Modelos mistos no mapeamento genético
de fatores de risco cardiovascular em
famílias brasileiras usando dados de SNPs

Mirian de Souza

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

Programa: Estatística

Orientadora: Profa. Dra. Júlia Maria Pavan Soler

*Durante a elaboração deste trabalho a autora
recebeu apoio financeiro da FAPESP.*

- São Paulo, julho de 2012 –

Modelos mistos no mapeamento genético
de fatores de risco cardiovascular em
famílias brasileiras usando dados de SNPs

Este exemplar corresponde à redação
final da dissertação devidamente corrigida e
defendida por Mirian de Souza.

São Paulo, 11 de julho de 2012.

*“Discovery consists of seeing what everybody has seen
and thinking what nobody has thought.”*

Albert Szent-Gyorgyi

*À minha mãe, meu pai e meu irmão (in memoriam),
aos meus tios, primos e amigos.*

Agradecimentos

Agradeço primeiramente a Deus pela vida, saúde e força para superar as dificuldades que apareceram ao longo da elaboração deste trabalho.

À minha mãe, meu pai, meu irmão (*in memorian*), meus tios e primos pelo carinho e apoio, além dos inúmeros amigos, que me encorajaram e ajudaram durante todo o caminho.

À minha orientadora, Profa. Dra. Júlia Maria Pavan Soler, pela compreensão e orientação que recebi ao longo deste trabalho.

Ao laboratório de Cardiologia e Genética Molecular do InCor/USP pela autorização na utilização do banco de dados e pela vinculação da bolsa ao Projeto Corações de Baependi (Processo Fapesp 2007/58150-7).

À Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) pela concessão de bolsa mestrado.

Aos meus amigos Rafael Jerônimo, Rafael Braz, Tiago Maia, Michel, Manuel, Cristian, Maurício e Núbia pela amizade e ajuda que sempre me ofereceram.

À equipe de bioestatística do AC Camargo, em especial a Aline.

A todos, obrigada.

Resumo

O estudo de doenças complexas, tais como hipertensão e glicemia, é de grande importância na área médica, pois essas doenças afetam muitas pessoas no mundo e seu padrão de variação envolve componentes ambientais, genéticos e suas possíveis interações. Para o mapeamento de genes a amostragem do genoma humano é feita por meio de plataformas de marcadores moleculares e, em geral, destacam-se duas classes de marcadores: os do tipo microsatélites e os SNPs (do inglês, *Single Nucleotide Polimorphisms*). Os dados de famílias são comumente analisados via modelos mistos e marcadores microsatélites de efeitos aleatórios, sendo que os estudos caso-controle com indivíduos não relacionados têm sido vinculados a dados de SNPs. Neste contexto, surge a problemática de como modelar o SNP em dados de famílias, pois o mesmo pode ser modelado como um fator fixo ou aleatório. Com a finalidade de trazer contribuições a esta discussão, um dos objetivos deste trabalho é propor um exercício de simulação e análise de dados genéticos que facilite o ensino e o entendimento de conceitos de genética e do mapeamento de genes modelados a partir de efeitos fixos ou aleatórios utilizando o *software R*. Além disso, na análise de dados envolvendo mapas densos de SNPs é necessário contornar o problema de múltiplos testes, e a proposta em multiestágios de Aulchenko et al. (2007) é uma alternativa de análise, na qual o efeito do SNP é modelado como um fator fixo e associado a um componente residual. Logo, surge também como desafio deste trabalho, aplicar o modelo em multiestágios para o mapeamento dos genes e discutir suas vantagens e limitações.

Abstract

The study of complex diseases such as hypertension and glucose is of great importance in the medical field because these diseases affect many people in the world and its pattern of variation involves environmental and genetics components and their possible interactions. For genes mapping the human genome sampling is performed by means of molecular markers platforms, generally including two kinds of markers: the type microsatellite and SNPs (Single Nucleotide Polimorphisms). The family data is commonly analyzed by mixed models and random effects microsatellite markers and the case-control studies with unrelated individuals have been linked to data from SNPs. In this context the question arises of how to model the SNP on family data because it can be modeled as a fixed or random factor. In order to bring contributions to this discussion, one of the objectives of this study is to propose a simulation exercise and analysis of genetic data to facilitate the teaching and understanding of concepts of genetics and gene mapping modeled from fixed or random effects using software R. Furthermore, analysis of data involving dense maps of SNPs is necessary to overcome the problem of multiple tests, and the proposal multistage Aulchenko et al. (2007) is an alternative analysis in which the effect of SNP is modeled as a fixed factor and associated with a residual component. So there is also a challenge of this study to apply the multistage model for the mapping of genes and discuss their advantages and limitations.

Sumário

1. Introdução	1
2. Conceitos em Genética	11
2.1 Alguns Conceitos	12
2.2 Dependência entre locos genéticos	18
2.2.1 Eventos de Recombinação	18
2.3 Funções de Distância Genética	21
2.3.1 Função de Mapeamento de Haldane	23
2.3.2 Função de Mapeamento de Kosambi	23
2.4 Leis de Segregação	24
2.4.1 Equilíbrio de Hardy-Weinberg (dentro do loco)	24
2.4.2 Desequilíbrio de Ligação (entre locos)	26
2.5 Efeitos Genéticos	28
2.6 Análise de Doenças Complexas	31
2.7 Delineamentos com famílias	33
2.8 Genética de populações humanas	35
2.8.1 Marcadores no DNA – Microsatélites	36
2.8.2 Marcadores no DNA – SNPs	37
3. Modelos Mistos	40

3.1 Inferência Estatística	44
3.2 Análise de Resíduos	46
3.2.1 Tipos de Resíduos	46
3.3 Modelos Estatísticos Genéticos	49
3.3.1 Modelo Poligênico	49
3.3.2 Estimação da Herdabilidade	50
3.3.3 Modelo Oligogênico	52
3.3.4 Modelo em Multiestágios	53
4. Simulação de dados Genéticos e Propostas de Exercícios	57
4.1 Material para fins acadêmicos	58
5. Aplicação – Análise em Multiestágios	70
5.1 Descrição da simulação dos dados de famílias	70
5.2 Aplicação do processo em multiestágios	72
6. Considerações Finais	76
7. Apêndice A	79
8. Apêndice B	84
9. Apêndice C	97
10. Referências Bibliográficas	110

Capítulo 1

Introdução

A Genética é o ramo da biologia que estuda a transferência das características físicas, químicas e biológicas de geração para geração. Um dos desafios da genética é compreender como as partes que compõem a vida funcionam em conjunto e são influenciadas pelos fatores ambientais, criando e operando sistemas vivos dinâmicos (Falconer et al., 1996). Nesta área, graças a estudos com mapeamento de genes, é possível descobrir a causa de muitas doenças e, assim, mecanismos de prevenção poderão ser desenvolvidos a partir das informações obtidas sobre os genes que as afetam (Farah, 1997).

Um dos principais objetivos da pesquisa em Genética humana é o estudo de doenças complexas, ou fenótipos quantitativos, tais como a hipertensão, glicemia e obesidade, os quais, num contexto estatístico, são variáveis aleatórias com distribuição contínua. Em contraste com as doenças mendelianas, que resultam de uma mutação em um único gene, fenótipos complexos são produtos da ação de múltiplos genes e fatores ambientais, bem como suas interações. Como produto de muitas pesquisas a comunidade genética tem alcançado grande sucesso em encontrar os genes que são responsáveis por uma ampla gama de doenças mendelianas, mas a busca por genes associados a doenças complexas não tem alcançado tal ritmo, apesar de intenso esforço de pesquisa dos setores acadêmico e comercial (Pritchard and Cox, 2002). Vale ressaltar que a diferença entre doenças simples e complexas é relativamente arbitrária, porque não existe uma clara transição de uma para outra (Ziegler et al., 2008).

A pesquisa genômica se propõe a estudar a natureza física e o funcionamento dos genomas, isto é, do material genético contido no conjunto de cromossomos de cada espécie (Lee & Lee, 2000). Um procedimento essencial para esta pesquisa é o mapeamento genético, o qual visa localizar locos genéticos associados com um fenótipo ou variável de interesse, que se encontram nos cromossomos em posição linear e estimar seus efeitos. A localização destes locos, em geral, é feita por meio da taxa de recombinação gênica, que possibilita revelar a posição citológica entre locos (Farah, 1997). Em especial, deseja-se localizar com maior precisão regiões cromossômicas envolvidas com a regulação de fenótipos e mensurar seu efeito em certas características de interesse específico, tais como aquelas que causam doenças em humanos ou, genes que controlam fenótipos de importância econômica em plantas e animais (Liu, 1998). Para que seja possível o mapeamento de genes foram desenvolvidos os mapas de marcadores moleculares, que são definidos como uma amostra do genoma, isto é, regiões de localização cromossômica conhecida e genotipáveis, ou seja, são caracteres de herança ou locos cromossômicos aos quais se pode atribuir um genótipo (por exemplo, *aa*, *Aa* ou *AA*). Atualmente, mapas de marcadores moleculares estão disponíveis para mapeamento de genes de muitas espécies (humanos, macacos, ratos, trigo, milho, soja, bactérias, entre outras) (Duarte, 2007).

Para ilustração, na Figura 1.1 está representado o mapa genético de uma determinada espécie. Cada marcador ocupa uma posição fixa no cromossomo e tem um nome específico (documentado na coluna da direita). O mapa reflete a ordem dos locos ao longo do cromossomo e a distância relativa entre eles (ilustrado na coluna da esquerda), apontando, por exemplo, que quanto maior a distância entre dois locos maior a taxa de recombinação entre eles. Neste caso, a distância citogenética está em unidades de centMorgan (cM). A distância citogenética entre marcadores moleculares em um mapa é

obtida por meio da probabilidade de recombinação entre eles, ou seja, da probabilidade de ocorrência de permuta genética entre dois locos cromossômicos, que por sua vez pode ser calculada por meio das funções de mapeamento. Para mais detalhes, ver seção 2.3- Funções de Distância Citogenética.

Figura 1.1- Mapa genético de três cromossomos

Para o mapeamento de genes associados a doenças complexas, os avanços biotecnológicos têm sido inestimáveis no sentido de permitirem a construção de mapas densos de marcadores, isto é, a amostragem cada vez mais densa e completa do genoma humano. Neste processo, basicamente dois sistemas de amostragem estão envolvidos: amostragem de indivíduos de uma população de interesse e amostragem do genoma destes indivíduos. A amostragem de indivíduos, em geral, obedece a planos amostrais clássicos, como os estudos observacionais caso-controle e a amostragem de famílias (indivíduos e seus familiares). A amostragem do genoma dos indivíduos é feita por meio de plataformas de marcadores moleculares e, em geral, destacam-se duas classes de marcadores que têm sido usados em mapeamento de genes: os do tipo microsatélites e os SNPs (do inglês, *Single Nucleotide Polimorphisms*).

As plataformas de microsatélites são pouco densas (utilizam espaçamentos médios de 5 a 10 cM entre os locos), seus marcadores estratificam a população em muitas categorias genótípicas (são altamente polimórficos, isto é, podem ocorrer muitas formas alélicas no marcador gerando muitas classes genótípicas, por exemplo, A_1A_1 , A_1A_2 , A_1A_3 , A_2A_3 e assim por diante) e, em geral, cada marcador corresponde a uma grande região cromossômica (de centenas a milhares de bases). Deste modo, estes marcadores podem conter mais de uma variação na sequência de DNA, que seja comum na população ou mesmo rara, isto é, os dados destes marcadores promovem uma correlação entre indivíduos da mesma família, pois a chance de dois indivíduos de famílias diferentes apresentarem o mesmo perfil de marcadores é extremamente pequena, o contrário ocorrendo para indivíduos parentes.

Já os SNPs, ou polimorfismo de um único nucleotídeo, são definidos como uma diferença que ocorre em um par homólogo de nucleotídeos em determinado loco nos genomas de indivíduos da população geral (Griffiths et al., 2008). O termo polimorfismo, neste caso, é usado para definir uma variação genética em um único nucleotídeo dentro de um gene (ou na sequência de DNA), sendo que tal variação é comum, isto é, deve estar presente em pelo menos 1% da população (Ziegler et al., 2008). Cerca de 99% da sequência do genoma humano é idêntica para toda a população, contudo as pequenas variações na sequência do DNA podem ter um impacto importante sobre a forma de como os seres humanos respondem a diferentes condições. Isto dá aos SNPs um grande valor para a investigação biomédica. As plataformas atuais de SNPs têm aproximadamente 1 milhão de marcadores (no caso da plataforma Affymetric 6.0) e cobre cerca de 80% do genoma humano (Carlson et al., 2003). Estima-se que há de 10 a 15 milhões de variações do tipo

SNPs em todo genoma humano (Ziegler et al., 2008). Cada marcador SNP é dialélico, isto é, estratifica a população em até três grupos genotípicos (por exemplo, AA, Aa e aa).

Acredita-se que os SNPs podem capturar informações sobre a variabilidade genética do gene em estudo, embora não sirvam como verdadeiras variantes causadoras de doenças. Ou seja, os SNPs não são necessariamente funcionais, mas podem estar localizados dentro de regiões funcionais ou próximos de uma variante funcional, e por isso são úteis no mapeamento de genes. Considere, por exemplo, um cenário onde se deseja investigar a associação entre um gene e uma doença. Neste caso, o gene compreende uma região de ácido desoxirribonucleico (DNA), representando uma parte do genoma humano. Isto é ilustrado pelo retângulo sombreado na Figura 1.2.

Figura 1.2 - Marcadores de SNPs

Em um modelo simples, pode-se supor que uma mutação em um único lugar dentro desta região resulta na doença. Em geral, a localização precisa desta variante causadora de doenças não é conhecida. Estes SNPs tendem a estar associados ao loco com o verdadeiro genótipo causador da doença. A idéia subjacente a este fenômeno é que, ao longo do tempo

evolutivo (ou seja, através de muitas gerações de reprodução), a mutação ou o alelo da doença foi herdado junto com a variante neste loco. Isso ocorre quando a probabilidade de um evento de recombinação na região de DNA entre o loco da doença e o loco marcador é pequena. Dessa forma, entendendo a variabilidade nestes locos tenderá a capturar, indiretamente, a variabilidade no verdadeiro loco da doença (Foulkes, 2009).

O estudo da Genética humana pode ser feito com indivíduos não relacionados ou com agregação familiar, que envolve dados coletados em vários indivíduos dentro da mesma unidade familiar. Na pesquisa com famílias os indivíduos dentro da mesma família são mais prováveis de serem mais semelhantes uns com os outros do que com pessoas de famílias diferentes. Este fenômeno é conhecido como *clustering* e implica em uma correlação entre variáveis observadas em indivíduos dentro da família (Foulkes, 2009). Delineamentos com famílias têm sido usados com sucesso no mapeamento de genes de doenças comuns, como o projeto *Northern Manhattan Family Study* (Sacco et al., 2007), que investigou a herdabilidade da síndrome metabólica em 803 pacientes de 89 famílias caribenhas hispânicas, ou o projeto *San Antonio Family Heart Study* (Comuzzie et al., 2003) que considerou uma amostra de 1400 membros em mais de 40 famílias americanas mexicanas que vivem no Texas. Nestes estudos, dados foram coletados em indivíduos e seus familiares, estendendo os graus de parentesco tanto quanto possível para inclusão de relacionamentos verticais (ao longo de gerações, pais, filhos, netos) e horizontais (tios, sobrinhos, primos).

