatores de risco genéticos associados à doença (neste caso, uma cardiopatia congênita) o mapa de marcadores SNP disponível para estudo foi analisado via a metodologia em múltiplos estágios. Alguns resultados importantes foram notados. Para cada cromossomo fez-se primeiramente uma análise de simetria uniloco, por meio da qual alguns SNPs se mostraram “significantes”. Após esta análise o método CUSUM foi aplicado no intuito de caracterizar sub-regiões do genoma associadas com fatores de risco genético para a doença. Para cada cromossomo foram encontrados blocos de SNPs consecutivos significantes, por exemplo, nos cromossomos 1, 2, 3, 4, 5 e 10. Posteriormente, escolheu-se uma destas região, neste caso uma região do cromossomo 5 de maior interesse aos médicos do InCor e que tem sido alvo de muitos estudos, na qual foram realizadas análises intervalares, para pares de SNPs adjacentes (que geram dados no formato de tabelas 4×4), via teste exato e TDT generalizado. Como resultado, para estes dados, por exemplo, na análise uniloco alguns SNPs evidenciaram associação com a doença, como os SNPs **rs12517710** na posição 8703479 BP e o SNP **rs12187730**, na posição 8703923 BP. Ainda,

considerando a análise intervalar via testes exatos e TDT generalizado, foi possível identificar duas sub-regiões de associação com a doença, a região em torno dos SNPs **rs12517710** ao **rs10053640** e a região em torno dos SNPs **rs992319** ao **rs1158727**. A utilidade das alternativas de testes exatos pôde ser destacada na análise intervalar pois, dentre os 99 pares de SNPs analisados na região do cromossoma 5 para as duas hipóteses de simetria de interesse em tabelas 4×4 , o que engloba 198 situações, somente 20 situações apresentaram tamanhos amostrais apropriados à análise assintótica. Também, vários locos de SNPs somente foram identificados em associação com a doença na análise intervalar, mostrando que esta proposta é mais sensível do que a análise uniloco.

Considerando trabalhos futuros pode-se fazer uma avaliação simultânea de múltiplos locos genéticos como fatores de risco para doenças em tabelas de contingência de dimensão superior a 4×4 , ou seja, uma extensão das análises vistas neste trabalho. Outro ponto de interesse refere-se à construção das tabelas de contingência provenientes dos dados de trios. Neste trabalho, admitimos configurações específicas da distribuição dos alelos nos pares de locos, isto é, a configuração *cis*, mas existem na literatura propostas que levam em conta as possíveis ambiguidades na definição da função de verossimilhança.

Ainda como proposta futura, outros modelos estruturais de associação podem ser considerados quando se tem delineamentos com trios. Aqui se fez uma adaptação do modelo logístico para este tipo de dados, uma possível proposta alternativa seria considerar modelos log-lineares. Wan et al. (2010), considerando estudos caso-controle e efeitos de interação entre genes, obtêm uma estatística de teste de associação mais fácil de implementar computacionalmente explorando as equivalências entre as formulações dos modelos logísticos e log-lineares.

Outra questão de interesse na identificação de regiões de associação é a utilização de um procedimento CUSUM não paramétrico, por meio, por exemplo, do uso do método *bootstrap* (Efron and Tibshirani, 1994; Davison and Hinkley, 1997) para construir os limites de confiança.

Apêndice A

Genética - Conceitos Básicos e Revisão

Seguem alguns conceitos os quais podem facilitar a leitura 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 sequência de DNA, um loco genético, identificável no cromossomo em termos de localização, 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 sequências 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 influencia características hereditárias (Andrade e Pinheiro, 2002). Formas alternativas de um gene em um dado loco são chamados de **alelos**, que 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 frequência 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 Apêndice, será apresentado o coeficiente de desequilíbrio de ligação, um parâmetro importante no estudo de associação entre locos genéticos e no mapeamento de genes. Serão, também, considerados outros conceitos, como equilíbrio de Hardy-Weinberg, fração de recombinação e análise de ligação.

A.1 Equilíbrio de Hardy-Weinberg

Para um único loco genético, um conceito importante em genética é o Equilíbrio de Hardy-Weinberg (EHW), que descreve as probabilidades genotípicas em termos de probabilidades alélicas, supondo independência na combinação dos alelos paternos que definem o genótipo (vide, Figura A.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 A.1 Ilustração do Equilíbrio de Hardy-Weinberg.

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

$$p_{ij} = p_i p_j [\mathbb{1}(i = j) + 2\mathbb{1}(i < j)], \quad (\text{A.1})$$

em que $\mathbb{1}(C)$ representa a função indicadora referente ao conjunto C .

Considerando locos dialélicos, segue na Tabela A.1, a relação entre probabilidades dos alelos e genótipos sob EHW (Falconer and 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 em que se observam desvios do EHW, possivelmente, devido a erros de genotipagem.

Tabela A.1 *Relação entre probabilidades dos alelos e genótipos.*

	Alelos		Genótipos		
	A	a	AA	Aa	aa
Probabilidades	p	q	p^2	$2pq$	q^2

Na análise de associação genética de uma população, descrevem-se seus possíveis genótipos, e podem-se considerar duas abordagens de análise de dados: no nível de genótipo (indivíduo) e cromossômico (alélico). Quando se consideram os dados de marcadores na forma de genótipo (AA , Aa , aa , 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 com trios, ilustrado no Capítulo 2, podem-se utilizar estas duas análises.

A.2 Análise de ligação

Considerando pares de locos genéticos, os conceitos de ligação e desequilíbrio de ligação são bastante importantes. Na análise de ligação, estudam-se os eventos de recombinação entre dois locos, sejam eles genes, marcadores moleculares, aberrações cromossômicas, etc, em que se utiliza a fração de recombinação (denotada por θ), definida como a probabilidade de que ocorra um evento de recombinação (*crossing-over*)¹ 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.

¹ *Crossing-over*: troca recíproca entre segmentos correspondentes de cromossomos homólogos, a qual ocorre na primeira divisão da meiose.

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:

$$\text{LOD} = Z(\hat{\theta}) = \log_{10} \frac{L(\hat{\theta})}{L(0,5)}, \quad (\text{A.2})$$

em que, $L(\hat{\theta})$ 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, $H_0: \theta=0,5$. Para a maioria das aplicações L é definida em função do modelo de probabilidades Binomial em que, 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 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, $4,6 \times Z(\hat{\theta})$ segue aproximadamente uma distribuição qui-quadrado com 1 grau de liberdade, $\chi^2_{(1)}$. O valor crítico LOD=3, o qual tem sido adotado como um critério de decisão, está associado a um nível descritivo² menor do que 10^{-4} .

Pode ser estabelecida uma relação entre a fração de recombinação (θ) e distância citogenética (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) e Ott (1991).

A.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

² O nível descritivo associado a LOD=3 é igual a 0,0002016645.