Considerando variáveis quantitativas avaliadas em indivíduos e seus familiares, mesmo sem qualquer informação de marcadores moleculares, é possível obter estimativas de herdabilidade poligênica que informam se há componente genético envolvido na variabilidade da resposta de interesse. Neste caso, tem sido adotado o modelo de

componentes de variância (Almasy and Blangero, 1998; de Andrade et al., 1999) para análise de dados familiares, sendo o coeficiente de herdabilidade definido como a correlação intra-classe (sob a formulação geral de um modelo misto), representando a proporção da variância total que é devida a componentes genéticos.

Atualmente sabe-se que doença cardiovascular é uma doença complexa cujos fatores de risco abrangem obesidade central, hipertensão arterial sistêmica, entre outros. Agregação familiar, adoção, gemelaridade e estudos de segregação indicam que os fatores de risco cardiocascular têm importante componente hereditário, havendo, todavia, interação entre fatores genéticos e ambientais. Diversos estudos em diferentes populações analisaram a contribuição de fatores genéticos para a agregação familiar dos fatores de risco cardiovascular e seus componentes. Contudo, na população brasileira estes dados permanecem desconhecidos. A partir do interesse em analisar estes dados na população brasileira surgiu o Projeto Corações de Baependi, gerenciado pelo Laboratório de Cardiologia e Genética Molecular do InCor, cujo objetivo consiste em realizar o mapeamento genético de genes associados a fatores de risco cardiovascular na população brasileira utilizando painel de genotipagem contendo 1.000.000 de marcadores polimórficos do tipo SNP espalhados pelo genoma humano. Para cumprir os requisitos mínimos de estimar a prevalência de fatores de risco cardiovascular nesta população de interesse foram examinadas 1700 pessoas, estratificadas pela densidade populacional do município de Baependi – MG, provenientes de 119 famílias selecionadas aleatoriamente. Até o momento apenas as análises que não dependem dos dados de marcadores moleculares foram feitas, por exemplo, Oliveira et al. (2008) analisando dados de Baependi obtiveram estimativas de herdabilidades poligênicas para diferentes fenótipos associados a fatores de risco cardiovascular e Giolo et al. (2009), considerando dados da mesma população,

obtiveram estimativas de componentes de variância poligênico para a idade de diagnóstico de hipertensão, glicemia e alto colesterol. Os resultados destes dois trabalhos mostraram herdabilidades poligênicas (para mais detalhes, ver seção 3.2.2 – Estimação da Herdabilidade) significantes nos fenótipos avaliados, justificando o interesse pelo mapeamento dos genes envolvidos nestes efeitos. Nesse contexto, surgiu a motivação para o desenvolvimento do presente trabalho cujo um dos objetivos é formular e implementar computacionalmente modelos estatísticos para o mapeamento de genes relacionados a variáveis quantitativas considerando a análise de dados de delineamentos com famílias e plataformas de SNPs.

Para análises de mapeamento de genes com dados de famílias existem na literatura vários métodos que têm sido desenvolvidos explorando tanto a variação das respostas dentro como também entre as famílias. Mas muitos destes métodos são computacionalmente exigentes (principalmente para famílias com um número grande de indivíduos e muitos SNPs a serem avaliados) e demandam muito tempo computacional. A metodologia em multiestágios proposta por Aulchenko et al. (2007) sugere que para detectar o efeito do SNP seja feita uma análise de mapeamento de genes em vários estágios usando modelos mistos e regressão clássica, o que reduz drasticamente o tempo computacional, por isso, surge também como desafio deste trabalho aplicar o modelo em multiestágios para o mapeamento dos genes dos dados do Projeto Corações de Baependi. Além disso, este trabalho propõe discutir as limitações e vantagens da formulação de modelos mistos sob o procedimento em multiestágios para análise de dados de famílias e plataformas de marcadores SNPs, proposto por Aulchenko et al. (2007), pois sabe-se que os dados de SNPs sempre estiveram vinculados a estudos caso-controle com indivíduos não relacionados, os quais é esperado que compartilhem pequenas regiões genômicas em

comum. Já os dados de famílias têm sido comumente analisados via modelos mistos e marcadores microsatélites de efeitos aleatórios. Neste cenário, surge a temática de como modelar o SNP no mapeamento de genes usando dados de famílias, pois o mesmo pode ser modelado como efeito fixo ou aleatório e esta decisão, de modelar um efeito genético como um fator fixo ou aleatório, não é trivial e vários autores têm levantado esta discussão na literatura (Almasy and Blangero, 1998; Blangero et al., 2004). Com a finalidade de trazer ilustrações e contribuições a esta discussão um segundo objetivo do presente trabalho é propor um exercício de simulação e análise de dados genéticos que facilite o ensino e o entendimento de conceitos de genética e do mapeamento de genes modelados a partir de efeitos fixos ou aleatórios. Deste modo, optou-se pela alternativa didática, que foi gerar computacionalmente dados de genótipos de SNPs e variáveis fenotípicas para dados de famílias utilizando recursos do aplicativo estatístico *R*. Os genótipos de cada indivíduo não relacionado foram gerados respeitando a estrutura familiar e o equilíbrio de Hardy-Weinberg (H-W), que descreve as probabilidades genotípicas em termos das probabilidades alélicas (para mais detalhes ver seção 2.4.1 - Equilíbrio de Hardy-Weinberg). Posteriormente foram gerados os fenótipos dos indivíduos seguindo distribuição normal multivariada com diferentes estruturas para os parâmetros do vetor de média e da matriz de covariâncias tendo como objetivo estimar o efeito do SNP sob diferentes contextos de simulação.

A possibilidade de gerar o banco de dados e analisá-los possibilita entender como alguns conceitos genéticos se aplicam na prática e também como a carga genética é transferida de geração para geração. Este material pode ser usado para diferentes finalidades acadêmicas atendendo a dificuldades tanto de estudantes da área de Estatística,

que desejam modelar dados genéticos, quanto de estudantes da Genética, interessados em entender a modelagem de dados genéticos.

Para introduzir conceitos envolvidos no mapeamento de genes e compreender de que maneira os dados de famílias e SNPS podem ser simulados e quais condições devem ser respeitadas durante a simulação, são introduzidos no capítulo 2 alguns tópicos de genética, análise de doenças complexas e tipos de marcadores. Os diferentes modelos estatísticos em Genética, a metodologia proposta por Aulchenko et al. (2007) e o uso de seus resíduos são apresentados no capítulo 3. O capítulo 4 descreve os exercícios didáticos que podem ser utilizados para diferentes finalidades acadêmicas. Os diferentes cenários de simulação de dados, a aplicação do modelo em multiestágios nos dados simulados e a discussão dos resultados em cada um dos contextos são apresentados no capítulo 5. As considerações finais com os principais resultados e contribuições da dissertação, bem como algumas sugestões para pesquisas futuras são apresentados no capítulo 6.

Capítulo 2

Conceitos em Genética

Este capítulo é relevante para uma melhor compreensão e desenvolvimento da metodologia e da simulação abordadas neste trabalho. Serão introduzidos alguns conceitos de Genética.

O crescimento e a compreensão da área da Genética só foram possíveis graças ao redescobrimto dos trabalhos de Gregor Johann Mendel em 1900. As pesquisas de Mendel com cruzamentos de plantas permitiram concluir que as unidades hereditárias que eram transmitidas de geração a geração eram denominadas genes. Entre as contribuições de Mendel destacam-se suas leis de segregação:

- 1- Primeira Lei de Mendel ou Lei da segregação igual - Cada característica é determinada por um par de fatores genéticos denominados alelos. Estes, na formação dos gametas, são separados e, desta forma, pai e mãe transmitem apenas um de seus alelos para seu descendente.
- 2- Segunda Lei de Mendel ou Lei da segregação independente - Os pares alélicos localizados em diferentes cromossomos segregam independentemente em gametas e conjuntamente em zigotos.

Maiores detalhes sobre este assunto são discutidos em Haines & Pericak-Vance (1998), Griffiths et al. (2008) e Crow & Franklin (1970) e informações históricas em

Sturtevant (2001). É importante salientar que as leis de Mendel se aplicam a modelos de herança monogênicos, isto é, ao estudo de características qualitativas controladas por um único gene (por exemplo, a hemofilia A e a anemia falciforme), e não podem ser utilizadas diretamente no mecanismo de herança de características quantitativas (por exemplo, doenças complexas como hipertensão arterial e obesidade), uma vez que estas podem ser controladas por muitos genes que interagem entre si e com o ambiente (Falconer & Mackay 1996).

A partir de então, foi desenvolvida uma teoria quantitativa, baseada na Genética Mendeliana, criando subsídios para responder a uma das perguntas de maior interesse em Genética Quantitativa, ou seja, quanto da variabilidade de uma ou mais respostas de interesse numa certa população é devido a causas genéticas e quanto a causas ambientais (Lynch & Walsh, 1998) e quais genes estão envolvidos no processo de herança.

2.1 Alguns Conceitos

Todas as células humanas, exceto as células vermelhas do sangue, carregam em seu núcleo a informação genética em forma de cromossomo (Ziegler et al., 2008). Os cromossomos são encontrados exclusivamente no núcleo das células e são compostos por ácido desoxirribonucléico (DNA) e proteínas.

O DNA especifica as funções que serão realizadas pelas células que compõem o organismo e determina, juntamente com o ambiente, suas características físicas nos diferentes níveis (bioquímico, fisiológico, etc). O DNA é composto de dois filamentos enrolados um no outro em uma longa dupla hélice (Griffiths et al., 2008). Cada filamento de DNA é uma grande molécula feita de sequências lineares de subunidades chamadas

nucleotídeos. Cada nucleotídeo é formado por três partes: açúcar (desoxirribose), base nitrogenada e fosfato. Existem quatro tipos diferentes de bases nitrogenadas: adenina (A), timina (T), guanina (G) e citosina (C) (Sham, 1998). O DNA é uma cadeia de nucleotídeos mantidos juntos por pontes entre A e T e entre G e C (Griffiths et al., 2008), como ilustrado na Figura 2.1.

Figura 2.1: A estrutura do DNA

Embora existam apenas quatro tipos de nucleotídeos em um único filamento de DNA, estes podem estar em qualquer ordem. As sequências de nucleotídeos correspondem a genes (...TTACGGACCT...), os quais codificam proteínas e estão distribuídos de forma linear ao longo do DNA. Cada gene ocupa uma região no cromossomo chamada de loco (do inglês, *locus*). De acordo com o *National Human Genome Research Institute*, o genoma

humano contém aproximadamente de 20000 a 25000 genes. Cerca de 1% de todo o DNA é composto por genes (Ziegler et al., 2008), a parte restante é formada por *introns* que têm funções reguladoras da expressão genética, estabilidade química, etc.

Gene é a unidade fundamental física e funcional da hereditariedade que leva informação de uma geração para a seguinte (Griffiths et al., 2008). Uma das formas diferentes de um gene que pode existir em um único loco são chamadas de alelos. O grupo de alelos de diferentes genes ou locos no mesmo cromossomo recebido de um dos pais do indivíduo é chamado de haplótipo. O ser humano é diplóide porque possui duas cópias de cada cromossomo, ou equivalentemente, possui duas cópias de cada gene (uma cópia herdada do pai e a outra da mãe).

A informação genética humana está distribuída em 23 pares de cromossomos, sendo 22 deles autossômicos e um sexual (gametas), ou seja, o conjunto de diferentes cromossomos (cariótipo, ver Figura 2.2) é formado por 46 cromossomos. O conjunto formado por uma das cópias de todos os cromossomos é denominado genoma humano, que é composto por 23 cromossomos.

A constituição genética de um indivíduo, ou mais especificamente, os alelos presentes em um dos locos do cromossomo representam o genótipo do indivíduo com respeito a esse loco. O genótipo não é observável, porém pode ser deduzido por meio de processos indiretos (de cruzamentos laboratoriais entre outros), sendo definido no momento da fecundação e não se altera, salvo alguma mutação, durante toda a vida do indivíduo. Já o conjunto de características observáveis nesse indivíduo no nível bioquímico, fisiológico ou morfológico (com respeito a esse loco) é conhecido como o fenótipo e pode ser determinado pela interação de sua constituição genética (o genótipo) em associação a outros locos e/ou a fatores ambientais. O fenótipo é tudo aquilo que é visível ou mensurável

de alguma forma e que está associado a alguma característica do indivíduo, sendo que alguns fenótipos podem se alterar com o passar do tempo, por exemplo, altura, hipertensão arterial, peso entre outros. Neste contexto usaremos fenótipos para denominar uma variável aleatória avaliada no indivíduo.

O fato do ser humano ser diplóide e receber uma cópia de cada gene de um dos pais foi abordado na simulação de dados considerada neste trabalho, mais especificamente na construção do genótipo dos filhos, onde cada gene dos pais foi doado seguindo uma distribuição Uniforme [0,1]. Maiores detalhes estão apresentados na seção 4.1- Descrição da simulação dos dados de famílias.

Figura 2.2: Cariótipo Humano

Quando os alelos num determinado loco são idênticos, o genótipo do indivíduo é caracterizado como homozigoto com respeito a esse loco (por exemplo, AA ou aa). Caso apresente no mesmo loco alelos diferentes, em cada cromossomo homólogo (por exemplo, Aa), chama-se heterozigoto.

Há três formas de expressão fenotípica de um heterozigoto diplóide (Griffiths et al., 2008):

- a) Dominância Total ou completa: neste caso indivíduos com genótipo, Aa , são fenotipicamente iguais a indivíduos com genótipo AA , mas diferentes de indivíduos com genótipo aa , pois o alelo A é dominante para o alelo a , ou, equivalentemente, a é recessivo para o alelo A . No caso dominante basta uma cópia do alelo para a característica se expressar. No caso recessivo a expressão do fenótipo só ocorre na presença de duas cópias do alelo. Logo, têm-se três possíveis genótipos (AA , Aa e aa) e duas classes fenotípicas, como ilustrado na Figura 2.3.

Figura 2.3: Ilustração da expressão fenotípica Dominância Total

- b) Dominância Incompleta: a expressão fenotípica do heterozigoto é intermediária. O genótipo Aa conduz a um fenótipo que é o meio termo entre os fenótipos AA e aa , então não há dominância de um dos alelos e os três genótipos produzem, respectivamente, três fenótipos (ver Figura 2.4). O padrão de expressão fenotípica neste caso é linear ou aditivo.

Figura 2.4: Ilustração da expressão fenotípica Dominância Incompleta

- c) Superdominância: fuga da linearidade em que a expressão do heterozigoto envolve uma interação entre os alelos alternativos, A e a .

Figura 2.5: Ilustração da expressão fenotípica Superdominância

- d) O heterozigoto mostra propriedades de ambos os alelos. Um exemplo é visto no grupo sanguíneo ABO de humanos, em que os indivíduos com genótipo AA ou AO possuem sangue tipo A, onde A é o alelo dominante, aqueles com genótipo BB ou BO possuem sangue tipo B e o alelo B é dominante e, finalmente, indivíduos com genótipo AB possuem sangue tipo AB e os alelos A e B são

codominantes. As combinações dos três alelos resultam em seis genótipos e quatro fenótipos diferentes:

Tabela 2.1: Relação dos Genótipos e Fenótipos do sistema ABO

Genótipo	Tipo Sanguíneo (Fenótipo)
AA, AO	A
BB, BO	B
AB	AB
OO	O

2.2 Dependência entre locos genéticos

Para descrever a configuração genética de uma população, deve-se primeiramente especificar seus possíveis genótipos e, então, as probabilidades genotípicas (probabilidades dos indivíduos serem portadores de cada genótipo) e as probabilidades alélicas (probabilidades de ocorrência de cada alelo nesta população). Em amostras estimam-se estas probabilidades por meio das correspondentes proporções observadas. A estimação dessas probabilidades, assim como a interação de diferentes alelos no mesmo ou em diferentes locos e o grau com o qual tais interações governam as probabilidades genéticas, é de grande interesse para a análise genética de populações. A seguir é apresentado um dos conceitos fundamentais envolvidos com a estrutura de dependência entre alelos e genótipos nos locos genéticos.

2.2.1- Eventos de Recombinação

Durante a meiose, processo de divisão celular que leva à formação de gametas, as cromátides homólogas ou irmãs se duplicam e se emparelham, podendo partir-se em

determinados pontos e juntar-se novamente, porém com as porções dos filamentos trocados (quiasma), tornando possível o surgimento de novos tipos de gametas, que são chamados de gametas recombinantes. Este mecanismo é conhecido como Recombinação Genética ou *Crossing over* e está ilustrado na Figura 2.6.

Figura 2.6- Representação do crossing-over e formação de gametas recombinantes

A transmissão de locos como eventos independentes é geralmente verdadeira para genes que estão em diferentes cromossomos. Vale ressaltar que quando dois locos genéticos estão muito próximos e no mesmo cromossomo, a ocorrência de recombinação entre eles é improvável, e os genes que compõem estes locos tendem a ser transmitidos juntos, como um grupo, e íntegros, em cada meiose, mas quando dois locos estão distantes

no mesmo cromossomo diz-se “não ligados” e é provável que ocorra *crossing over* em algum ponto entre os genes, produzindo alelos recombinantes. Nota-se que este raciocínio parte do pressuposto de que os genes se distribuem linearmente ao longo dos cromossomos, ocupando posições bem definidas (Amabis et al., 1997). Em vista disso, verifica-se que a ocorrência de recombinação depende da distância entre os locos em questão.

A probabilidade de que ocorra evento de recombinação entre dois locos quaisquer na meiose é conhecida como fração de recombinação e será representada neste texto por θ . A fração de recombinação está diretamente relacionada à distância citogenética entre locos num mesmo cromossomo, isto é, quanto mais próximos dois locos estão, menor é a probabilidade de ocorrência de recombinação entre eles (Ott, 1991). Logo, a fração de recombinação é função da distância citogenética (Ziegler et al., 2008). O parâmetro θ pode ser expresso pela fórmula de Mather, como a seguir (Lange, 1997):

$$\theta = \frac{1}{2}P(\lambda > 0) = \frac{1}{2}[1 - P(\lambda = 0)] \quad (2.1)$$

onde λ indica o número de eventos de recombinação entre dois locos no mesmo cromossomo. Desse modo, o intervalo dentro do qual θ pode variar é de 0 a 0,5. Espera-se que o número máximo de recombinantes existentes seja a metade do número de gametas possíveis, isto é, considerando que durante a meiose ocorreu 100% de *crossing over*, 50% dos gametas formados serão parentais e os outros 50% serão recombinantes, de modo que a probabilidade de recombinação é sempre menor ou igual a 50%, ou seja, $\theta \leq 0,5$ (Bearzoti, 1997).

Se $\theta = 0,5$ não existe ligação genética, isto é, $P(\lambda = 0) = 0$. Se $\theta < 0,5$, ou, em especial se θ está muito próximo de zero, isto significa que há ligação genética entre os dois

locos em questão, sendo estes denominados não recombinantes. No procedimento de inferência estatística sobre o parâmetro θ é usual definir a hipótese nula como sendo $H_0: \theta = 0,5$ (locos não ligados) e a hipótese alternativa como $H_1: \theta < 0,5$. No teste da hipótese H_0 , se a evidência amostral apontar para a rejeição de H_0 , conclui-se que há ligação entre os locos sob análise.

2.3 Funções de Distância Genética

Embora o método para medir distância entre dois locos seja definido em termos da probabilidade de recombinação, θ , nem sempre tais probabilidades fornecem medidas diretas dessas distâncias. A presença de um *crossing over*, ou permuta de material genético, numa determinada região pode diminuir a probabilidade de ocorrência de permutas em outras regiões e, conseqüentemente diferentes modelos têm sido adotados para estimar a distância dos locos em função de θ . A probabilidade de recombinação não é uma boa medida de comprimento genético por si só, porque esta não pode exceder 50% (Birren et al., 1999). Por exemplo, suponha que os locos A , B e C estejam dispersos em um cromossomo e que $\theta_{AB} = 0,3$ e $\theta_{BC} = 0,4$, onde θ_{AB} é a probabilidade de recombinação entre os locos A e B e θ_{BC} , entre os locos B e C . Nesse caso, θ_{AC} não pode ser igual a 0,7, pois a fração de recombinação tem como limite superior o valor 0,5.

Os mapas de ligação mais precisos são aqueles obtidos da adição da probabilidade de recombinação de um grande número de pequenos intervalos genéticos. Alternativamente, o mapa destas distâncias pode ser obtido corrigindo a probabilidade de recombinação observada e assumindo algum modelo de *crossing over* (Birren et al., 1999). Como a fração de recombinação não é uma medida aditiva fez-se necessário o

desenvolvimento das funções de mapeamento. Estas funções são utilizadas para converter as probabilidades de recombinação (θ) em medidas de distância entre locos genéticos cujas propriedades permitam o ordenamento de locos no cromossomo.

Denote a medida da distância (citogenética) por λ a qual é definida como o número médio de pontos de permuta genética no segmento cromossômico sob estudo. A unidade dessa escala de medida foi denominada Morgan (M), de forma que um Morgan (1M) corresponde a um segmento cromossômico no qual ocorre, em média, uma permuta genética por gameta em cada geração. É comum utilizar esta medida em centiMorgans (cM), a qual corresponde à distância entre dois locos cuja recombinação ocorre em 1% dos casos, logo $1 \text{ cM} = 0,01\text{M}$.

Um mapa de ligação mostra a ordem dos locos em um cromossomo, sendo esta ordem baseada nos dados de fração de recombinação entre os locos. Ao se consultar um mapa de ligação, as informações de interesse são a probabilidade de recombinação, θ , existente entre os locos bem como a posição dos locos; portanto é preciso que exista uma relação que permita obter θ a partir de λ e vice-versa. Essa regra de associação é chamada de função de mapeamento (ou função de distância citogenética). A função de distância citogenética mais simples é conhecida como função identidade e foi proposta por Morgan (1928):

$$\theta = \lambda \quad (2.2)$$

apropriada para pequenos intervalos onde a probabilidade de múltiplos *crossover* é praticamente nula, o que vem a ser garantido para genes próximos e ligados geneticamente com $\theta < 0,1$ (Ott, 1991). Outras funções matemáticas de mapeamento têm sido desenvolvidas (Haldane, 1919; Kosambi, 1944; Carter & Falconer, 1951; Rao et al. 1977; Karlin, 1984; Ott, 1991), sendo as funções de Haldane e Kosambi as mais utilizadas para se

relacionar distância genética λ entre dois locos no mapa e a probabilidade de recombinação θ . Mais detalhes sobre a formalização de funções de distância genética são encontrados em Lange (1997) e Ott (1991).

2.3.1 Função de Mapeamento de Haldane

A função de Haldane é apropriada quando a distância entre dois genes é assumida maior, múltiplos *crossovers* podem ocorrer entre locos ao longo do genoma e, além disso, não assume quase nenhuma interferência, ou seja, as ocorrências de permutas genéticas são independentes. Para tal função assume-se também que os *crossovers* se distribuem no cromossomo, segundo uma distribuição de Poisson. Para que haja recombinação entre dois locos separados por uma distância λ é necessário que ocorra um número ímpar de permutas genéticas, pois, com a ocorrência de um número par de permutas genéticas a configuração original é restaurada. Como o número de *crossovers* em um segmento de tamanho λ pode ser modelado segundo a distribuição de Poisson, com média λ , tem-se:

$$\theta = \frac{1}{2}(1 - e^{-2\lambda}), \quad (2.3)$$

cuja função inversa se dá por:

$$\lambda = -\{\log_e(1 - 2\theta)\}/2. \quad (2.4)$$

2.3.2 Função de Mapeamento de Kosambi

A função de Kosambi admite interferência, ou seja, as ocorrências de permutas genéticas não são eventos independentes. Neste caso, a relação entre a distância λ com a probabilidade de recombinação θ pode ser expressa como sendo:

$$\theta = \frac{1}{2} \tanh(2\lambda) = \frac{1}{2} \left(\frac{e^{4\lambda} - 1}{e^{4\lambda} + 1} \right), \quad (2.5)$$

em que *tanh* representa a função tangente hiperbólica. A expressão acima tem como inversa:

$$\lambda = \log_e \{ (1 + 2\theta) / (1 - 2\theta) \} / 4. \quad (2.6)$$

Bearzoti (2000) e Lynch & Walsh (1998) comparam as funções de mapeamento de Morgan, Haldane e de Kosambi de acordo com as distâncias em centimorgans (cM) e as probabilidades de recombinação variando de 0 a 0,5 (ver Figura 2.7). Afirmam que para $\theta < 0,15$ as três funções fornecem estimativas de $\lambda \cong \theta$, diferindo para probabilidades de recombinações maiores. Para valores de $\theta \cong 0,5$ as funções de Haldane e Kosambi tendem a aproximar-se.

Os conceitos de probabilidade de recombinação e funções de distância são úteis na construção de mapas de marcadores moleculares que são locos com posição citogenética (em cM) conhecida no genoma. As plataformas de microsatélites são um exemplo disso. Já mapas de marcadores do tipo SNP são construídos considerando distâncias físicas em pares de bases (bp).

2.4 Leis de Segregação

2.4.1 Equilíbrio de Hardy- Weinberg (dentro do loco)

Um marco histórico em Genética ocorreu no início de 1900 quando biólogos evolucionistas tentaram explicar porque a ocorrência de fenótipos dominantes ou doenças na população não aumentam, até que, ao longo de muitas gerações, toda a população esteja afetada. A resposta para essa questão foi dada, independentemente, por Hardy (1908) e Weinberg

(1908), que previram o comportamento de alelos em uma população usando a distribuição binomial. Sua prova, denominada de teorema ou lei de Hardy- Weinberg (H-W), mostra que em uma grande população se a frequência de traços genotípicos depender de genes autossômicos o equilíbrio é alcançado após uma geração de cruzamentos aleatórios. Para traços genotípicos que dependem de genes sexuais a obtenção de equilíbrio exigirá mais de uma geração (Haines & Pericak-Vance, 1998). Processos como seleção, migração, mutação e oscilação podem alterar as probabilidades genotípicas.

Figura 2.7- Comparação entre diferentes funções de distância

Supondo independência na combinação dos alelos paternos que definem o genótipo do filho, as probabilidades genotípicas podem ser escritas como o produto das probabilidades alélicas correspondentes, isto é:

$$P(A_i A_j) = P(A_i)P(A_j), \quad (2.7)$$

onde $P(A_i A_j)$ é a probabilidade do genótipo $A_i A_j$ e $P(A_i)$ e $P(A_j)$ são as probabilidades dos alelos A_i e A_j ($i, j = 1, 2$), respectivamente (em locos dialélicos). Tal fatoração pode ser chamada de independência horizontal.

Considerando o loco dialélico A com probabilidades de ocorrência dos alelos A_1 e A_2 sendo p e q , respectivamente, ($p + q = 1$). Sob equilíbrio de H-W as probabilidades genotípicas para $A_1 A_1$, $A_1 A_2$ e $A_2 A_2$ satisfazem as relações apresentadas na Tabela 2.2 (Falconer & Mackay, 1996).

Tabela 2.2 Relação entre probabilidades dos alelos e genótipos

	Alelos		Genótipos		
	A_1	A_2	$A_1 A_1$	$A_1 A_2$	$A_2 A_2$
Probabilidades	p	q	p^2	$2pq$	q^2

De maneira geral, pode-se dizer que o equilíbrio de H-W 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, a união entre indivíduos de estratos diferentes são menos prováveis de ocorrer do que cruzamentos entre indivíduos de mesmo estrato, além disso, a endogamia pode conduzir a desvios do equilíbrio (Sham, 1998).

2.4.2 Desequilíbrio de Ligação (entre locos)

O desequilíbrio de ligação (D) é definido como qualquer desvio das probabilidades alélicas conjuntas em relação às probabilidades alélicas esperadas sob independência, indicando a

existência de uma associação entre dois locos, neste caso, loco A com alelos A_i e loco B com alelos B_j . Quando a população está em equilíbrio de ligação a expressão é dada por:

$$P(A_i B_j) = P(A_i)P(B_j), \quad (2.8)$$

onde $P(A_i B_j)$ representa a probabilidade conjunta do haplótipo $A_i B_j$; $P(A_i)$ e $P(B_j)$ representam as probabilidades marginais dos alelos A_i e B_j , respectivamente. Este equilíbrio define independência entre locos no mesmo cromossomo e pode ser chamado de independência vertical (entre locos) para diferenciar do equilíbrio de H-W, uma independência horizontal (na constituição dos alelos dentro do loco). Caso haja desequilíbrio de ligação, este é medido entre os alelos A_i e B_j , sendo que esse desvio pode ser representado por:

$$D = P(A_i B_j) - P(A_i)P(B_j). \quad (2.9)$$

O estado de equilíbrio se dá quando $D = 0$ e é chamado de equilíbrio de ligação.