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

$$D = p_{A_i B_k} - p_{A_i} p_{B_k} = \text{COV}(\mathbf{1}_{A_i}, \mathbf{1}_{B_k}), \quad (\text{A.3})$$

em que, $p_{A_i B_k}$ representa a probabilidade do haplótipo $A_i B_k$; 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 A.2), têm-se os alelos A , a no loco A , os alelos B e b no loco B , e os possíveis haplótipos AB , Ab , aB e ab . Considerando o haplótipo AB , a medida de desequilíbrio gamético é dada por:

$$D_{AB} = p_{AB} - p_A p_B, \quad (\text{A.4})$$

em que, $D_{AB} \in [D_{\min}, D_{\max}]$, tal que, $D_{\min} = -\max[p_A p_B, (1-p_A)(1-p_B)]$ e $D_{\max} = \min[p_A(1-p_B), (1-p_A)p_B]$, 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 A.2 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,

$$D = \text{Cov}[\mathbf{1}(A), \mathbf{1}(B)], \quad (\text{A.5})$$

em que:

$\text{Cov}[\mathbb{1}(A), \mathbb{1}(B)]$ é a covariância entre as variáveis $\mathbb{1}(A)$ e $\mathbb{1}(B)$, que é dado por $\text{Cov}[\mathbb{1}(A), \mathbb{1}(B)] = E[\mathbb{1}(A)\mathbb{1}(B)] - E[\mathbb{1}(A)]E[\mathbb{1}(B)]$ e

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

Tem-se:

$$E[\mathbb{1}(A)] = p(\mathbb{1}(A) = 1) = p_A,$$

$$E[\mathbb{1}(B)] = p(\mathbb{1}(B) = 1) = p_B \text{ e}$$

$$E[\mathbb{1}(A)\mathbb{1}(B)] = p(\mathbb{1}(A) = 1, \mathbb{1}(B) = 1) = p_{AB}.$$

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 A.3 mostra o decréscimo do desequilíbrio de ligação por gerações para diferentes frações de recombinação entre os locos, o desequilíbrio torna-se muito pequeno com o número de gerações e/ou torna-se grande com a fração de recombinação (Jorde, 1995).

Figura A.3 *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.

Apêndice B

Demonstrações de Algumas Expressões

B.1 Demonstrações referentes às expressões do Capítulo 3, Seção 3.2.2

$$P(\mathbf{n}|n_{..}, \boldsymbol{\pi}) = P(n_{11}, n_{12}, n_{13}, n_{14}, \dots, n_{44}|n_{..}, \boldsymbol{\pi})$$

$$P(\mathbf{n}|n_{..}, \boldsymbol{\pi}) = P[(n_{12}, n_{21}), (n_{13}, n_{31}), (n_{14}, n_{41}), (n_{23}, n_{32}), (n_{24}, n_{42}), (n_{34}, n_{43}), n_{1.}, n_{2.}, n_{3.}|n_{..}, \boldsymbol{\pi}] =$$

$$= P(n_{11}, n_{22}, n_{33}, n_{12}, n_{13}, n_{14}, n_{23}, n_{24}, n_{34}, n_{12} + n_{21}, n_{13} + n_{31}, n_{14} + n_{41}, n_{23} + n_{32}, n_{24} + n_{42}, n_{34} +$$

$$n_{43}|n_{..}, \boldsymbol{\pi}) = P(n_{12} + n_{21}|n_{..}, \boldsymbol{\pi})P(n_{11}|n_{12} + n_{21}, n_{..}, \boldsymbol{\pi})P(n_{12}|n_{11}, n_{12} + n_{21}, n_{..}, \boldsymbol{\pi})$$

$$P(n_{13} + n_{31}|n_{11}, n_{12}, n_{12} + n_{21}, n_{..}, \boldsymbol{\pi})P(n_{13}|n_{11}, n_{12}, n_{12} + n_{21}, n_{13} + n_{31}, n_{..}, \boldsymbol{\pi})$$

$$P(n_{14} + n_{41}|n_{11}, n_{12}, n_{13}, n_{12} + n_{21}, n_{13} + n_{31}, n_{..}, \boldsymbol{\pi})$$

$$P(n_{14}|n_{11}, n_{12}, n_{13}, n_{12} + n_{21}, n_{13} + n_{31}, n_{14} + n_{41}, n_{..}, \boldsymbol{\pi})$$

$$P(n_{22}|n_{11}, n_{12}, n_{13}, n_{14}, n_{12} + n_{21}, n_{13} + n_{31}, n_{14} + n_{41}, n_{..}, \boldsymbol{\pi})$$

$$P(n_{23} + n_{32}|n_{11}, n_{12}, n_{13}, n_{14}, n_{22}, n_{12} + n_{21}, n_{13} + n_{31}, n_{14} + n_{41}, n_{..}, \boldsymbol{\pi})$$

$$P(n_{23}|n_{11}, n_{12}, n_{13}, n_{14}, n_{22}, n_{12} + n_{21}, n_{13} + n_{31}, n_{14} + n_{41}, n_{23} + n_{32}, n_{..}, \boldsymbol{\pi})$$

$$P(n_{24} + n_{42}|n_{11}, n_{12}, n_{13}, n_{14}, n_{22}, n_{23}, n_{12} + n_{21}, n_{13} + n_{31}, n_{14} + n_{41}, n_{23} + n_{32}, n_{..}, \boldsymbol{\pi})$$

$$P(n_{24}|n_{11}, n_{12}, n_{13}, n_{14}, n_{22}, n_{23}, n_{12} + n_{21}, n_{13} + n_{31}, n_{14} + n_{41}, n_{23} + n_{32}, n_{24} + n_{42}, n_{..}, \boldsymbol{\pi})$$

$$P(n_{33}|n_{11}, n_{12}, n_{13}, n_{14}, n_{22}, n_{23}, n_{24}, n_{12} + n_{21}, n_{13} + n_{31}, n_{14} + n_{41}, n_{23} + n_{32}, n_{24} + n_{42}, n_{..}, \boldsymbol{\pi})$$

$$P(n_{34} + n_{43}|n_{11}, n_{12}, n_{13}, n_{14}, n_{22}, n_{23}, n_{24}, n_{33}, n_{12} + n_{21}, n_{13} + n_{31}, n_{14} + n_{41}, n_{23} + n_{32}, n_{24} + n_{42}, n_{..}, \boldsymbol{\pi})$$

$$P(n_{34}|n_{11}, n_{12}, n_{13}, n_{14}, n_{22}, n_{23}, n_{24}, n_{33}, n_{12} + n_{21}, n_{13} + n_{31}, n_{14} + n_{41}, n_{23} + n_{32}, n_{24} + n_{42}, n_{34} + n_{43}, n_{..}, \boldsymbol{\pi})$$

Considere a seguinte notação para as correspondentes funções de verossimilhança associadas às expressões anteriores:

$$L(\boldsymbol{\pi}) = L_1(\pi_{12} + \pi_{21})$$

$$L_2(\pi_{11}, \pi_{12}, \pi_{21})$$

$$L_3(\pi_{12}, \pi_{21})$$

$$L_4(\pi_{11}, \pi_{12}, \pi_{21}, \pi_{13}, \pi_{31})$$

$$L_5(\pi_{13}, \pi_{31})$$

$$L_6(\pi_{11}, \pi_{12}, \pi_{21}, \pi_{13}, \pi_{31}, \pi_{14}, \pi_{41})$$

$$L_7(\pi_{14}, \pi_{41})$$

$$L_8(\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}, \pi_{13}, \pi_{31}, \pi_{14}, \pi_{41})$$

$$L_9(\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}, \pi_{13}, \pi_{31}, \pi_{14}, \pi_{41}, \pi_{23}, \pi_{32})$$

$$L_{10}(\pi_{23}, \pi_{32})$$

$$L_{11}(\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}, \pi_{13}, \pi_{31}, \pi_{14}, \pi_{41}, \pi_{23}, \pi_{32}, \pi_{24}, \pi_{42})$$

$$L_{12}(\pi_{24}, \pi_{42})L_{13}(\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}, \pi_{13}, \pi_{31}, \pi_{33}, \pi_{14}, \pi_{41}, \pi_{23}, \pi_{32}, \pi_{24}, \pi_{42})$$

$$L_{14}(\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}, \pi_{13}, \pi_{31}, \pi_{33}, \pi_{14}, \pi_{41}, \pi_{23}, \pi_{32}, \pi_{24}, \pi_{42}, \pi_{34}, \pi_{43})$$

$$L_{15}(\pi_{34}, \pi_{43}).$$

Observações:

I) As demonstrações para as verossimilhanças $L_4(\pi_{11}, \pi_{12}, \pi_{21}, \pi_{13}, \pi_{31})$, $L_6(\pi_{11}, \pi_{12}, \pi_{21}, \pi_{13}, \pi_{31}, \pi_{14}, \pi_{41})$,

$L_8(\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}, \pi_{13}, \pi_{31}, \pi_{14}, \pi_{41})$, $L_9(\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}, \pi_{13}, \pi_{31}, \pi_{14}, \pi_{41}, \pi_{23}, \pi_{32})$,

$L_{11}(\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}, \pi_{13}, \pi_{31}, \pi_{14}, \pi_{41}, \pi_{23}, \pi_{32}, \pi_{24}, \pi_{42})$,

$L_{13}(\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}, \pi_{13}, \pi_{31}, \pi_{33}, \pi_{14}, \pi_{41}, \pi_{23}, \pi_{32}, \pi_{24}, \pi_{42})$ e

$L_{14}(\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}, \pi_{13}, \pi_{31}, \pi_{33}, \pi_{14}, \pi_{41}, \pi_{23}, \pi_{32}, \pi_{24}, \pi_{42}, \pi_{34}, \pi_{43})$ seguem a mesma formulação da verossimilhança $L_2(\pi_{11}, \pi_{12}, \pi_{21})$.

II) As demonstrações para as verossimilhanças $L_7(\pi_{14}, \pi_{41})$, $L_{10}(\pi_{23}, \pi_{32})$, $L_{12}(\pi_{24}, \pi_{42})$ e $L_{15}(\pi_{34}, \pi_{43})$ seguem a mesma formulação como nas verossimilhanças $L_3(\pi_{12}, \pi_{21})$ e $L_5(\pi_{13}, \pi_{31})$.

- Seguem as distribuições para cada verossimilhança.

1) $L_1(\pi_{12} + \pi_{21})$:

$$n_{12} + n_{21} | n_{..}, \boldsymbol{\pi} \sim \text{Bin}(n_{..} - (n_{12} + n_{21}), \pi_{12} + \pi_{21})$$

2) $L_2(\pi_{11}, \pi_{12}, \pi_{21})$:

$$n_{11} | n_{12} + n_{21}, n_{..}, \boldsymbol{\pi}$$

$$P(n_{11} | n_{12} + n_{21}) = \frac{P(n_{11}, n_{12} + n_{21})}{P(n_{12} + n_{21})} =$$

$$= \frac{\frac{n_{..}! \pi_{11}^{n_{11}} (\pi_{12} + \pi_{21})^{n_{12} + n_{21}} (1 - \pi_{11} - \pi_{12} - \pi_{21})^{n_{..} - n_{11} - n_{12} - n_{21}}}{n_{11}! (n_{12} + n_{21})! (n_{..} - n_{11} - n_{12} - n_{21})!}}{\frac{n_{..}! (\pi_{12} + \pi_{21})^{n_{12} + n_{21}} (1 - \pi_{12} - \pi_{21})^{n_{..} - n_{12} - n_{21}}}{(n_{12} + n_{21})! (n_{..} - n_{12} - n_{21})!}}$$

$$= \frac{\pi_{11} (1 - \pi_{11} - \pi_{12} - \pi_{21})^{n_{..} - n_{11} - n_{12} - n_{21}} (n_{..} - n_{12} - n_{21})!}{n_{11}! (n_{..} - n_{11} - n_{12} - n_{21})! (1 - \pi_{12} - \pi_{21})^{n_{..} - n_{12} - n_{21} - n_{11} + n_{11}}}$$

$$= \frac{(n_{..} - n_{12} - n_{21})!}{n_{11}! (n_{..} - n_{11} - n_{12} - n_{21})!} \left(\frac{\pi_{11}}{1 - \pi_{12} - \pi_{21}} \right)^{n_{11}} \left(\frac{1 - \pi_{11} - \pi_{12} - \pi_{21}}{1 - \pi_{12} - \pi_{21}} \right)^{n_{..} - n_{11} - n_{12} - n_{21}}$$

Portanto,

$$n_{11} | n_{12} + n_{21}, n_{..}, \boldsymbol{\pi} \sim \text{Bin}\left(n_{..} - (n_{12} + n_{21}), \frac{\pi_{11}}{1 - \pi_{12} - \pi_{21}}\right)$$

3) $L_3(\pi_{12}, \pi_{21})$:

$$n_{12} | n_{11}, n_{12} + n_{21}, n_{..}, \boldsymbol{\pi}$$

Seja $n_{12} = x$ e $n_{12} + n_{21} = s$.

$$P(n_{12} | n_{11}, n_{12} + n_{21}) = \frac{P(n_{12}, n_{11}, n_{12} + n_{21})}{P(n_{11}, n_{12} + n_{21})} =$$

$$= \frac{P(n_{12}=x, n_{11}, n_{12}+n_{21}=s)}{P(n_{11}, n_{12}+n_{21}=s)} = \frac{P(n_{12}=x, n_{11}, n_{21}=s-x)}{P(n_{11}, n_{21}=s-x)} =$$

$$= \frac{\frac{n_{..}! \pi_{12}^x \pi_{21}^{s-x} \pi_{11}^{n_{11}} (1 - \pi_{12} - \pi_{21} - \pi_{11})^{n_{..} - x - s + x - n_{11}}}{n_{12}! (s-x)! n_{11}! (n_{..} - n_{12} - s + x - n_{11})!}}{\frac{n_{..}! (\pi_{12} + \pi_{21})^s \pi_{11}^{n_{11}} (1 - \pi_{12} - \pi_{21} - \pi_{11})^{n_{..} - s - n_{11}}}{s! n_{11}! (n_{..} - n_{12} - n_{21} - n_{11})!}} = \frac{s! \pi_{12}^x \pi_{21}^{s-x}}{n_{12}! (s-x)! (\pi_{12} + \pi_{21})^{s+x-x}} =$$

$$\begin{aligned}
&= \binom{s}{x} \left(\frac{\pi_{12}}{\pi_{12} + \pi_{21}} \right)^x \left(1 - \frac{\pi_{12}}{\pi_{12} + \pi_{21}} \right)^{s-x} = \\
&= \binom{n_{12} + n_{21}}{n_{12}} \left(\frac{\pi_{12}}{\pi_{12} + \pi_{21}} \right)^{n_{12}} \left(1 - \frac{\pi_{12}}{\pi_{12} + \pi_{21}} \right)^{n_{21}}
\end{aligned}$$

Portanto,

$$n_{12}|n_{11}, n_{12} + n_{21}, n_{..}, \boldsymbol{\pi} \sim \text{Bin}\left(n_{12} + n_{21}, \frac{\pi_{12}}{\pi_{12} + \pi_{21}}\right)$$

4) $L_4(\pi_{11}, \pi_{12}, \pi_{21}, \pi_{13}, \pi_{31})$:

$$n_{13} + n_{31}|n_{11}, n_{12}, n_{12} + n_{21}, n_{..}, \boldsymbol{\pi} \sim \text{Bin}\left(n_{..} - n_{11} - n_{12} - n_{21}, \frac{\pi_{13} + \pi_{31}}{1 - \pi_{11} - \pi_{12} - \pi_{21}}\right)$$

5) $L_5(\pi_{13}, \pi_{31})$:

$$n_{13}|n_{11}, n_{12}, n_{12} + n_{21}, n_{13} + n_{31}, n_{..}, \boldsymbol{\pi}$$

Seja $n_{12} = x$, $n_{13} = g$, $n_{12} + n_{21} = s$ e $n_{13} + n_{31} = y$.