Considerando o modelo de desequilíbrio baseado em dois locos dialélicos A com os alelos A_1 e A_2 , e B , com os alelos B_1 e B_2 , os possíveis haplótipos são: $A_1 B_1$, $A_1 B_2$, $A_2 B_1$, $A_2 B_2$. Ao considerar o haplótipo $A_1 B_1$, a medida do desequilíbrio gamético é dada por:

$$D_{A_1 B_1} = P(A_1 B_1) - P(A_1)P(B_1). \quad (2.10)$$

O desequilíbrio de ligação também pode ser escrito em função dos quatro alelos envolvidos como (Ott, 1999):

$$D = P(A_1 B_1) P(A_2 B_2) - P(A_1) P(B_1) P(A_2) P(B_2),$$

com limites dados por:

$$D_{min} = \max(-P(A_1)P(B_1), -P(A_2)P(B_2))$$

e

$$D_{max} = \min(P(A_1)P(B_2), P(A_2)P(B_1)).$$

Este parâmetro de desequilíbrio de ligação (D) é também conhecido como medida de associação alélica entre locos no mesmo cromossomo e tem sido explorado em estudos genéticos para se melhorar a localização de genes por procedimentos de mapeamento refinado (Cardon & Abecasis, 2000).

2.5 Efeitos Genéticos

Efeitos genéticos (ou valores genéticos) são as unidades métricas pelas quais uma característica é medida. O fenótipo é uma (ou conjunto de) característica(s) que um indivíduo possui e, em geral, ele é resultado de interações entre genótipo (constituição genética do indivíduo) e o ambiente (Falconer & Mackay, 1996). A equação a seguir expressa este relacionamento:

$$Y = \mu + G + e \quad (2.11)$$

em que Y representa os valores fenotípicos (colesterol, glicemia, entre outros) e é definido em termos de valores quantitativos, μ é a média populacional da resposta Y , G é o valor ou efeito genotípico exercido sobre Y e e é o componente de erro, que representa todas as circunstâncias não genéticas que influenciam o valor fenotípico. É conhecido que na procura de determinantes genéticos que regulem a variável resposta Y , mapas de marcadores moleculares do genoma humano podem auxiliar na sua identificação.

Os valores Y , G e e podem ser expressos em qualquer unidade que represente uma propriedade biológica que possa ser medida de forma discreta ou contínua, tal como peso, altura, teor de gordura, pressão arterial, número de grãos, etc. Quando não existe efeito do

ambiente (modelado no fator e) sobre determinada característica o fenótipo apresentado é idêntico para todos os indivíduos que possuem o mesmo genótipo, diferindo pelos locos em G .

O efeito genético linear está, em geral, associado a uma variável categorizada assumindo valores 0, 1, 2, indicando o genótipo do indivíduo, por exemplo, A_1A_1 , A_1A_2 , A_2A_2 , respectivamente. Vale destacar que no caso dialélico o componente genético de um gene corresponde a um fator em 3 níveis e, portanto, tem-se até 2 graus de liberdade para estudar um efeito. Assim,

$$G = G_a + G_d \quad (2.12)$$

onde G_a representa o efeito aditivo, isto é, um valor fenotípico que pode ser predito linearmente por meio do número de alelos de um certo tipo que definem o genótipo; e G_d é o efeito de dominância, ou seja, o valor fenotípico que não pode ser explicado linearmente (resíduo genético devido ao efeito de interação entre os alelos A_1 e A_2 no mesmo loco). Tanto o efeito aditivo como de dominância foram descritos e ilustrados na seção 2.1 (Figuras 2.3, 2.4 e 2.5). O valor genotípico de um indivíduo não é mensurável na prática, exceto quando estamos interessados em um loco onde os genótipos são fenotipicamente distinguíveis.

O seguinte modelo linear tem sido adotado para representar efeitos genéticos como componentes do valor esperado de Y :

$$Y_j = \mu + \underbrace{aX_{a_j} + dX_{d_j}}_G + e_j \quad (2.13)$$

em que Y_j é o fenótipo do indivíduo j ; μ é a média fenotípica geral, e_j é variável aleatória definindo o erro e

$$X_{a_j} = \begin{cases} -1 & , \text{ se o indivíduo } j \text{ carregar o genótipo } A_1A_1 \\ 0 & , \text{ se o indivíduo } j \text{ carregar o genótipo } A_1A_2 \\ 1 & , \text{ se o indivíduo } j \text{ carregar o genótipo } A_2A_2 \end{cases}$$

e

$$X_{d_j} = \begin{cases} 0 & , \text{ se o indivíduo } j \text{ carregar o genótipo } A_1A_1 \\ 1 & , \text{ se o indivíduo } j \text{ carregar o genótipo } A_1A_2 \\ 0 & , \text{ se o indivíduo } j \text{ carregar o genótipo } A_2A_2 \end{cases}$$

neste caso, tem-se:

$$E(Y_i | G_j = A_1A_1) = \mu_{A_1A_1} = \mu - a$$

$$E(Y_i | G_j = A_1A_2) = \mu_{A_1A_2} = \mu + d$$

$$E(Y_i | G_j = A_2A_2) = \mu_{A_2A_2} = \mu + a$$

logo,

$$a = \frac{\mu_{A_2A_2} - \mu_{A_1A_1}}{2} \quad (2.14)$$

$$\begin{aligned} d &= \mu_{A_1A_2} - \mu \\ &= \mu_{A_1A_2} - \left(\frac{\mu_{A_1A_1} + \mu_{A_2A_2}}{2} \right) \end{aligned} \quad (2.15)$$

Na Figura 2.8 estes efeitos são representados no gráfico de perfis de médias fenotípicas em função da classe genotípica.

A Figura 2.9 indica a escala de valores dos genótipos A_1A_1, A_1A_2, A_2A_2 , com ($\mu = 0$).

Se $d = 0$, não há dominância de nenhum alelo em relação a outro e a interação alélica é denominada aditiva. Caso $d > 0$, A_1 domina A_2 e caso $d < 0$, A_2 domina A_1 . Se $d = a$, indica interação alélica de dominância completa (ou superdominância) de A_1 em relação a A_2 e se, em oposição, $d = -a$, há dominância completa de A_2 em relação a A_1 . O grau de

dominância é obtido pela razão $\frac{d}{a}$ cujo valor dá uma ideia do tipo de interação alélica para o caráter estudado. A Tabela 2.3 apresenta, resumidamente, a escala de valores genotípicos.

Figura 2.8- Perfis de médias fenotípicas

A_2A_2		A_1A_2	A_1A_1
↓			
$-a$	0	d	a

Figura 2.9- Valores Genotípicos

Tabela 2.3 - Tipos de Interações Alélicas

$ d / a = 0$	$ d / a = 1$	$0 < d / a < 1$	$ d / a > 1$
Aditiva	Dominância Completa	Dominância Parcial	Sobredominância

2.6 Análise de doenças Complexas

Doenças complexas ou comuns, ou fenótipos quantitativos, são aquelas que afetam 30% da população humana e são, provavelmente, decorrentes da ação de múltiplos locos, do

ambiente e de possíveis interações, nenhum dos quais é individualmente necessário ou suficiente para a expressão da doença, logo, não seguem o padrão de herança Mendeliano (Birren et al., 1999). Estas desordens podem ser poligênicas (múltiplos genes envolvidos) ou multifatoriais (múltiplos genes interagindo com o meio ambiente). Ao contrário de fenótipos monogênicos, onde a variação genética de um fenótipo é devida principalmente à variação de um alelo em um único loco, a expressão fenotípica nestes casos poligênicos é devida à ação de dois ou mais locos. A interação destes múltiplos locos da doença, seus efeitos resultantes e a expressão fenotípica é usualmente explicada por modelos genéticos “aditivos” ou “multiplicativos”. Em outras palavras, o efeito de dois ou mais locos pode ser resultado da soma de seus efeitos marginais (aditivo) ou os locos podem interagir de tal maneira que resulta em um risco ainda maior do que aquele gerado independentemente por cada um dos locos (Haines & Pericak, 1998).

Nas duas últimas décadas, as causas genéticas de doenças complexas comuns têm sido cada vez mais enfatizadas como um meio para compreender melhor sua patogênese, com o objetivo da melhoria de estratégias preventivas, de ferramentas de diagnóstico e de tratamento. Um esforço considerável está sendo despendido na tentativa de detectar locos genéticos que contribuem para doenças complexas. Estudos de associação (caso-controle) e de ligação (dados de famílias) compõem as duas estratégias de coletas de dados dominantes na literatura: estudos de associação são, em geral, adotados com o objetivo de encontrar genes comuns que predisõem a doenças na população geral; e estudos de ligação têm foco em doenças com segregação familiar e muitas vezes em identificar genes raros que cosegregam em poucas famílias. Embora ambas as estratégias tenham vantagens convincentes, análises de associação são mais amplamente feitas devido à maior facilidade de coleta de dados.

As doenças complexas ou comuns têm de moderada a alta evidência genética e participação de fatores ambientais e, em geral, apresentam agregação familiar de casos, o que justifica o uso de delineamentos com famílias para seus estudos. A principal causa de mortalidade em países desenvolvidos é a doença arterial coronária (DAC), sendo responsável não apenas por uma parcela significativa dos orçamentos de saúde destes países, mas também por uma perda significativa de anos de vida produtiva nos indivíduos acometidos (Russel et al., 1998). Logo, por conta desta e de outras doenças complexas é válido o interesse por estudos de ligação. Encontrar vários casos de doenças complexas em qualquer grande família não é muito difícil, desde que estas desordens sejam comuns. Exemplos de doenças complexas comuns que são agregadas por famílias são: glicemia, hipertensão, obesidade, doenças cardiovasculares e muitos tipos de câncer.

2.7 Delineamentos com famílias

Estudos com membros de uma mesma família são a base para a determinação da etiologia genética de um fenótipo complexo. A abordagem com estudo de famílias implica a captura de informações sobre a doença em conjuntos específicos de parentes (por exemplo, irmãos, pai-filho, primo de primeiro grau, meio-irmãos), ou em famílias inteiras (famílias nucleares ou famílias estendidas) (Haines & Pericak, 1998).

As famílias que serão incluídas em um estudo de doenças complexas são selecionadas, em geral, por meio da identificação de um caso- referência (probando), que é o indivíduo afetado, e por meio deste os demais parentes (afetados e não afetados) são identificados e convidados a fazerem parte do estudo. Os casos-referência podem ser identificados a partir de registros de doenças que muitas vezes são de base populacional, de

admissões consecutivas a uma clínica, ou de registros médicos sobre pacientes, que não são de base populacional e, portanto, podem não ser representativos do conjunto de pacientes. Nem todos os indivíduos afetados em um área geográfica podem ser casos-referência, e a proporção que pode ser identificada é chamada de “*Ascertainment Probability*”. Estudos com famílias tipicamente usam esta regra para identificar os casos-referência e seus familiares. Estas regras envolvem uma estrutura de amostragem fixa, tal como a seleção do núcleo familiar no qual está inserido o caso-referência, ou uma amostragem variável na qual é selecionado o núcleo familiar e também toda a linha parental do caso-referência (família estendida). O uso deste tipo de amostragem (pseudo aleatória) pode afetar a distribuição dos fenótipos observados nas famílias amostradas e o desprezo destas regras de amostragem durante as análises pode causar conclusões errôneas com respeito à base genética e herança dos fenótipos. Várias “correções” no procedimento de inferência, para que haja uma correta interpretação genética dos dados de famílias amostradas, têm sido propostos por Fisher (1934), Morton (1959), Li and Mantel (1968) e Birren et al., (1999).

Metodologias estatísticas que envolvem “*Ascertainment Probability*”, embora importantes na análise de segregação de um fenótipo não são, em geral, consideradas na análise de ligação. Isto porque o objetivo da análise de segregação é determinar se existe evidência estatística significativa de que fatores genéticos que estão segregando em algumas famílias contribuem para a etiologia de um fenótipo e não estimar prevalências de medidas de risco, por exemplo. Por esta razão, as famílias são amostradas de forma diferenciada para conter um grande número de parentes afetados. Deste modo, a análise de ligação está preocupada com a co-segregação (tendência de genes estreitamente ligados e marcadores genéticos serem herdados em conjunto) de um fenótipo de interesse com alelos de um marcador, uma análise que é independente da distribuição do fenótipo na população geral.

Na verdade, o objetivo em amostragem de famílias para a análise de ligação é propositadamente (e não aleatoriamente) escolher famílias que maximizam a chance de detecção de ligação (Birren et al., 1999). Neste sentido, o plano amostral que mais tem sido usado é a amostragem aleatória de indivíduos de uma população alvo, os quais abrirão a coleta de dados de cada família.

A seção 4.2 trata da análise de dados de famílias para a finalidade de mapeamento de genes e explora a estrutura de correlação entre os indivíduos da mesma família, bem como a resposta média dos indivíduos independente da agregação familiar.

2.8 Genética de Populações Humanas

O estudo da genética de populações é baseado na descrição da variabilidade genética existente e na investigação dos mecanismos evolutivos que a mantêm ou modificam. A variação genética é um processo natural, gerada a partir de mutações e recombinação da variação pré-existente (Futuyma, 1998). No caso da espécie humana, a evolução, isto é, as mudanças nas frequências alélicas ao longo do tempo, é decorrente dos fatores evolutivos, como fluxo gênico, deriva genética¹ e seleção natural (em menor escala) e dependente tanto de processos biológicos como também demográficos e históricos (Chakravarti, 1999; Futuyma, 1998).

Regiões do genoma que apresentam variação e cujo alelo mais (menos) comum apresenta uma proporção populacional de no máximo 99% (mínimo 1%) são denominadas marcadores genéticos polimórficos. Os marcadores são uma importante ferramenta no entendimento das relações populacionais e da história evolutiva humana, principalmente no que tange à reconstrução de rotas e à determinação das origens de povoamento.

A avaliação da constituição genética de uma população permite fazer inferências sobre acontecimentos do passado, como efeitos de gargalos populacionais, migrações e expansões demográficas (Cavalli-Sforza et al., 1994).

A utilização de metodologias de análise de dados e mapas de marcadores genéticos apropriados permite avaliar a diversidade genética intra e interpopulações e ajuda na compreensão da evolução humana, além do mapeamento de genes envolvidos na regulação de doenças.

2.8.1 Marcadores no DNA - Microsatélites

Microsatélites ou STRs (do inglês, *Short Tandem Repeat Sequences*) são sequências formadas pela repetição ininterrupta de uma sequência simples, com no máximo seis bases nucleotídicas. Esses marcadores são encontrados nos genomas de inúmeros organismos e apresentam altos níveis de polimorfismos, o que os torna interessantes marcadores genéticos. Cada microsatélite constitui um loco genético multialélico, uma vez que seus alelos podem ser detectados e discriminados na população e possuem expressão codominante (Ferreira & Grattapaglia, 1995). Os mapas de marcadores microsatélites são utilizados nos estudos de famílias, pois promovem uma correlação entre indivíduos do mesmo núcleo familiar.

¹Deriva genética, ou derivação genética, é um mecanismo que ocorre ao acaso e atua em consonância com a seleção natural, modificando as características das espécies ao longo do tempo. É um processo estocástico, atuante sobre as populações, modificando a frequência dos alelos e a predominância de certas características na população. É mais frequente ocorrer em populações pequenas e as alterações induzidas poderão não ser adaptativas, mas podem ser benéficas ou até mesmo nulas.

Os microsátélites são bastante úteis em estudos de genética de populações humanas, com o objetivo de inferir alguns aspectos histórico-demográficos, como eventos de migração e mistura genética, pois são abundantes no genoma e possuem alta taxa de mutação, comparado com outros tipos de polimorfismos. Assim, em um curto período evolutivo um número elevado de eventos mutacionais acumulados pode ser observado em populações divergentes (Callegari- Jacques et al., 2003).

2.8.2 Marcadores no DNA - SNPs

São marcadores mais amplamente distribuídos no genoma humano, estimando-se que eles sejam responsáveis por cerca de 90% das variações nas sequências de base no DNA (Collins et al., 1998). Refletem mutações ocorridas no passado e que, em sua maioria, foram eventos únicos. Dessa forma, dois indivíduos que compartilham um determinado alelo apresentam uma herança evolutiva comum (Stoneking, 2001). Esses fatos, associados à ampla ocorrência de SNPs no genoma humano, quando comparada à de outros marcadores, os tornam ideais para análises genômicas.

Os SNPs humanos normalmente não são compartilhados com outros primatas, mas a maioria (~85%) é comum a todas as populações humanas, com apenas 15% ou menos sendo exclusivos de algumas populações (Barbujani et al., 1997). Assim, a maior parte da variação do genoma humano é devida, principalmente, a variações dentro das populações, e não a variações entre populações, e, conseqüentemente, apenas uma pequena fração da diversidade genética é exclusiva a alguns grupos. Populações africanas, entretanto, apresentam uma diversidade maior do que outros grupos, o que reflete a sua antiguidade em relação a populações nativas da Europa, Ásia e Américas (Chakravarti, 1999).

Os SNPs são marcadores interessantes, pois podem marcar ou rastrear a presença de outras diferenças genéticas que geram fenótipos de interesse. Além disso, eles são úteis ao estudo de taxas de mutação e história evolutiva (Pedrosa, 2006).

Há SNPs em alta frequência em todas as populações humanas e eles são abundantes em todo o genoma. Estimativas sugerem que a taxa de SNPs é de 1 a cada 200 – 300 pares de bases do genoma. Portanto, considerando todo o genoma humano, deve existir ao menos cerca de 15 milhões de SNPs (Stephens et al., 2001). Deste modo, eles são amostradores do genoma com cobertura densa. A Figura 2.10 compara a amostragem do genoma com microsátélites (medida de distância citogenética) e com SNPs (medida de distância física em número de bases), onde percebe-se que os microsátélites aparecem em menor quantidade e abrangem grandes regiões cromossômicas ao contrário dos SNPs.

Os mapas de SNPs atuais consistem de um milhão de SNPs (por exemplo, a plataforma da Affymetric 6.0) e são utilizados principalmente nos estudos de associação (caso-controle) com indivíduos não parentes que compartilham pequenas regiões genômicas em comum. Os dados de famílias têm sido comumente analisados utilizando marcadores microsátélites de efeito aleatório. Com a utilização dos marcadores SNPs para análise dos dados de famílias dois problemas surgem: se o efeito do SNP deve ser modelado como um fator fixo ou aleatório e, além disso, o problema de múltiplos testes ganha destaque, pois simultaneamente um grande número de modelos precisam ser ajustados, e isto demanda um grande tempo computacional. Várias estratégias para contornar este problema têm sido propostas na literatura, entre elas a de Aulchenko et al. (2007) na qual, para detectar o efeito do SNP utilizando dados de famílias é feita uma análise de mapeamento de genes em vários estágios usando modelos mistos e regressão clássica, e o efeito do SNP é modelado como um fator fixo e associado a componente

residual. A descrição completa do modelo em multiestágios bem como suas limitações e vantagens é apresentada na seção 3.2.4.

Figura 2.10- Comparação da amostra com microsatélites e com SNPs

Capítulo 3

Modelos Mistos

De maneira geral, um modelo linear que apresenta somente fatores (variáveis independentes) de efeitos fixos, que modelam o valor esperado da variável dependente, além do erro experimental, que é aleatório, é denominado modelo de efeitos fixos. Esse tipo de modelo tem sido amplamente estudado, existindo inúmeros livros abordando seus aspectos teóricos e aplicados, em vários níveis de complexidade (pode-se citar, Graybil, 1976 e Neter et al., 1985).

O modelo que apresenta apenas fatores de efeitos aleatórios, exceto a constante μ , que é sempre fixa, é denominado modelo de componentes de variância (Searle et al, 1992).

Estendendo a classe de modelos lineares tem-se um modelo misto, que é aquele que apresenta tanto fatores de efeitos fixos como aleatórios, além do erro experimental e da constante μ . Estes modelos são muito flexíveis pois permitem estruturar os componentes do valor esperado, da variância e covariância associados à variável dependente sob estudo (Searle et al., 1992 e Demidenko, 2004).

Outro motivo para adotar um modelo linear misto é a possibilidade de fazer a predição de efeitos aleatórios, na presença de efeitos fixos, através dos BLUP's (*Best Linear Unbiased Prediction*) que são de grande valia em genética e estratégias de melhoramento. Além disso, o modelo misto é útil para modelar medidas repetidas e, por vezes, é caracterizado em termos de um modelo hierárquico. Ocasionalmente, o modelo

misto é usado para análise de *clusters* ou painéis de dados e, muito frequentemente, para modelar dados longitudinais (Demidenko, 2004).

Na estatística clássica uma típica suposição é que as observações advindas da mesma população geral são independentes e identicamente distribuídas. Os dados analisados sob modelos mistos têm uma complexa e hierárquica estrutura com múltiplos níveis. Observações entre níveis ou *clusters* são assumidas independentes, mas observações dentro de cada *cluster* são dependentes, porque elas pertencem à mesma subpopulação. Consequentemente, há duas fontes de variação: entre *clusters* e dentro dos *clusters*.

Em geral, o modelo linear de efeitos mistos (ou modelo misto) é escrito como

$$Y_i = X_i\beta + Z_i\gamma_i + \varepsilon_i, \quad i = 1, \dots, c \quad (3.1)$$

em que Y_i representa um vetor ($n_i \times 1$) de respostas da i -ésima unidade experimental ($i = 1, 2, \dots, c$); X_i é uma matriz ($n_i \times p$) de especificação dos efeitos fixos, a qual é conhecida e de posto completo; β é um vetor ($p \times 1$) de parâmetros dos efeitos fixos; Z_i é uma matriz ($n_i \times q$) de especificação dos efeitos aleatórios, também conhecida e de posto completo; γ_i é um vetor ($q \times 1$) de variáveis latentes, comumente denominadas efeitos aleatórios, que refletem o comportamento individual da i -ésima unidade experimental; ε_i é um vetor ($n_i \times 1$) de erros aleatórios (Searle et al., 1992).

Fazendo,

$$Y = (Y_1^T, \dots, Y_c^T)^T,$$

$$X = (X_1^T, \dots, X_c^T)^T,$$

$$Z = \text{diag}(Z_1, \dots, Z_c),$$

$$\gamma = (\gamma_1^T, \dots, \gamma_c^T)^T,$$

$\varepsilon = (\varepsilon_1^T, \dots, \varepsilon_c^T)^T$, o modelo (3.1) pode ser escrito como,

$$Y = \mathbf{X}\beta + \mathbf{Z}\gamma + \varepsilon. \quad (3.2)$$

Em geral, assume-se que $E(\gamma) = 0$, $E(\varepsilon) = 0$ com

$$\text{Cov} \begin{pmatrix} \gamma \\ \varepsilon \end{pmatrix} = \begin{pmatrix} \Delta & 0_{cq \times n} \\ 0_{n \times cq} & \Sigma \end{pmatrix}, \quad (3.3)$$

em que $\mathbf{0}_{k_1 \times k_2}$ representa uma matriz nula de ordem $k_1 \times k_2$, Δ e Σ são matrizes quadradas de ordens cq e $n = \sum_{i=1}^c n_i$, positivas definidas, que correspondem, respectivamente, às matrizes de covariância dos vetores γ e ε . No modelo (3.2), os efeitos fixos são usados para modelar o valor esperado da variável resposta Y , enquanto que os efeitos aleatórios são utilizados para modelar sua estrutura de covariâncias. Usualmente, assume-se que γ e ε seguem uma distribuição normal cq e (n) -variada, respectivamente.

Fazendo $\zeta = \mathbf{Z}\gamma + \varepsilon$ obtém-se o modelo marginal,

$$Y = \mathbf{X}\beta + \zeta, \quad (3.4)$$

com

$$\zeta = \begin{bmatrix} \zeta_1 \\ \vdots \\ \zeta_c \end{bmatrix} = \begin{bmatrix} \varepsilon_1 + Z_1\gamma_1 \\ \vdots \\ \varepsilon_c + Z_c\gamma_c \end{bmatrix},$$

em que ζ tem distribuição normal n -variada com vetor de médias 0_n e matriz de covariâncias

$$\mathbf{V} = \mathbf{Z}\Delta\mathbf{Z}^T + \Sigma. \quad (3.5)$$

Assim, sob (3.4) o valor esperado de Y é $E(Y) = X\beta$ e sua matriz de covariâncias é V . Usualmente, Δ e Σ não são totalmente conhecidas, então para estimá-las é comum colocar um parâmetro de dispersão σ^2 em evidência, ou seja, $\Delta = \sigma^2 D$ e $\Sigma = \sigma^2 R$, sendo D e R matrizes positivas definidas. Assim,

$$V = \sigma^2 (ZDZ^T + R). \quad (3.6)$$

Diferentes estruturas para D e R podem ser encontradas na literatura (por exemplo, Searle et al., 1992; Wolfinger, 1993).

Pode-se também definir,

$$M = \sigma^2 V^{-1} = (ZDZ^T + R)^{-1}, \quad (3.7)$$

tal que M é simétrica. Uma fórmula de inversão útil é,

$$M = R^{-1} - R^{-1}ZC^{-1}Z^T R^{-1}, \quad (3.8)$$

com

$$C = D^{-1} + Z^T R^{-1}Z. \quad (3.9)$$

Note que C é uma matriz quadrada de dimensão q . É mais fácil calcular M (3.8) do que inverter V . Além disso, é melhor calcular $C^{-1} = [I_q + DZ^T R^{-1}Z]^{-1}D$ do que inverter C diretamente de (3.9) (Hilden-Minton, 1995). Contudo, em situações em que $Z = I$ esta facilidade não se aplica.

Neste trabalho as matrizes M e C são usadas, posteriormente, para estimar o BLUE e o BLUP de γ e β no contexto do ajuste de modelos genéticos.

3.1 Inferência Estatística

Existem vários métodos de estimação dos parâmetros (γ e β) do modelo (3.2) disponíveis na literatura, entre eles os métodos Bayesianos, o método da Máxima Verossimilhança (MV), Máxima Verossimilhança Restrita (MVR) e o método de Mínimos Quadrados (MQ), que podem ser encontrados em Searle et al. (1992) e Demidenko (2004), por exemplo.

Harville (1976) obteve o melhor estimador linear não viesado (“BLUE- *Best Linear Unbiased Estimator*”) para β e o melhor preditor não viesado (“BLUP- *Best Linear Unbiased Predictor*”) para o vetor de efeitos aleatórios γ . Sejam $\hat{\gamma}$ e $\hat{\beta}$, respectivamente, o BLUP e o BLUE de γ e β . Como principais resultados envolvendo estimadores e preditores obtidos sob o modelo (3.2) têm-se (Harville, 1990):

- $\hat{\gamma}$ e $\hat{\beta}$ são funções lineares de Y ;
- $E[\hat{\gamma} - \gamma] = 0$ e $E[\hat{\beta} - \beta] = 0$, ou seja, $\hat{\gamma}$ e $\hat{\beta}$ são não viesados, respectivamente, para γ e β ;
- $\hat{\gamma}$ é o melhor preditor de γ dentro da classe dos preditores lineares, no sentido que minimizam o erro quadrático médio (EQM) de estimação $E[(\hat{\gamma} - \gamma)^T(\hat{\gamma} - \gamma)]$;
- $\hat{\beta}$ é o melhor estimador de β dentro da classe dos estimadores lineares no sentido que minimizam o EQM de estimação $E[(\hat{\beta} - \beta)^T(\hat{\beta} - \beta)]$;
- Na classe dos estimadores lineares, o BLUP $\hat{\gamma}$ maximiza a correlação entre γ^T e qualquer outro preditor γ^* , cujo valor máximo é $\rho_{(\gamma^T, \hat{\gamma})}$;
- $Cov[\hat{\gamma}, \hat{\gamma}^T] = Var(\hat{\gamma})$.

Outras propriedades de $\hat{\beta}$ e $\hat{\gamma}$ são dadas em Henderson (1975) e em McLean et al. (1991).

Como convenção, estimadores de efeitos aleatórios são chamados de preditores enquanto para os de efeitos fixos o termo estimadores se mantém. Discussões a respeito do emprego do termo “estimadores” tanto para efeitos fixos como para efeitos aleatórios são feitas em Robinson (1991). Além disso, este autor descreve o BLUP e o BLUE como os “Estimadores de Máxima Verossimilhança” de γ e β obtidos por meio da densidade conjunta do vetor aleatório (Y, γ) , sob a suposição de normalidade de γ e ε , tratada como uma “função de verossimilhança”. Hodges (1994) demonstrou como obter o BLUE e BLUP a partir da equação do modelo misto (ver Apêndice A.1).

A partir da obtenção do BLUE e o BLUP apresentada no apêndice A.1, tem-se que

$$\hat{\beta} = (X^T M X)^{-1} X^T M Y = (X^T V^{-1} X)^{-1} X^T V^{-1} Y, \quad (3.10)$$

$$\hat{\gamma} = C^{-1} Z^T R^{-1} (Y - X \hat{\beta}) = D Z^T Q Y, \quad (3.11)$$

com

$$Var(\hat{\beta}) = \sigma^2 [X^T M X]^{-1}, \quad (3.12)$$

$$Var(\hat{\gamma} - \gamma) = \sigma^2 [D - D Z^T Q Z D], \quad (3.13)$$

e

$$Cov(\hat{\beta}, \hat{\gamma} - \gamma) = -[X^T M X]^{-1} X^T M Z D, \quad (3.14)$$

sendo,

$$Q = M - M X (X^T M X)^{-1} X^T M, \quad (3.15)$$

como em (A.13), e $\mathbf{Q}$ simétrica de ordem n , tal que $\mathbf{QM}^{-1}\mathbf{Q} = \mathbf{Q}$, $\mathbf{QX} = \mathbf{0}$ e o posto de $\mathbf{Q}$ é $n - p$.