$$P(n_{13}|n_{11}, n_{12}, n_{12} + n_{21}, n_{13} + n_{31}) = \frac{P(n_{13}, n_{11}, n_{12}, n_{12} + n_{21}, n_{13} + n_{31})}{P(n_{11}, n_{12}, n_{12} + n_{21}, n_{13} + n_{31})} =$$

$$\begin{aligned}
&= \frac{P(n_{13}=g, n_{11}, n_{12}=x, n_{12}+n_{21}=s, n_{13}+n_{31}=y)}{P(n_{11}, n_{12}=x, n_{12}+n_{21}=s, n_{13}+n_{31}=y)} = \frac{P(n_{13}=g, n_{11}, n_{12}=x, n_{21}=s-x, n_{31}=y-g)}{P(n_{11}, n_{12}=x, n_{21}=s-x, n_{13}+n_{31}=y)} = \\
&= \frac{\frac{n_{..}! \pi_{13}^g \pi_{11}^{n_{11}} \pi_{12}^x \pi_{21}^{s-x} \pi_{31}^{y-g} (1 - \pi_{11} - \pi_{12} - \pi_{21} - \pi_{13} - \pi_{31})^{n_{..} - g - n_{11} - x - s + x - y + g}}{g! n_{11}! x! (s-x)! (y-g)! (n_{..} - g - n_{11} - x - s + x - y + g)!}}{\frac{n_{..}! \pi_{11}^{n_{11}} \pi_{12}^x \pi_{21}^{s-x} (\pi_{13} + \pi_{31})^y (1 - \pi_{11} - \pi_{12} - \pi_{21} - \pi_{13} - \pi_{31})^{n_{..} - n_{11} - x - s + x - y}}{n_{11}! x! (s-x)! y! (n_{..} - n_{11} - x - s + x - y)!}} = \\
&= \frac{y! \pi_{13}^g \pi_{31}^{y-g}}{g! (y-g)! (\pi_{13} + \pi_{31})^{y+g-g}} = \\
&= \binom{y}{g} \left(\frac{\pi_{13}}{\pi_{13} + \pi_{31}} \right)^g \left(1 - \frac{\pi_{13}}{\pi_{13} + \pi_{31}} \right)^{y-g} = \\
&= \binom{n_{13} + n_{31}}{n_{13}} \left(\frac{\pi_{13}}{\pi_{13} + \pi_{31}} \right)^{n_{13}} \left(1 - \frac{\pi_{13}}{\pi_{13} + \pi_{31}} \right)^{n_{31}}
\end{aligned}$$

Portanto,

$$n_{13}|n_{11}, n_{12}, n_{12} + n_{21}, n_{13} + n_{31}, n_{..}, \boldsymbol{\pi} \sim \text{Bin}\left(n_{13} + n_{31}, \frac{\pi_{13}}{\pi_{13} + \pi_{31}}\right)$$

6) $L_6(\pi_{11}, \pi_{12}, \pi_{21}, \pi_{13}, \pi_{31}, \pi_{14}, \pi_{41})$:

$$n_{14}+n_{41}|n_{11}, n_{12}, n_{13}, n_{12}+n_{21}, n_{13}+n_{31}, n_{..}, \boldsymbol{\pi} \sim \text{Bin}\left(n_{..} - n_{11} - n_{12} - n_{21} - n_{13} - n_{31}, \frac{\pi_{14}+\pi_{41}}{1-\pi_{11}-\pi_{12}-\pi_{21}-\pi_{13}-\pi_{31}}\right)$$

7) $L_7(\pi_{14}, \pi_{41})$:

$$n_{14}|n_{11}, n_{12}, n_{13}, n_{12} + n_{21}, n_{13} + n_{31}, n_{14} + n_{41}, n_{..}, \boldsymbol{\pi} \sim \text{Bin}\left(n_{14} + n_{41}, \frac{\pi_{14}}{\pi_{14}+\pi_{41}}\right)$$

8) $L_8(\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}, \pi_{13}, \pi_{31}, \pi_{14}, \pi_{41})$:

$$n_{22}|n_{11}, n_{12}, n_{13}, n_{14}, n_{12} + n_{21}, n_{13} + n_{31}, n_{14} + n_{41}, n_{..}, \boldsymbol{\pi} \\ \sim \text{Bin}\left(n_{..} - n_{11} - n_{12} - n_{21} - n_{13} - n_{31} - n_{14} - n_{41}, \frac{\pi_{22}}{1-\pi_{11}-\pi_{12}-\pi_{21}-\pi_{13}-\pi_{31}-\pi_{14}-\pi_{41}}\right)$$

9) $L_9(\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}, \pi_{13}, \pi_{31}, \pi_{14}, \pi_{41}, \pi_{23}, \pi_{32})$:

$$n_{23} + n_{32}|n_{11}, n_{12}, n_{13}, n_{14}, n_{22}, n_{12} + n_{21}, n_{13} + n_{31}, n_{14} + n_{41}, n_{..}, \boldsymbol{\pi} \\ \sim \text{Bin}\left(n_{..} - n_{11} - n_{12} - n_{21} - n_{13} - n_{31} - n_{14} - n_{41} - n_{22}, \frac{\pi_{23}+\pi_{32}}{1-\pi_{11}-\pi_{12}-\pi_{21}-\pi_{13}-\pi_{31}-\pi_{14}-\pi_{41}-\pi_{22}}\right)$$

10) $L_{10}(\pi_{23}, \pi_{32})$:

$$n_{23}|n_{11}, n_{12}, n_{13}, n_{14}, n_{22}, n_{12} + n_{21}, n_{13} + n_{31}, n_{14} + n_{41}, n_{23} + n_{32}, n_{..}, \boldsymbol{\pi} \sim \text{Bin}\left(n_{23} + n_{32}, \frac{\pi_{23}}{\pi_{23}+\pi_{32}}\right)$$

11) $L_{11}(\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}, \pi_{13}, \pi_{31}, \pi_{14}, \pi_{41}, \pi_{23}, \pi_{32}, \pi_{24}, \pi_{42})$:

$$n_{24} + n_{42}|n_{11}, n_{12}, n_{13}, n_{14}, n_{22}, n_{23}, n_{12} + n_{21}, n_{13} + n_{31}, n_{14} + n_{41}, n_{23} + n_{32}, n_{..}, \boldsymbol{\pi} \\ \sim \text{Bin}\left(n_{..} - n_{11} - n_{12} - n_{21} - n_{13} - n_{31} - n_{14} - n_{41} - n_{23} - n_{32} - n_{22}, \frac{\pi_{24}+\pi_{42}}{1-\pi_{11}-\pi_{12}-\pi_{21}-\pi_{13}-\pi_{31}-\pi_{14}-\pi_{41}-\pi_{23}-\pi_{32}-\pi_{22}}\right)$$

12) $L_{12}(\pi_{24}, \pi_{42})$:

$$n_{24}|n_{11}, n_{12}, n_{13}, n_{14}, n_{22}, n_{23}, n_{12}+n_{21}, n_{13}+n_{31}, n_{14}+n_{41}, n_{23}+n_{32}, n_{24}+n_{42}, n_{..}, \boldsymbol{\pi} \sim \text{Bin}\left(n_{24} + n_{42}, \frac{\pi_{24}}{\pi_{24}+\pi_{42}}\right)$$