Discussões a respeito do processo de estimação quando $\mathbf{X}$ não tem posto completo e as matrizes $\mathbf{D}$ e $\mathbf{R}$ não são positivas definidas podem ser encontradas em Harville (1976). O BLUE e o BLUP, sob a formulação de modelos mistos com efeitos genéticos são utilizados no procedimento em multiestágios proposto por Aulchenko et al. (2007) para modelar dados de famílias e plataformas de SNPs. Logo, conhecer as propriedades das matrizes descritas nesta seção e no Apêndice A.1 são de grande valia para a simulação, implementação computacional e análise dos dados considerados neste trabalho.

3.2 Análise de Resíduos

Para a aplicação e entendimento da limitação do modelo de Aulchenko et al. (2007) faz-se necessário analisar os resíduos de modelos mistos. Resíduos, em geral, são utilizados para avaliar a validade das suposições impostas por modelos estatísticos. Por exemplo, no caso linear normal clássico, utilizam-se os resíduos padronizados para verificar homocedasticidade, existência de pontos discrepantes e normalidade. Para ajuste do modelo (3.1) é possível utilizar diferentes tipos de resíduos.

3.2.1 Tipos de Resíduos

Considerando a equação (3.1) do modelo misto existem três tipos de vetores de erros que podem ser definidos (Hilden-Minton, 1995):

- Erros condicionais, ε ,
- efeitos aleatórios, $Z\gamma$ (ou γ para $Z = \mathbf{I}$),

- erros marginais: $\zeta = \mathbf{Z}\gamma + \varepsilon$.

Cada tipo de vetor de erro está associado a um diagnóstico de interesse, mas estes vetores não são diretamente observáveis. Na seção 3.1 foram definidos o BLUE de β e o BLUP de γ . Uma vez estimado um modelo misto com BLUE e BLUP, parece natural definir resíduos em termos destes estimadores, tal como,

$$\hat{\varepsilon} = Y - \mathbf{X}\hat{\beta} - \mathbf{Z}\hat{\gamma}, \quad (3.16)$$

denominado resíduo condicional. Também pode-se definir

$$d = \mathbf{Z}\hat{\gamma}, \quad (3.17)$$

chamado de resíduo de efeito aleatório. Supondo que há interesse em checar a linearidade da resposta para as variáveis em X utiliza-se o modelo marginal,

$$\hat{\zeta} = Y - \mathbf{X}\hat{\beta}, \quad (3.18)$$

em que $\hat{\zeta}$ é resíduo composto, pois $\hat{\zeta} = d + \hat{\varepsilon}$, dado que $\zeta = \mathbf{Z}\gamma + \varepsilon$.

Note que o resíduo condicional avalia o comportamento das observações depois de subtraído os efeitos das covariáveis ($\mathbf{X}\hat{\beta}$) e do efeito aleatório ($\mathbf{Z}\hat{\gamma}$). Já nos resíduos marginais o comportamento das observações é avaliado depois de subtraído apenas os efeitos das covariáveis.

Utilizando (3.8) conclui-se que

$$\begin{aligned} \mathbf{RQY} &= \mathbf{RM}(I_n - \mathbf{X}[\mathbf{X}^T \mathbf{M} \mathbf{X}]^{-1} \mathbf{X}^T \mathbf{M}) \mathbf{Y} \\ &= \mathbf{RM}(Y - \mathbf{X}\hat{\beta}) \end{aligned}$$

$$\begin{aligned}
&= (I_n - \mathbf{Z}\mathbf{C}^{-1}\mathbf{Z}^T\mathbf{R}^{-1})(\mathbf{Y} - \mathbf{X}\hat{\beta}) \\
&= \mathbf{Y} - \mathbf{X}\hat{\beta} - \mathbf{Z}\mathbf{C}^{-1}\mathbf{Z}^T\mathbf{R}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\beta}) \\
&= \mathbf{Y} - \mathbf{X}\hat{\beta} - \mathbf{Z}\hat{\gamma}.
\end{aligned}$$

Assim,

$$\hat{\varepsilon} = \mathbf{R}\mathbf{Q}\mathbf{Y},$$

$$d = \mathbf{Z}\mathbf{D}\mathbf{Z}^T\mathbf{Q}\mathbf{Y},$$

$$\hat{\zeta} = \mathbf{M}^{-1}\mathbf{Q}\mathbf{Y},$$

e suas variâncias são, respectivamente,

$$\text{Var}[\hat{\varepsilon}] = \sigma^2\mathbf{R}\mathbf{Q}\mathbf{R}, \quad (3.19)$$

$$\text{Var}[d] = \sigma^2\mathbf{Z}\mathbf{D}\mathbf{Z}^T\mathbf{Q}\mathbf{Z}\mathbf{D}\mathbf{Z}^T, \quad (3.20)$$

$$\text{Var}[\hat{\zeta}] = \sigma^2\mathbf{M}^{-1}\mathbf{Q}\mathbf{M}^{-1}. \quad (3.21)$$

Resíduo puro é aquele que depende apenas das componentes fixas do modelo e do respectivo erro do qual ele é preditor. Já um resíduo que depende de dois ou mais erros é denominado resíduo confundido (Hilden- Minton, 1995). Sob a validade do modelo (3.2), tem-se também que

$$\hat{\varepsilon} = \mathbf{R}\mathbf{Q}\varepsilon + \mathbf{R}\mathbf{Q}\mathbf{Z}\gamma, \quad (3.22)$$

$$d = \mathbf{Z}\mathbf{D}\mathbf{Z}^T\mathbf{Q}\mathbf{Z}\gamma + \mathbf{Z}\mathbf{D}\mathbf{Z}^T\mathbf{Q}\varepsilon, \quad (3.23)$$

$$\hat{\zeta} - \zeta = -\mathbf{X}[\mathbf{X}^T\mathbf{M}\mathbf{X}]^{-1}\mathbf{X}^T\mathbf{M}\zeta. \quad (3.24)$$

De (3.22) e (3.23) conclui-se que $\hat{\varepsilon}$ e d são resíduos confundidos pela presença de γ e ε , respectivamente. Se $\mathbf{Z} \in \mathbb{C}(\mathbf{X})$, com $\mathbb{C}(\mathbf{X})$ representando o subespaço gerado pelas

colunas da matriz $\mathbf{X}$, então $\mathbf{QZ} = \mathbf{0}$ e, nesse caso, os resíduos são puros. Quando o interesse é verificar a suposição de normalidade para o erro ε , não é aconselhável utilizar $\hat{\varepsilon}$, porque ele é confundido por γ ; logo, quando γ se afasta muito da normalidade, $\hat{\varepsilon}$ pode não apresentar características de normalidade, mesmo quando ε segue uma distribuição normal (Nobre, 2004).

3.3 Modelos Estatísticos Genéticos

Na modelagem de dados de famílias os modelos mistos representam uma ferramenta estatística útil para a identificação de genes e mensuração dos seus efeitos sobre uma variável resposta (Blangero et al., 2004; de Andrade et al., 1999). Diferentes modelos estatísticos têm sido utilizados na análise destes dados no intuito de incorporar no modelo o efeito de diferentes tipos de marcadores, por exemplo, microsatélite e/ ou SNPs. A seguir são apresentadas diferentes formulações de modelos mistos usados em Genética.

3.3.1 Modelo Poligênico

Considerando variáveis quantitativas avaliadas em indivíduos e seus familiares, mesmo sem qualquer informação de marcadores moleculares é possível modelar o efeito genético utilizando a estrutura familiar, isto é, modelando apropriadamente componentes de variância e covariâncias entre os indivíduos. Sob os mesmos pressupostos do modelo (3.1) no contexto de dados genéticos, para a finalidade de mapeamento de genes, descreve-se o modelo poligênico definido por:

$$Y_f = \mu + \mathbf{X}_f\beta + g_f + e_f, \quad f = 1, \dots, F \quad (3.25)$$

em que Y_f representa um vetor ($n_f \times 1$) da variável fenotípica observada na f -ésima família; μ é média fenotípica geral, X_f é uma matriz de especificação dos efeitos fixos, β é o vetor de parâmetros dos efeitos fixos, a matriz Z_i de (3.1) equivale à matriz identidade, o efeito aleatório γ_i de (3.1) é denotado como variável aleatória g_f associada ao efeito poligênico e e_f é a variável aleatória definindo o erro. As variáveis aleatórias, g_f e e_f , são assumidas não correlacionadas e, em geral, normalmente distribuídas, com vetor de média zero e componentes de variância σ_g^2 e σ_e^2 , respectivamente. Em geral, o erro é suposto comum a cada indivíduo, enquanto o componente poligênico é compartilhado entre indivíduos, sendo proporcional ao seu grau de parentesco. A covariância entre as variáveis fenotípicas para os indivíduos i e i' é dada por:

$$Cov(y_{if}; y_{i'f'}) \begin{cases} \sigma_g^2 + \sigma_e^2 & \text{para } i = i', \\ 2\phi_{ii'}\sigma_g^2 & \text{para } i \neq i' \text{ e } f = f' \text{ (relacionados)}, \\ 0 & \text{para } i \neq i' \text{ e } f \neq f' \text{ (não-relacionados)}. \end{cases} \quad (3.26)$$

O parâmetro $2\phi_{ii'}$ é o coeficiente de relacionamento entre os indivíduos i e i' , sendo dado por $\left(\frac{1}{2}\right)^r$, r representando o grau de relacionamento. Deste modo, a variância de Y denotada por σ_y^2 é a soma dos componentes de variância poligênico e residual, $\sigma_y^2 = \sigma_g^2 + \sigma_e^2$. A função de verossimilhança considerando os dados dos membros de uma família é obtida a partir da distribuição normal multivariada, sendo $Y_f \sim Nn_f(\mu + X_f\beta; 2\phi_f\sigma_g^2 + I_f\sigma_e^2)$, com $2\phi_f$ a matriz de relacionamento conhecida, com elementos $2\phi_{ii'}$ e I_f a matriz identidade, correspondentes à família f .

3.3.2 Estimação da Herdabilidade

Como salientado anteriormente, as doenças complexas apresentam agregação familiar de casos e têm de moderada a alta evidência genética. Desta forma, faz-se necessário identificar as contribuições de fatores genéticos associados a fenótipos que descrevem tais doenças, o que é feito por meio do cálculo da medida de herdabilidade. A herdabilidade é um conceito importante em Genética sendo definida como a proporção da variância total da variável fenotípica que é devida a componentes genéticos. Sob o modelo misto poligênico (3.25) a variância fenotípica de Y (σ_y^2) pode ser particionada em componentes genéticos (σ_g^2) e associados ao termo de erro (σ_e^2), ou seja, $\sigma_y^2 = \sigma_g^2 + \sigma_e^2$. A herdabilidade, abreviada pelo símbolo h^2 , pode variar de 0 a 1, e quanto mais próxima de 1, mais herdável é esta característica, ou seja, mais importante é a contribuição do fator genético para sua manifestação. Pode ser escrita como,

$$h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2}, \quad (3.27)$$

que corresponde ao coeficiente de correlação intra-classe, isto é, ao coeficiente de correlação entre os fenótipos de indivíduos relacionados, da mesma família.

Sob o modelo (3.25) há interesse no teste de herdabilidade poligênica dado por $H_0 : \sigma_g^2 = 0$ vs $H_1 : \sigma_g^2 > 0$, o qual é feito com base, por exemplo, na estatística razão de verossimilhanças que, para o caso do modelo misto poligênico segue uma distribuição mistura de qui-quadrados $\frac{1}{2}\chi_1^2 + \frac{1}{2}\chi_0^2$ (Self & Liang, 1987).

Na simulação dos dados deste trabalho diferentes valores para a herdabilidade foram considerados com a finalidade de criar diferentes contextos para a análise da proposta de Aulchenko et al. (2007).

3.3.3 Modelo Oligogênico

Caso, no modelo (3.25), a herdabilidade poligênica seja significativa para a variável resposta Y , justifica-se o interesse na identificação dos específicos genes que a regulam. Para identificar estes genes é necessário adicionar a informação de marcadores moleculares, e extensões do modelo poligênico (3.25) podem ser feitas introduzindo um componente aleatório adicional. Para esta finalidade, o seguinte modelo misto é adotado,

$$Y_f = \mu + \mathbf{X}_f\beta + q_f + g_f + e_f, \quad f = 1, \dots, F \quad (3.28)$$

em que exceto q_f os demais termos do modelo estão definidos como em (3.25) e q_f é a variável aleatória que define o efeito de um oligogene ou variante genética de efeito individual maior sobre o fenótipo, sendo assumida como não correlacionada com as demais variáveis (do poligene e erro) e seguindo, em geral, uma distribuição Normal com média zero e componente de variância σ_q^2 . Em geral, este oligogene é conhecido na literatura como QTL (do inglês, *Quantitative Trait Locus*). Para se estudar o efeito de específicos QTLs sobre a variável resposta Y , sob o modelo (3.28) usa-se a informação de dados de marcadores microsatélites para se estimar a proporção de alelos compartilhados, no loco de QTL, por dois indivíduos relacionados. Esta proporção de alelos é conhecida como IBD (idênticos por descendência) e assume os valores 0, $\frac{1}{2}$ e 1 se os dois indivíduos compartilham 0, 1 ou 2 alelos idênticos no marcador considerado (Almasy and Blangero, 1998).

Os componentes da matriz de covariância dados em (3.26) podem ser estendidos para aplicação no modelo oligogênico e, neste caso, a covariância entre os indivíduos i e i' é dada por:

$$Cov(y_{if}; y_{i'f'}) \begin{cases} \sigma_g^2 + \sigma_q^2 + \sigma_e^2 & \text{para } i = i', \\ 2\phi_{ii'}\sigma_g^2 + \hat{\pi}_{ii'}\sigma_q^2 & \text{para } i \neq i' \text{ e } f = f' (\text{relacionados}), \\ 0 & \text{para } i \neq i' \text{ e } f \neq f' (\text{n\~ao} - \text{relacionados}), \end{cases} \quad (3.29)$$

em que $\hat{\pi}_{ii'}$ é a proporção de alelos compartilhados entre os indivíduos i e i' no loco de QTL (ou de um específico marcador). O teste do efeito do QTL é dado pela hipótese $H_0: \sigma_q^2 = 0$ vs $H_1: \sigma_q^2 > 0$ e é realizado com base na estatística razão de verossimilhanças, que, neste caso, segue uma distribuição de qui-quadrados $\frac{1}{2}\chi_2^2 + \frac{1}{2}\chi_1^2$ (Self & Liang, 1987; Giampaoli e Singer, 2009).