13) $L_{13}(\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}, \pi_{13}, \pi_{31}, \pi_{33}, \pi_{14}, \pi_{41}, \pi_{23}, \pi_{32}, \pi_{24}, \pi_{42})$:

$$n_{33}|n_{11}, n_{12}, n_{13}, n_{14}, n_{22}, n_{23}, n_{24}, n_{12} + n_{21}, n_{13} + n_{31}, n_{14} + n_{41}, n_{23} + n_{32}, n_{24} + n_{42}, n_{..}, \boldsymbol{\pi} \sim \\ \text{Bin}(n^*, \pi^*)$$

em que:

$$n^* = n_{..} - n_{11} - n_{12} - n_{21} - n_{13} - n_{31} - n_{14} - n_{41} - n_{23} - n_{32} - n_{24} - n_{42} - n_{22} \text{ e}$$

$$\pi^* = \frac{\pi_{33}}{1 - \pi_{11} - \pi_{12} - \pi_{21} - \pi_{13} - \pi_{31} - \pi_{14} - \pi_{41} - \pi_{23} - \pi_{32} - \pi_{24} - \pi_{42} - \pi_{22}}$$

$$\mathbf{14)} \text{ L}_{14}(\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}, \pi_{13}, \pi_{31}, \pi_{33}, \pi_{14}, \pi_{41}, \pi_{23}, \pi_{32}, \pi_{24}, \pi_{42}, \pi_{34}, \pi_{43}):$$

$$n_{34} + n_{43} | n_{11}, n_{12}, n_{13}, n_{14}, n_{22}, n_{23}, n_{24}, n_{33}, n_{12} + n_{21}, n_{13} + n_{31}, n_{14} + n_{41}, n_{23} + n_{32}, n_{24} + n_{42}, n_{..}, \boldsymbol{\pi} \sim \text{Bin}(n^{**}, \pi^{**})$$

em que:

$$n^{**} = n_{..} - n_{11} - n_{12} - n_{21} - n_{13} - n_{31} - n_{14} - n_{41} - n_{23} - n_{32} - n_{24} - n_{42} - n_{22} - n_{33}$$

$$\pi^{**} = \frac{\pi_{34} - \pi_{43}}{1 - \pi_{11} - \pi_{12} - \pi_{21} - \pi_{13} - \pi_{31} - \pi_{14} - \pi_{41} - \pi_{23} - \pi_{32} - \pi_{24} - \pi_{42} - \pi_{22} - \pi_{33}}$$

$$\mathbf{15)} \text{ L}_{15}(\pi_{34}, \pi_{43}):$$

$$n_{34} | n_{11}, n_{12}, n_{13}, n_{14}, n_{22}, n_{23}, n_{24}, n_{33}, n_{12} + n_{21}, n_{13} + n_{31}, n_{14} + n_{41}, n_{23} + n_{32}, n_{24} + n_{42}, n_{34} + n_{43}, n_{..}, \boldsymbol{\pi} \sim \text{Bin}\left(n_{34} + n_{43}, \frac{\pi_{34}}{\pi_{34} + \pi_{43}}\right)$$

Apêndice C

Rotinas Computacionais

C.1 Programas PLINK e R.

Dados dos TRIOS DO BRASIL.

O símbolo # refere-se a comentários.

```
#Exemplo do cálculo da estatística TDT no programa PLINK
```

```
#http://pngu.mgh.harvard.edu/purcell/plink/
```

```
#Teste TDT
```

```
#Primeiramente, ler os arquivos BRZ.ped e BRZ.map e construir tres arquivos:
```

```
#fam (as primeiras tres colunas de .ped),
```

```
#bim (2 colunas com o nome do alelo para cada SNP) e
```

```
#bed (dados brutos)
```

```
plink --file BRZ --make-bed --out BRZ1
```

```
#lendo o arquivo de leitura BRZ1 e salvando os resultados no arquivo resultdt
```

```
plink --bfile BRZ1 --tdt --out resutdt
```

```
*****
```

```
#Exemplo de carta CUSUM no programa R
```

```
#Exemplo para o Cromossomo 5
```

```
crom5<-read.table("C:/PLINK/resultdt2c5.txt", sep=",", dec=".", h=T)
```

```
#Inicio carta CUSUM
```

```
library(qcc)
```

```
cart <- qcc(crom5, type="xbar.one", nsigmas=3, plot=FALSE)
```

```
cusum(cart)
```

```
*****
```

C.2 Regiões candidatas (Método CUSUM)

*****Cromossomo 1*****

***** 1a. Região *****

SNP	BP
rs10127961	73559057
rs11210209	73565833
rs1885246	73566087
rs2208565	73582817
rs10789368	73586747
rs1160682	73589152
rs2340400	73593869
rs10890034	73596867
rs17568589	73597143
rs10789369	73597497
rs1923227	73619933
rs17568889	73620476
rs7546663	73620717
rs11210220	73622243
rs11210222	73622275
rs1923229	73622471
rs1923236	73626414
rs12044079	73626735
rs11210226	73626872
rs10518420	73633131
rs1885251	73638608
rs12035848	73643489
rs7543202	73645473
rs10465868	73653242
rs11210242	73670397
rs11210243	73670618
rs4350155	73684008
rs11210251	73685387
rs11210252	73692261
rs647959	73692513
rs647478	73692628

rs505725	73702993
rs12138620	73703212
rs11210255	73704325
rs12140865	73705474
rs12754690	73712486
rs6701520	73713306
rs9425120	73724103
rs7534840	73748346
rs7545661	73749076
rs12139989	73756917

***** 2a. Região *****

SNP	BP
rs17842210	165828994
rs6701488	165829440
rs6666091	165829636
rs6677532	165829696
rs6702092	165830039
rs12126826	165831285
rs2995081	165835065
rs11810018	165835646
rs2143305	165838785
rs4657679	165839182
rs6693156	165839289
rs6427095	165840614
rs4657686	165845545
rs1013057	165846973
rs6698572	165849887
rs926521	165859156
rs1229390	165873964

***** 3a. Região *****

SNP	BP
rs12726603	192908673
rs4253903	192914811
rs12141663	192917069
rs6700138	192926331

rs7520494 192926805
rs16837328 192931801

*****Cromossomo 2*****

***** 1a. Região *****

SNP	BP
rs6734278	26055737
rs6719890	26060347
rs12713741	26060925
rs12474530	26063001
rs12990137	26077051
rs1992300	26079784
rs7600349	26083183
rs7604184	26084366
rs6734513	26086025
rs4665809	26104865
rs6749855	26108767
rs7599132	26133314
rs11677219	26134743
rs1560868	26134963
rs4233709	26144997
rs6546736	26147382
rs10048683	26174023
rs6742226	26196041

***** 2a. Região *****

SNP	BP
rs6713572	75621585
rs10489977	75621606

***** 3a. Região *****

SNP	BP
rs6734682	193164502
rs1515349	193166824
rs10497732	193168479

rs1946877 193237081
rs1898917 193247065
rs13394988 193250072
rs1439664 193267259
rs1439663 193267387
rs1595720 193275937
rs10460318 193297208
rs934952 193299876
rs11687813 193301037
rs6730654 193311175
rs6730696 193311341
rs11686149 193337289
rs16834896 193342309
rs1439652 193347489
rs1439651 193347787
rs16834909 193347825
rs7567586 193348325
rs13419834 193348721
rs1125174 193384011
rs7598135 193386259
rs11894222 193389622
rs1596754 193401604
rs12328398 193404116
rs13003012 193414565
rs4850795 193421595
rs10197088 193432388
rs10179416 193437742
rs7602191 193440726
rs4850801 193441846
rs1373900 193444627
rs12465720 193449506
rs13014947 193451244
rs12617870 193454528
rs1979001 193457550
rs9288210 193473143
rs41376050 193474374
rs1596751 193516105

rs13011989 193521670
rs12995239 193525931
rs10497737 193532430
rs1979002 193547174
rs7574968 193554359
rs6750646 193580660
rs13001157 193611095
rs11679751 193625384
rs2357537 193662703
rs7577834 193664726
rs973885 193702523
rs1379470 193703561

*****Cromossomo 3*****

***** 1a. Região *****

SNP	BP
rs2881514	16528684
rs2346910	16528828
rs4685343	16528973
rs4684284	16535352
rs1529513	16538010
rs17042365	16540563

***** 2a. Região *****

SNP	BP
rs4683748	144345809
rs1900994	144345909

*****Cromossomo 4*****

***** 1a. Região *****

SNP	BP
rs509863	13579007
rs13136618	13579355
rs13113093	13579383
rs13138372	13579511

rs6857639	13580952
rs1444153	13588851
rs1795569	13589297
rs573996	13596832
rs7665661	13606387
rs544883	13607074
rs543907	13607197
rs543173	13607245
rs16889552	13611753
rs6839527	13617397
rs9291619	13617958
rs3843422	13619712
rs557282	13626858
rs7666675	13627492
rs477315	13635291
rs493198	13635806
rs480574	13638629
rs565463	13642067

***** 2a. Região *****

SNP	BP
rs1048329	186536752

*****Cromossomo 5*****

***** 1a. Região *****

SNP	BP
rs12517710	8703479
rs12187730	8703923
rs736970	8704432
rs7704554	8705400
rs200114	8705778
rs10076745	8706313
rs10053640	8706345
rs199195	8708624
rs200107	8708690
rs199194	8709986

rs200091	8715504
rs200090	8715681
rs6882860	8717747
rs200077	8719535
rs999419	8724855
rs200056	8734335
rs200055	8734628
rs200050	8738784
rs10060493	8740137
rs200044	8741962
rs200012	8748751
rs200001	8752765
rs199999	8753487

***** 2a. Região *****

SNP	BP
rs4301211	83659636
rs2133768	83664041
rs7720807	83668447
rs41509548	83675518
rs305653	83682567
rs10063982	83696884
rs10073440	83698401
rs305661	83716793
rs12659469	83718035
rs2941228	83722161

*****Cromossomo 6*****

***** 1a. Região *****

SNP	BP
rs4053720	87176209

***** 2a. Região *****

SNP	BP
rs2076511	52246138
rs2235714	52246185

rs6918669 52246400

***** 3a. Região *****

SNP	BP
rs7765167	115801529
rs4624905	115805692
rs4563753	115805961
rs9374538	115813397
rs9372436	115851140
rs12203679	115853725

***** 4a. Região *****

SNP	BP
rs12194182	160754505
rs2457572	160756396
rs3123636	160762527
rs2504929	160767976
rs2665357	160768157
rs9457925	160768733
rs3106164	160770263
rs376563	160771756
rs9456538	160773383
rs2457577	160774128
rs1567442	160777305
rs2292334	160778178
rs2457012	160778961

***** 5a. Região *****

SNP	BP
rs7773552	169855220
rs9478098	169864628
rs9478099	169872172
rs6942332	169873859
rs6459659	169900793
rs11966349	169904462
rs914968	169905116
rs9371171	169909543

rs9295024	169910367
rs6459664	169922532
rs7739088	169926308
rs6906751	169927392
rs3823466	169931851

*****Cromossomo 7*****

***** 1a. Região *****

SNP	BP
rs6963566	79537747
rs4731190	79538939

*****Cromossomo 8*****

***** 1a. Região *****

SNP	BP
rs831739	70313106
rs1809311	70314533
rs2655173	70318471
rs2655184	70325996
rs2615100	70337030
rs283323	70342532
rs283324	70342805
rs831728	70346885
rs283345	70352866
rs283344	70353003

*****Cromossomo 10*****

***** 1a. Região *****

SNP	BP
rs12264028	59948146
rs1649023	59955826
rs1658469	59962408
rs1649080	59963300
rs1303970	59968483

rs1619284	59972521
rs1649069	59973701
rs1658456	59974332
rs1649060	59980486
rs1649057	59988120
rs1649051	59992003
rs1649048	59994288
rs1658438	59996589
rs1649039	60000047
rs1649035	60002707
rs11006187	60003389
rs6481407	60005096
rs10763556	60011404
rs10763558	60011940
rs7899985	60030194
rs930368	60037777
rs7097457	60040147
rs17694070	60040878
rs10826205	60041612
rs7080572	60046825
rs7095923	60046958
rs11006208	60052411
rs10740740	60052500
rs4948528	60057978
rs2028205	60059007
rs7904807	60061588
rs12356237	60062443
rs920259	60066071
rs12569625	60087263

***** 2a. Região *****

SNP	BP
rs7904383	117465896
rs10510011	117468540

*****Cromossomo 12*****

***** 1a. Região *****

SNP	BP
rs11835989	90848884
rs7970039	90849044

*****Cromossomo 14*****

***** 1a. Região *****

SNP	BP
rs10150397	99091951
rs10150747	99092101
rs10150666	99092159
rs10150754	99092217
rs2144809	99092617
rs4445835	99093149
rs11845767	99094055
rs2400744	99096470
rs8015109	99097733
rs8016105	99100288
rs8004421	99105278
rs941558	99112092
rs8022133	99117244
rs17098856	99117470
rs17098860	99118556

*****Cromossomo 14*****

***** 1a. Região *****

SNP	BP
rs10150397	99091951
rs10150747	99092101
rs10150666	99092159
rs10150754	99092217
rs2144809	99092617
rs4445835	99093149
rs11845767	99094055
rs2400744	99096470

rs8015109	99097733
rs8016105	99100288
rs8004421	99105278
rs941558	99112092
rs8022133	99117244
rs17098856	99117470
rs17098860	99118556

*****Cromossomo 15*****

***** 1a. Região *****

SNP	BP
rs12439925	29386793
rs7175141	29396838
rs1465779	29397182
rs11070938	29397406
rs11070943	29402961
rs9920754	29416590
rs11853926	29417314
rs4779861	29417644
rs12439853	29422035
rs7174211	29425288
rs3803554	29430281
rs1807100	29432224
rs8036899	29434408
rs7168479	29438317
rs7174079	29438360
rs12592541	29439088
rs12915988	29440287
rs8031093	29442500
rs4268714	29462745
rs8042404	29467308
rs7169573	29484659
rs6493657	29487132
rs8035633	29487314
rs11630485	29492339
rs4354896	29494067

rs11071016	29510021
rs6493668	29532660
rs12592652	29534700
rs12592342	29536628
rs4779528	29537170
rs2219508	29538299
rs8038667	29541098
rs7403275	29541300
rs7403458	29541317
rs17228317	29544125
rs11071070	29545453
rs2125617	29548278
rs8035018	29549369
rs8023660	29549752
rs8023664	29549766
rs11071075	29550901
rs898213	29551070
rs2338684	29551691
rs1077168	29551703
rs1017720	29555821
rs4779532	29559726
rs1824354	29572022
rs17816007	29574720
rs7178180	29578755
rs16956801	29579883
rs16956805	29581980
rs7181743	29596545
rs16956820	29597533
rs4779889	29601495
rs7173874	29601759
rs16956838	29606987
rs17228394	29608201
rs7176771	29610400
rs7176903	29611068
rs12915549	29614490
rs2125620	29616176
rs12908729	29618031

rs10152711	29618492
rs17816055	29619386
rs12900301	29619936
rs12591927	29621203
rs4365252	29629663
rs4779892	29630334

*****Cromossomo 17*****

***** 1a. Região *****

SNP	BP
rs1619833	29758813
rs210837	29759282
rs159248	29761159
rs150512	29764523
rs210835	29766528

*****Cromossomo 18*****

***** 1a. Região *****

SNP	BP
rs8084454	48842943
rs8084270	48843374
rs4995148	48851527

***** 2a. Região *****

SNP	BP
rs70198	75556678
rs558700	75556859

*****Cromossomo 19*****

***** 1a. Região *****

SNP	BP
rs12975693	36345513
rs2279705	36350972

*****Cromossomo 22*****

***** 1a. Região *****

SNP	BP
rs5749969	33559081

***** 2a. Região *****

SNP	BP
rs8138968	45768799
rs6009087	45769450
rs6009092	45771831
rs3827413	45771899
rs9627641	45772802
rs7289525	45773164
rs6009096	45781228
rs12158956	45781888
rs6009099	45785428
rs11703540	45786323
rs713808	45791114
rs1005553	45791613
rs6009103	45792905
rs6009104	45794963
rs6009105	45795226