3.3.4 Modelo em Multiestágios

O modelo oligogênico (3.28) tem sido utilizado no mapeamento de genes em doenças complexas para dados de famílias, pois modela a estrutura familiar. Além disso, o modelo (3.28) faz a pesquisa por regiões muito grandes do genoma, por isso, sempre esteve vinculado a dados de marcadores moleculares do tipo microsatélite. Para dados com marcadores do tipo SNP a aplicação direta do modelo (3.28) para identificação de genes parece não ser eficiente, pois o marcador SNP é formado por um único nucleotídeo (a unidade física na sequência de DNA) e seus dados sempre estiveram ligados a estudos caso-controle com indivíduos não relacionados (ver Ziegler et al., 2008 para uma discussão detalhada).

Um dos pontos polêmicos na literatura (Blangero, 2004; Almasy and Blangero, 2008) é a modelagem do SNP no mapeamento de genes, pois questiona-se se o mesmo deve ser modelado estatisticamente como um componente fixo ou aleatório. Neste contexto, acredita-se que os dados de SNP, isoladamente, não modelam estrutura familiar, isto é, não

explicam a correlação entre indivíduos relacionados, sendo mais apropriado modelar seu efeito no componente fixo (Aulchenko et al., 2007; Amin et al., 2007). Contudo, é interessante notar que conjuntos adjacentes de SNPs, analisados juntos, podem trazer informações sobre regiões genômicas (maiores) associadas a doenças. Neste sentido, uma possível estratégia de análise de dados é mensurar o efeito de regiões cromossômicas por meio de blocos de SNPs, denominados haplótipos. Tais haplótipos quando identificados poderiam ser modelados via (3.28), isto é, considerando o efeito do bloco de SNPs como aleatório.

Considerando dados de famílias, dentre as propostas para modelagem do efeito do SNP como componente fixo a mais utilizada é a de Aulchenko et al. (2007) que propõem uma análise de mapeamento de genes em multiestágios. No passo inicial os dados são analisados considerando o seguinte modelo poligênico:

$$Y_i = \mu + X_{ij}\beta_j + g_i + e_i, i = 1, \dots, n; j = 1, \dots, M, \quad (3.30)$$

onde Y_i é o fenótipo do i -ésimo indivíduo, μ é a média geral fenotípica, X_{ij} é o valor da j -ésima covariável ou efeito fixo para o i -ésimo indivíduo, β_j é um estimador do j -ésimo efeito fixo ou covariável, g_i é a variável aleatória associada ao poligene e e_i é a variável aleatória definindo o erro. As hipóteses sobre os efeitos aleatórios g_i e e_i e a covariância entre as variáveis fenotípicas para os indivíduos i e i' são dadas como em (3.25) e (3.26). No **Estágio 1** do modelo em multiestágios é feito o ajuste do modelo poligênico (3.30) para se obter os resíduos condicionais ($\hat{e}_i$) que são dados por:

$$\hat{e}_i = y - (\hat{\mu} + \mathbf{X}_{ij}\hat{\beta}_j + \hat{g}_i) \quad (3.31)$$

em que $\hat{\beta}_j$ é o estimador (BLUE) do j -ésimo efeito fixo e $\hat{g}_i$ é o preditor (BLUP) do efeito aleatório do poligene para o indivíduo i . Os resíduos obtidos em (3.31) são usados no **Estágio 2** como uma variável dependente em um modelo de regressão clássico para cada SNP_j , tal que,

$$\hat{e}_i = \mu + \beta_{\text{SNP}_j} X_{\text{SNP}_{ij}} + u_i, \quad (3.32)$$

em que $\hat{e}_i$ é o resíduo condicional do modelo (3.30), μ é a média geral, β_{SNP_j} é o correspondente efeito do SNP_j , $X_{\text{SNP}_{ij}}$ representa o genótipo do marcador SNP_j para o indivíduo i , e assume os valores 0, 1 ou 2 para os genótipos aa , Aa e AA , respectivamente, e u_i é o termo de erro. Note que o efeito do SNP não é avaliado no modelo inicial (3.30), o qual é ajustado no intuito de eliminar efeitos de covariáveis e a estrutura de dependência familiar. Neste caso o SNP está sendo modelado para avaliar seu efeito no componente residual ($\hat{e}_i$) com g_i assumido fixo. Testes simultâneos de $H_0: \beta_{\text{SNP}_j} = 0$ ($j=1, 2, \dots, M$) com nível de significância, em geral, corrigido para os múltiplos testes (Storey and Tibshirani, 2001) são realizados para selecionar SNPs do conjunto total com evidência de algum efeito sobre a variável Y . Deste modo, o Estágio 2 tem a finalidade de filtrar SNPs do conjunto total sob análise (sendo M , em geral, da ordem de milhares). Subsequentemente, no **Estágio 3** os SNPs selecionados no Estágio 2 são avaliados em uma análise final por meio do seguinte modelo poligênico,

$$Y_i = \mu + X_{ik}\beta_k + X_{\text{SNP}_{ik}^*}\beta_{\text{SNP}_k^*} + g_i + e_i, \quad i = 1, \dots, n; k = 1, \dots, M^*. \quad (3.33)$$

em que SNP^* refere-se aos SNPs selecionados no Estágio 2 e $M^* \leq M$ refere-se a um subconjunto do total de M SNPs avaliados nos estágios anteriores.

Aulchenko et al. (2007) fazem as aplicações utilizando dados simulados em que o número de SNPs no estudo não ultrapassa 100000, o qual é um número relativamente pequeno quando comparado com as plataformas que incluem aproximadamente 1 milhão de SNPs a serem analisados. Como vantagem do modelo em multiestágios tem-se a redução no tempo computacional que seria exigido se fossem ajustados diretamente M modelos poligênicos do tipo (3.33).

Uma outra proposta recente (Amin et al., 2007) segue as mesmas premissas de Aulchenko et al. (2007) com a diferença que os elementos da matriz de relacionamento são obtidos por meio da informação dos SNPs e não da declaração da genealogia familiar. Estimar o grau de relacionamento entre indivíduos via dados de SNP (mapas densos) tem sido uma alternativa bastante recomendada na literatura que pode evitar possíveis erros de declaração (Thompson, 1969).

Capítulo 4

Simulação de dados Genéticos e Propostas de Exercícios

Como salientado anteriormente, os dados de famílias têm sido comumente analisados via modelos mistos e usando marcadores microsatélites que são modelados como efeitos aleatórios. Os SNPs, por sua vez, têm sido vinculados a estudos observacionais caso-controle com indivíduos não relacionados tendo seu efeito modelado como um fator fixo. Para dados de famílias e SNPs surge o problema, não trivial, de como modelar os SNPs no mapeamento de genes, questionando-se se o mesmo deve ser modelado como efeito fixo ou aleatório. Com o intuito de contribuir para esta discussão este trabalho apresenta exercícios de simulação e análise de dados genéticos que facilitam o ensino e o entendimento de conceitos de Genética e de mapeamento de genes modelados a partir de efeitos fixos ou aleatórios. Deste modo, optou-se pela alternativa didática, que foi gerar computacionalmente dados de genótipos de SNPs e variáveis fenotípicas para dados de famílias utilizando, em particular, os recursos do *software R*. Para realizar tal procedimento, conceitos abordados nos capítulos 2 e 3 foram utilizados tanto para simular os dados de famílias e SNPs quanto para criar diferentes cenários para aplicação de diferentes procedimentos de análise dos dados, inclusive o modelo proposto por Aulchenko et al. (2007). A partir da elaboração deste procedimento um material foi produzido e disponibilizado para diferentes finalidades acadêmicas o qual pode ser usado tanto para estudantes da área de Estatística, que desejam entender alguns conceitos em Genética e ter um contato inicial com o *software R* para simulação e modelagem de dados genéticos,

quanto para estudantes de Genética interessados em entender a modelagem estatística de dados genéticos.

4.1 Material para fins acadêmicos

Nesta seção são propostos alguns exercícios de simulação, os quais possibilitam visualizar na prática como alguns conceitos de Estatística e Genética se aplicam, e como podem ser usados para diferentes finalidades acadêmicas. Ao adotar estes exercícios torna-se possível o desenvolvimento dos códigos de programação necessários para a solução de cada exercício por meio de comandos simples ou pode-se utilizar a programação já elaborada, na qual basta executar as funções para solucionar os exercícios, sem ter contato com os códigos de programação. Para todos os exercícios foram desenvolvidos os comandos e funções para utilização no *software R*, sob os quais é necessário apenas mudar os parâmetros para se obter diferentes cenários (ver Apêndice B).

Exercício 1. Gerar aleatoriamente $2n$ sequências independentes de DNA (A, C, T, G) de tamanho m , com, por exemplo, $P(A) = P(T) = 0,4$ e $P(C) = P(G) = 0,1$ (Figura 4.1).

...CCTGGAATTC...
...TGAATCTGGA...

Figura 4.1- Modelo de sequências geradas

Motivação: Entendimento do genoma humano (2 sequências com cerca de 3 bilhões de bases) e visualização gráfica do comportamento da distribuição do número de alelos “A” de acordo com $P(A)$ e o tamanho de n .

Atividade lúdica: Amostragem (com reposição) de bolas com 4 cores diferentes de uma urna, onde cada cor representa uma letra (base) da sequência de DNA. Qual é uma possível composição da urna, para que as probabilidades das bases de DNA sejam as mesmas como propostas no Exercício 1 (Figura 4.2)?

Figura 4.2- Representação de uma possível composição de urna

Formalização Teórica: Simulação de variáveis aleatórias com distribuição de probabilidades Uniforme (Figura 4.3).

$Y \sim U[0,1]$; se $y \in [0; 0,4) \Rightarrow$ ocorre A

se $y \in [0,4; 0,5) \Rightarrow$ ocorre C

se $y \in [0,5; 0,9) \Rightarrow$ ocorre T

se $y \in [0,9; 1] \Rightarrow$ ocorre G

Figura 4.3- Representação da distribuição das variáveis aleatórias

Gráficos: Considerando $m = 1$, criar gráficos da distribuição do número de alelos “A” para 50 e 1000 indivíduos. Repetir este processo $k = 50$ vezes. O que é observado em cada gráfico (Figura 4.4)?

Os comandos em *R* para executar este exercício estão no Apêndice B.

Neste exercício é possível verificar como cada parâmetro influencia na distribuição do número de alelos e criar diferentes cenários mudando $P(A)$, $P(T)$, $P(C)$ ou $P(G)$, desde que a soma das probabilidades seja 1, é válido ressaltar que nem todos os cenários respeitarão o equilíbrio de Hardy-Weinberg (descrito na seção 2.4.1) considerando, por exemplo, a ocorrência de “A” e “não A” em alguma posição da sequência gerada, mas serão úteis para entendimento do exercício. Pode-se também mudar os parâmetros para a construção dos gráficos por meio da mesma função desenvolvida no aplicativo *R* (Apêndice B) e verificar a distribuição do número de alelos “G”, ou pode-se ainda criar gráficos com diferentes valores para $2n$ ou k , onde k representa o número de execuções das simulações (Figura 4.5).

Figura 4.4- Distribuição do número de “alelos” A, $P(A) = 0,4$

Figura 4.5- Distribuição do número de “alelos” G, $P(G)=0,1$

Exercício 2. Considerando o cenário proposto no Exercício 1, faça o pareamento e a codificação das sequências, onde as letras A e T recebem código 1 e as letras C e G recebem código 0, segundo o modelo da Figura 4.6. Deste modo obtenha a matriz de genótipos codificados $G_{n \times m} = \{0, 1, 2\}$, em que n é o número de indivíduos e m é o tamanho da sequência, ou seja, o número de marcadores tipo SNP.

Motivação: Entendimento do equilíbrio de Hardy-Weinberg (seção 2.4.1) na distribuição dos genótipos dos indivíduos e entendimento da codificação de um marcador do tipo SNP.

Formalização Teórica: Geração de uma matriz de genótipos (sob o equilíbrio de H-W), $G_{n \times m} = \{0, 1, 2\}$, em que n é o número de indivíduos e m é o tamanho da sequência tal que:

$P(0) = P(aa) = P(a)^2 = 0,04$; aa: ocorrência dos pares de bases GG, CC, GC, CG.

$P(1) = P(Aa) = 2P(A)P(a) = 0,32$; Aa: ocorrência dos pares de bases AG, AC, TG, TC, GA, CA, GT, CT.

$P(2) = P(AA) = P(A)^2 = 0,64$; AA: ocorrência dos pares de bases AA, AT, TA, TT.

Figura 4.6- Modelo para codificação das sequências

Gráficos: Criar gráficos da distribuição dos genótipos para $n = 50$ e $n = 1000$ indivíduos.

Repetir este processo $k = 50$ vezes. O que é observado em cada gráfico (Figura 4.7)? Os alelos estão em equilíbrio de H-W?

Figura 4.7- Gráfico de perfis com a distribuição de probabilidades de genótipos para indivíduos não relacionados

O Apêndice B apresenta os comandos em *R* para executar este exercício.

Neste exercício pode-se calcular (manualmente) as probabilidades de $P(AA)$, $P(Aa)$ e $P(aa)$ para diferentes cenários e analisar, graficamente, o comportamento dos genótipos para as probabilidades obtidas. Além disso, pode ser verificado se o equilíbrio de H-W está satisfeito. Uma construção gráfica interessante para representar estes dados é construir o simplex $[0, 1]^3$ com a dispersão dos pontos. Esta visualização permite discutir a estrutura composicional das probabilidades de ocorrência dos valores 0, 1 e 2 (Figura 4.8).

Figura 4.8- Gráfico Simplex $[0, 1]^3$

Exercício 3. Formação do núcleo familiar- gerar aleatoriamente quatro seqüências para os pais e obter, condicionalmente, as seqüências de descendentes. O seguinte heredograma/pedigree pode ser adotado:

Figura 4.9- Heredograma de um núcleo familiar (6 membros)

Motivação: Discutir as probabilidades genótípicas dos filhos para casamentos aleatórios e o Equilíbrio de Hardy-Weinberg:

$$P(0)^1 = P(aa)^1 = P[(aa)^1 \cap (aa \times aa)^0] + P[(aa)^0 \cap (aa \times Aa)^1] + P[(aa)^0 \cap (Aa \times Aa)^1] = 0,04$$

$P(1)^1 = P(Aa)^1 = 0,32$; $P(2)^1 = P(AA)^1 = 0,64$, em que o sobre-escrito indica a geração dos pais (= 0) e dos filhos (= 1).

Formalização Teórica: Geração de uma matriz de genótipos (para m locos) para dados de F famílias (*cluster*) $G_{6F \times m} = \{0, 1, 2\}$, onde em cada um dos locos dos genótipos dos pais foi selecionado um dos genes para serem doados ao genótipo dos filhos segundo $Y \sim U[0,1]$, se $y \in [0; 0,5)$ é selecionado o gene da primeira sequência, se $y \in [0,5; 1]$ é selecionado o gene da segunda sequência.