Referências Bibliográficas

- [1] Agresti, A. (2002). *Categorical data analysis, 2nd Ed.* New York: Wiley.
- [2] Allen, A.S. and Satten, G.A. (2009). Genome-wide association analysis of rheumatoid arthritis data via haplotype sharing. *BMC Proceeding* **3** Suppl 7: S30.
- [3] Altshuler, D.; Daly, M.J. and Lander, E.S. (2008). Genetic mapping in human disease. *Science* **v. 322**, 881-888.
- [4] Ambrosius W.T.; Lange E.M. and Langefeld C.D. (2004). Power for genetic association studies with random allele frequencies and genotype distributions. *Am. J. Hum. Genet.* **74**, 683-693.
- [5] Amos, C.I; Wu, X.; Broderick, P.; Gorlov, I. P.; Gu, J.; Eisen, T.; Dong, Q.; Zhang, Q.; Gu, X.; Vijayakrishnan, J.; Sullivan, K.; Matakidou, A.; Wang, Y.; Mills, G.; Doheny, K.; Tsai, Y.; Chen, W. V.; Shete, S.; Spitz, M.R. and Houlston, R.S. (2008). Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. *Nat. Genet.* **40(5)**, 616-622.
- [6] Andrade, M. e Pinheiro, H.P. (2002). *Métodos estatísticos aplicados em genética humana*. 15° SINAPE, São Paulo: ABE.
- [7] Aschard, H.; Guedj, M. and Demenais, F. (2007). A two-step multiple-marker strategy for genome-wide association studies. *BMC proceedings* **I**, S134.
- [8] Aulchenko, Y.S.; Ripke, S.; Isaacs, A. and van Duijn, C.M. (2007). GenABEL: an R library for genome-wide association analysis. *Bioinformatics* **23**, 1294-1296.
- [9] Barrett J.C.; Fry B.; Maller J. and Daly M.J. (2005). Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263-265.
- [10] Batista, M.J.; Giolo, S.R.; Pereira, A.C.; and Soler, J.M.P. (2008). Evidence for SNP Effect through Supervised Association Analysis. *Proceedings of the Genetic Analysis Workshop 16 Problem 1*, St. Louis, USA.
- [11] Batista, M.J. (2006). *Análise de associação aplicada ao mapeamento genético de doenças*. Dissertação de mestrado, São Paulo: IME/USP.
- [12] Basu, D. (1975). Statistical information and likelihood (with discussions). *Sankyã, Series A* **37**, 1-71.

- [13] Basu, D. (1977). On the elimination of nuisance paramaters. *J. Am. Statist. Assoc.* **72**, 355-366.
- [14] Basu, D. (1979). Discussion of Berkson's paper "In dispraise of the exact test". *J. Statist. Plan. Inf.* **3**, 189-197.
- [15] Bergen, A.W.; van den Bree, M.B.M.; Yeager, M.; Welch, R.; Ganjei, J.K.; Haque, K.; Bacanu, S.; Berrettini, W.H.; Grice, D.E.; Goldman, D.; Bulik, C.M.; Klump, K.; Fichter, M.; Halmi, K.; Kaplan, A.; Strober, M.; Treasure, J.; Woodside, B. and Kaye, W.H. (2003). Candidate genes for anorexia nervosa in the 1p33-36 linkage region: serotonin 1D and delta opioid receptor loci exhibit significant association to anorexia nervosa. *Molecular Psychiatry* **8**, 397-406.
- [16] Clayton, D.G. and Jones, H. (1999). Transmission/disequilibrium tests for extended marker haplotypes. *Am. J. Hum. Genet.* **65**, 1161-1169.
- [17] Clayton, D. (2003). Population association. In *Handbook of Statistical Genetics, 2nd Ed.* New York: Wiley, 939-960.
- [18] Conti, D.V. and Gauderman, J. (2004). SNPs, haplotypes, and model selection in a candidate gene region: the SIMPlE analysis for multilocus data. *Genetic Epidemiology* **27**, 429-441.
- [19] Correa J.B.; Dellazzana, J.E.; Sturm, A.; Leite, D.M.; de Oliveira Filho, G.R. and Xavier, R.G. (2009). Using the Cusum curve to evaluate the training of orotracheal intubation with the Truview EVO2 laryngoscope. *Rev. Bras. Anesthesiol.* **59(3)**, 321-331.
- [20] Davison, A.C. and Hinkley, D.V. (1997). *Bootstrap methods and their application*. Cambridge: Cambridge University Press.
- [21] Devlin, B.; Roeder, K. and Wasserman, L. (2003). Analysis of multilocus models of association. *Genetic Epidemiology* **25**, 36-47.
- [22] Duncan, C.T., Haile, R.W. and Duggan, D. (2005). Recent development in genome-wide association scans: a workshop summary and review. *Am. J. Hum. Genet.* **77**, 337-345.
- [23] Efron, B. and Tibshirani, R.J. (1994). *An introduction to the bootstrap*. Boca Raton: Chapman & Hall/CRC.
- [24] Ewens, W.J. and Spielman, R.S. (2003). The transmission/disequilibrium test. In *Handbook of statistical genetics, 2nd Ed.* New York: Wiley, 961-972.
- [25] Falconer, D.S. and Mackay, T.F.C. (1996). *Introduction to quantitative genetics*. London: Prentice Hall.

-
- [26] Farah, S.B. (1997). *DNA segredos e mistérios*. São Paulo: Sarvier.
- [27] Farias, R.; Moreno, G. and Patriota, A. (2009). Reduction of models in the presence of nuisance parameters, *Revista Colombiana de Estadística* **32**, 99-121.
- [28] Guedj, M.; Aschard, H.; Nuel, G.; Wojcik, J. and Demenais, F. (2006). Picking-up candidate genomic regions in genome-wide association studies through use of the Local Score statistic, *GAW 15-Genetic Analysis Workshop Group* **15**, 15-19.
- [29] Gianola, D. and Sorensen, D. (2002). *Likelihood, Bayesian and MCMC Methods in Quantitative Genetics*. New York: Springer.
- [30] Giolo, S.R.; Soler, J.M.P.; Batista, M.J.; Almeida, M.A.A. and Pereira, A.C. (2011). Evidence of SNP effect on the risk of rheumatoid arthritis: effects of covariate adjustment upon association results. *Revista Brasileira de Biometria* **29**, 47-59.
- [31] Greenspan G. and Geiger D. (2006). Modeling haplotype block variation using Markov chains. *Genetics* **172**, 2583-2599.
- [32] Horvath, S. and Baur, M.P. (2000). Future directions of research in statistical genetics, *Statist. Med.* **19**, 3337-3343.
- [33] Huang, B.E.; Amos, C.I. and Lin, D.Y. (2007). Detecting haplotype effects in genomewide association studies, *Genetic Epidemiology* **31**, 803-812.
- [34] International HapMap Project (2003). *Nature* **426**, 789-796.
- [35] Jorde, L.B. (2000). Linkage disequilibrium and the search for complex disease genes. *Genome Res.* **10**, 1435-1444.
- [36] Jorde, L.B. (1995). Linkage disequilibrium as a gene-mapping tool. *Am. J. Hum. Genet.* **56**, 11-14.
- [37] Kao, C.H.; Zeng, B.Z. and Teasdale, R.D. (2002). Multiple interval mapping for quantitative trait loci. *Genetics* **152**, 1203-1216.
- [38] Lange, K. (1997). *Mathematical and statistical methods for genetic analysis*. New York: Springer.
- [39] Lazzeroni, L.C. and Lange, K. (1998). A conditional inference framework for extending the transmission/disequilibrium test. *Hum. Hered.* **48**, 67-81.
- [40] Lewontin, R.C. (1964). The interation of selection and linkage. I General considerations, heterotic models. *Genetics* **49**, 49-67.
- [41] Liu, B.H. (1998). *Statistical genomics: linkage, mapping and QTL analysis*. New York: CRC Press.

- [42] Mather, K. (1951). *The measurement of linkage in hereditary*. London: Methuen.
- [43] Mاتيoli, S.R. (2002). TDT2l (A program for performing TDT tests with two close linked markers in trios). <http://adi-38.bio.ib.usp.br/tdt2l/index.html>.
- [44] McCarthy, M.I.; Abecasis, G.R.; Cardon, L.R.; Goldstein, D.B.; Little, J.; Ioannidis, J.P.A. and Hirschhorn J.N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature* **9**, 356-369.
- [45] Millstein, J.; Conti, D.V.; Gilliland, F.D. and Gauderman, W.J. (2006). A testing for identifying susceptibility genes in the presence of epistasis. *Am. J. Hum. Genet.* **78**, 15-27.
- [46] Montgomery, D.C. (2008). *Introduction to statistical quality control, 6nd Ed.* New York: Wiley.
- [47] Morton, N.E. (1955). Sequential tests for the detection of linkage. *Am. J. Hum. Genet.* **7**, 277-318.
- [48] Narain, P. (2007). A theoretical treatment of interval mapping of a disease gene using transmission disequilibrium tests. *J. Biosci.* **32**, 1317-1324.
- [49] Nielsen, R.; Paul, J.S.; Albrechtsen, A and Song, Y.S. (2011). Genotype and SNP calling from next-generation sequencing data. *Nature* **12**, 443-451.
- [50] Ott, J. (1991). *Analysis of Human Genetic Linkage*. London: Johns Hopkins.
- [51] Page, E.S. (1954). Continous Inspection Schemes. *Biometrika* **41**, 100-115.
- [52] Papachristou C. and Lin S. (2006). Microsatellites versus single-nucleotide polymorphisms in confidence interval estimation of disease loci. *Genetic Epidemiology* **30**, 3-17.
- [53] Paulino, C.D. e Singer, J.M. (2006). *Análise de Dados Categorizados*. São Paulo: Edgard Blücher.
- [54] Pena, S.D.; Di Pietro, G.; Fuchshuber-Moraes, M.; Genro, J.P.; Hutz, M.H.; Kehdy Fde, S.; Kohlrausch, F.; Magno, L.A.; Montenegro, R.C.; Moraes, M.O.; de Moraes, M.E.; de Moraes, M.R.; Ojopi, E.B.; Perini, J.A.; Racciopi, C.; Ribeiro-Dos-Santos, A.K.; Rios-Santos, F.; Romano-Silva, M.A.; Sortica, V.A. and Suarez-Kurtz, G. (2011). The Genomic Ancestry of Individuals from Different Geographical Regions of Brazil Is More Uniform Than Expected. *PLoS ONE* **6**, e17063.
- [55] Pender, S.L.F.; Croucher, P.J.P.; Mascheretti, S.; Prothero, J.D.; Fisher, S.A.; MacDonald, T.T.; Schreiber, S. and Shu Ye (2004). Transmission disequilibrium test of stromelysin-1 gene variation in relation to Crohn's disease. *J. Med. Genet.* **41**:e112.
- [56] Pereira, C.A.B. (1980). *Bayesian solutions to some classical problems of statistics*. Thesis (Ph. D.), Florida State University.

- [57] Pritchard, J.K. and Feldman, M.W. (1996). Statistics for microsatellite variation based on coalescence. *Theoretical Population Biology* **50**, 325-344.
- [58] Purcell, S.; Sham, P.C. and Daly, M.J. (2005). Parental phenotypes in family-based association analysis. *Am. J. Hum. Genet.* **76**, 249-259.
- [59] R Development Core Team (2011). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org>.
- [60] Rubinstein, P.; Walker, M.; Carpenter, C.; Carrier, C.; Krassner, J.; Falk, C. and Ginsberg, F. (1981). Genetics of HLA disease associations: the use of the haplotype relative risk (HRR) and the “haplo-delta” (Dh) estimates in juvenile diabetes from three racial groups. *Hum. Immunol.* **3**, 384.
- [61] Sasieni, P.D. (1997). From genotypes to genes: doubling the sample size. *Biometrics* **53**, 1253-1261.
- [62] Sham, P.C. (1998). *Statistics in human genetics*. New York: Arnold.
- [63] Slatkin, M. (1995). A measure of population subdivision based on microsatellite allele frequencies. *Genetics* **139**, 457-462.
- [64] Spielman, R.S.; McGinnis, R.E. and 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.
- [65] Stokes, M.; Davis, C.S. and Koch, G.G. (2000). *Categorical Data Analysis Using the SAS System, 2nd Ed.* North Carolina: SAS Publishing.
- [66] Sykes, N.H.; Toma, C.; Wilson, N.; Volpi, E.V.; Sousa, I.; Pagnamenta, A.T.; Tancredi, R.; Battaglia, A.; Maestrini, E.; Bailey, A.J.; Monaco A.P. and International Molecular Genetic Study of Autism Consortium (IMGSAC) (2009). Copy number variation and association analysis of SHANK3 as a candidate gene for autism in the IMGSAC collection. *European Journal of Human Genetics* **17**, 1347-1353.
- [67] Terwilliger, J.D. and Ott, J. (1994). *Handbook of human genetic linkage*. London: Johns Hopkins.
- [68] Terwilliger, J.D. and Ott, J. (1992). A haplotype-based haplotype relative risk statistic. *Hum. Hered.* **42**, 337-346.
- [69] Wan, X.; Yang, C.; Yang, Q.; Xue, H.; Fan, X.; Tang, N.L.S. and Yu, W. (2010). BOOST: A Fast Approach to Detecting Gene-Gene Interactions in Genome-wide Case-Control Studies. *Am. J. Hum. Genet.* **87**, 325-340.

-
- [70] Watkins, H. (2004). A trio family study showing association of the lymphotoxin- α N26 (804A) allele with coronary artery disease. *European Journal of Human Genetics* **12**, 770-774.
- [71] Weir, B.S. (1996). *Genetic data analysis II*. Sunderland: Sinauer Associates.
- [72] Weir, B.S.; Hill, W.G. and Cardon, L.R. (2004). Allelic association patterns for a dense SNP map. *Genet. Epidemiol.* **27**, 442-450.
- [73] Yang, J.; Benyamin, B.; McEvoy, B.P.; Gordon, S.; Henders, A.K.; Nyholt, D.R.; Madden, P.A.; Heath, A.C.; Martin, N.G.; Montgomery, G.W.; Goddard, M.E. and Visscher, P.M. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics* **42**, 565-569.
- [74] Yi, G.; Coleman, S. and Ren Q. (2006). CUSUM method in predicting regime shifts and its performance in different stock markets allowing for transaction fees. *Journal of Applied Statistics* **33**, 647-661.
- [75] Zhao, H.; Zhang, S.; Merikangas, K.R.; Trixler, M.; Wildenauer, D.B; Sun, F. and Kidd, K.K. (2000). Transmission/Disequilibrium Tests Using Multiple Tightly Linked Markers. *Am. J. Hum. Genet.* **67**, 936-946.
- [76] Ziegler, A.; König, I.R. and Thompson, J.R. (2008). Biostatistical aspects of genome-wide association studies. *Biometrical Journal* **50**, 8-28.