Gráficos: Criar gráficos da distribuição dos genótipos em famílias para 10 (60 indivíduos) e 1000 (6000 indivíduos) famílias. Repetir este processo 50 vezes. O que é observado em cada gráfico (Figura 4.10)?

O Apêndice B apresenta os comandos em *R* para a execução deste exercício.

$F = 10$ famílias, $n = 60$ indivíduos

$k = 50$ simulações

$F = 1000$ famílias, $n = 6000$ indivíduos

$k = 50$ simulações

Figura 4.10- Distribuição de probabilidades de genótipos para dados de famílias

Para este exercício pode-se calcular manualmente as probabilidades condicionais para os genótipos dos filhos em diferentes contextos. Além disso, é possível criar diferentes núcleos familiares estendendo as gerações verticalmente (netos e bisnetos) ou horizontalmente (primos e tios) e verificar o comportamento da distribuição dos genótipos na amostra total para cada cenário. É interessante notar que sob casamento aleatório e equilíbrio de H-W as probabilidades genotípicas (0, 1 ou 2) seguem o mesmo padrão tanto para $n = 60$ indivíduos independentes como para $n = 6000$ indivíduos dispostos em estrutura familiar. Contudo, no caso de famílias os gráficos da Figura 4.10 mostram um variabilidade maior em relação aos gráficos da Figura 4.7.

Exercício 4. Aplique o modelo de regressão linear para estimar efeitos genéticos (aditivo e de dominância) em indivíduos não relacionados.

Motivação: Entendimento do modelo genético aditivo ou linear (com efeitos genéticos fixos) (seções 2.1 e 2.5).

Formalização Teórica: Simulação de variáveis aleatórias seguindo a distribuição normal:

$$Y_{ij} \sim N(\mu_j + aX_{a_{ij}} + dX_{d_{ij}}, \sigma^2); i = 1, 2, \dots, n; j = aa, Aa, AA, \quad (4.1)$$

Efeito Aditivo

Efeito de dominância

em que μ_{AA} , μ_{Aa} e μ_{aa} estão definidos como na seção (2.5)

Gráficos: Resultados para $n = 60$ indivíduos, processo repetido $k = 50$ vezes, adote $\mu_j = 100 + 20X_j$, $X_j = 0, 1, 2$ com $a = 20$, $d = 0$, $\sigma^2 = 5$ (Figura 4.11).

No Apêndice B estão os comandos em *R* para executar este exercício.

Neste exercício é possível estimar o valor aditivo a via o ajuste de modelos de regressão clássicos, discutir diferentes gráficos de perfis de médias genotípicas para casos com efeitos de dominância ou mesmo sem qualquer efeito genético no padrão de variação das médias $a = d = 0$. É interessante criar diferentes cenários segundo as expressões fenotípicas de Dominância Total e Superdominância apresentadas na seção 2.1 e calcular o grau de dominância para cada contexto (seção 2.5).

Exercício 5. Utilize o modelo misto poligênico para simular os fenótipos de dados de família (com matriz 2ϕ conhecida) e verifique o comportamento dos dados para diferentes valores da herdabilidade.

Modelo Aditivo
 $n = 60$ indivíduos, $k = 50$ simulações

Valor Aditivo
 $n = 60$ indivíduos, $k = 50$ simulações

Figura 4.11- Resultados Modelo Aditivo

Motivação: Entendimento do modelo Poligênico (seção 3.3.1) (efeitos genéticos aleatórios) e aplicação do conceito de herdabilidade (seção 3.3.2).

Formalização Teórica: Simulação de variáveis aleatórias seguindo a distribuição normal multivariada com efeito genético no componente de variância.

$$\mathbf{Y}_{f(6 \times 1)} \sim N_6(\boldsymbol{\mu}; \boldsymbol{\Sigma}), \quad \boldsymbol{\Sigma}_{6 \times 6} = 2\underset{\substack{\uparrow \\ \text{Matriz de relacionamento}}}{\boldsymbol{\Phi}}\sigma_g^2 + \mathbf{I}\sigma_e^2 \quad (4.2)$$

σ_g^2 : componente de variância poligênico

σ_e^2 : componente de variância residual

Gráficos: Verifique a relação entre o valor da herdabilidade e o comportamento dos dados fenotípicos (Figura 4.12) considerando $k = 50$ execuções e $F = 100$ famílias todas com estrutura da Figura 4.6.

O Apêndice B apresenta os comandos em *R* para a execução deste exercício. Note que as herdabilidades, sob ajuste do modelo misto (4.2), foram estimadas usando os resultados da seção (3.3.2). A Figura 4.11 mostra estas estimativas. Além disso, é apresentado um gráfico de dispersão dos valores de Y para uma simulação de 10 famílias (60 indivíduos). É interessante visualizar a correlação intra classe, sob a qual todos os indivíduos de uma mesma família tendem a ter valores de Y na mesma direção (acima ou abaixo da média). Isto é evidente para herdabilidades altas e para $h_g^2 = 0,01$ tal padrão desaparece.

Neste exercício pode-se discutir a utilidade do modelo misto para modelar cenários mais gerais em que o efeito genético influencia tanto o componente da média da variável fenotípica Y como o componente de variância.

Os exercícios propostos abordam diversos conteúdos da área de Estatística, entre eles, probabilidade condicional e independência, distribuição uniforme, distribuição normal, distribuição normal multivariada, modelos de regressão clássicos, modelos mistos, estimação de parâmetros, construção de gráficos, a importância do tamanho da amostra em uma coleta de dados, componentes da variância, correlação intra classe, entre outros. Também contemplam assuntos da área de Genética que foram apresentados neste trabalho e destacados no decorrer dos exercícios. Por isso, podem ser usados de maneira complementar à matéria teórica dada em sala de aula, pois é possível visualizar na prática

alguns assuntos estatísticos e genéticos, além de permitir um contato com a linguagem de programação no *software R*.

$$h_g^2 = 0.80$$

Estimativa da herdabilidade

Correlação intraclasse

$$h_g^2 = 0.01$$

Estimativa da herdabilidade

Correlação intraclasse

Figura 4.12- Associação entre o valor da herdabilidade e a correlação intra classe

Capítulo 5

Aplicação – Análise em Multiestágios

Neste capítulo é apresentado diferentes cenários de simulação de dados de famílias e SNPs utilizados para aplicação do modelo proposto por Aulchenko et al. (2007). Os fenótipos dos indivíduos foram gerados seguindo distribuição normal multivariada com diferentes estruturas para os parâmetros do vetor de média e da matriz de covariâncias tendo como objetivo estimar o efeito do SNP sob diferentes contextos de simulação. Os genótipos de cada indivíduo não relacionado foi gerado respeitando o equilíbrio de Hardy-Weinberg e a estrutura familiar foi imposta. Os comandos utilizados no *software R* para o desenvolvimento deste capítulo estão no Apêndice C.

5.1 Descrição da simulação dos dados de famílias

Para simular os dados de famílias, os genótipos (*AA*, *Aa* e *aa*) dos indivíduos não relacionados foram gerados segundo o equilíbrio de Hardy-Weinberg (descrito na seção 2.4.1). Em todos os cenários abordados neste trabalho considera-se a probabilidade genotípica como sendo $p = P(A) = 0,8$, sendo *a* o suposto alelo de risco de um loco de SNP. Com isto tem-se os seguintes valores descritos na Tabela 5.1.

Como um particular cenário os núcleos familiares foram formados com 14 indivíduos segundo o heredograma da Figura 5.1.

Os genótipos dos indivíduos não relacionados foram criados respeitando o equilíbrio de Hardy-Weinberg. Para formar uma família os indivíduos 1, 2, 4 e 5 do

heredograma foram selecionados aleatoriamente formando os pais do primeiro e do segundo núcleo familiar, respectivamente. Como dito na seção 2.1, o ser humano é diploíde, por isso, em cada um dos locos dos genótipos dos pais foi selecionado um dos genes para serem doados ao genótipo dos filhos (filho 3 para o primeiro núcleo e filho 6 para o segundo núcleo) segundo $Y \sim U[0,1]$, se $y \in [0; 0,5)$ é selecionado o gene da primeira sequência, se $y \in [0,5; 1]$ é selecionado o gene da segunda sequência. Para formar os genótipos dos indivíduos 7 ao 14 este mesmo critério para doação de genes foi utilizado, neste caso, os pais são os indivíduos 3 e 6. No próximo passo, os genótipos AA , Aa e aa foram codificados em 2, 1 e 0, respectivamente, formando os SNPs para cada indivíduo. Em todos os cenários os dados genotípicos foram gerados para 150 famílias, formando um total de 2100 indivíduos.

Tabela 5.1 - Relação entre diferentes probabilidades dos alelos e genótipos

Probabilidade		Genótipos		
$P(A)$	$P(a)$	AA	Aa	aa
0,8	0,2	0,64	0,32	0,04

Os fenotípicos das famílias foram gerados seguindo uma distribuição normal multivariada com diferentes estruturas para os parâmetros do vetor de média e da matriz de covariâncias tendo como objetivo estimar o efeito do SNP sob diferentes cenários de simulação.

Após a criação dos genótipos (considerando $m = 1$ loco) e fenótipos das famílias, os dados foram agrupados em uma matriz $B_{n \times 5}$, onde n representa o número de indivíduos, $n = 1, \dots, 2100$, com as seguintes identificações para as colunas: FAMID, ID, FA, MO, Y, SNP. Em FAMID é colocada a identificação da família, $f = 1, \dots, 150$, em ID é colocado a

identificação dos indivíduos, em FA é colocada a identificação do pai do indivíduo, em MO a identificação da mãe, em Y os dados fenotípicos e em SNP os dados genotípicos já codificados em 0, 1 e 2. Após a elaboração da matriz, os dados estão aptos para serem submetidos ao modelo em multiestágios proposto por Aulchenko et al. (2007).

Figura 5.1- Heredograma familiar

5.2 Aplicação do processo em multiestágios

Em nossa programação, o primeiro passo para aplicar o modelo em multiestágios no *software R* é disponibilizar o pacote *Kinship*, com ele é possível construir a matriz de relacionamento entre os indivíduos (a partir dos dados FAMID, ID, FA e MO), necessária para realizar o ajuste do modelo poligênico e obter os resíduos condicionais no Estágio 1 (os Estágios foram descritos na seção 3.3.4). No Estágio 2 os resíduos condicionais são utilizados como uma variável dependente em um modelo de regressão simples para cada SNP. Os SNPs que forem significantes (ao nível de significância 5 %) são selecionados

para o Estágio 3 e avaliados em uma análise final, neste caso também os que forem significantes ao nível de significância 5 % são considerados com influência sobre a variável Y .

Para verificar a funcionalidade do modelo proposto por Aulchenko et al. (2007) em todos os cenários, os dados de famílias e o SNP são fixos e, em cada simulação, apenas os dados fenotípicos da variável Y são gerados de acordo com os cenários descritos a seguir.

Cenário 1- $Y \sim N_{14}(\mu, 2\phi\sigma_g^2 + I\sigma_e^2)$; $\mu = 120 + \beta X_{SNP}$; $\beta = 6$, $X_{SNP} = 0, 1$, ou 2 , de acordo com os genótipos aa , Aa ou AA , respectivamente; σ_g^2 e σ_e^2 apresentam diferentes valores, apresentados na Tabela 5.2, com o intuito de analisar os dados com valores diferentes para herdabilidade, 2ϕ é a matriz de relacionamento conhecida e I a matriz identidade. Em todos os casos $\sigma_y^2 = 20$, onde $\sigma_y^2 = \sigma_g^2 + \sigma_e^2$ e $h_g^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2}$ é a herdabilidade. Para todos os casos apresentados a seguir os dados foram simulados 500 vezes.

Tabela 5.2- Resultados referentes ao cenário 1

Porcentagem de SNPs significantes	$h_g^2 = 0,1$	$h_g^2 = 0,35$	$h_g^2 = 0,75$
	$\sigma_e^2 = 18, \sigma_g^2 = 2$	$\sigma_e^2 = 13, \sigma_g^2 = 7$	$\sigma_e^2 = 5, \sigma_g^2 = 15$
Estágio 2	96,2%	93,6%	91,6%
Estágio 3	93%	91,4%	88,8%

Os valores da Tabela 5.2 mostram os valores percentuais em relação ao total de simulações. Indicam que quanto maior o valor de σ_e^2 maior é a porcentagem de SNPs significantes nos Estágios 2 e 3. Além disso, o modelo proposto por Aulchenko et al.

(2007) mostrou-se eficiente para este cenário, pois conseguiu captar uma porcentagem considerável de SNPs significantes no Estágio 3.

Cenário 2- $Y \sim N_{14}(\mu, 2\phi\sigma_g^2 + \pi\sigma_q^2 + I\sigma_e^2)$, onde $\mu = 120$, σ_g^2 , σ_q^2 e σ_e^2 apresentam diferentes valores, apresentados na Tabela 5.3, com o intuito de analisar os dados com valores diferentes para herdabilidade e π é a proporção de alelos compartilhados entre os indivíduos no loco de QTL. Em todos os casos $\sigma_y^2 = 20$, onde $\sigma_y^2 = \sigma_g^2 + \sigma_q^2 + \sigma_e^2$ e $h_g^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_q^2 + \sigma_e^2}$ e $h_q^2 = \frac{\sigma_q^2}{\sigma_g^2 + \sigma_q^2 + \sigma_e^2}$. Para todos os casos apresentados a seguir os dados foram simulados 500 vezes.

Tabela 5.3-Resultados referentes ao cenário 2

Porcentagem de SNPs significantes	$h_g^2 = 0,05$ $h_q^2 = 0,05$	$h_g^2 = 0,05$ $h_q^2 = 0,3$	$h_g^2 = 0,05$ $h_q^2 = 0,7$
	$\sigma_e^2 = 18, \sigma_q^2 = 1, \sigma_g^2 = 1$	$\sigma_e^2 = 13, \sigma_q^2 = 6, \sigma_g^2 = 1$	$\sigma_e^2 = 5, \sigma_q^2 = 14, \sigma_g^2 = 1$
Estágio 2	60,2%	51,2%	42,6%
Estágio 3	55,2%	44%	37,2%

Os dados da Tabela 5.3 mostram que conforme os valores de σ_e^2 diminuem a porcentagem de SNPs significantes no Estágio 2 diminuem consideravelmente, mostrando que para este cenário o modelo proposto por Aulchenko et al. (2007) não é muito eficiente. Além disso, comparando com o Cenário 1, o comportamento do procedimento multiestágio foi muito ruim indicando que quando o SNP traz informação sobre a correlação familiar o resíduo condicional não é uma boa estatística a ser adotada, sozinha, na seleção de SNPs.

Cenário 3- Foi gerado $X_{SNP} = 0, 1, \text{ ou } 2$, de acordo com os genótipos aa, Aa ou AA , respectivamente, de forma aleatória para todos os 2100 indivíduos, sem respeitar a estrutura familiar, isto é, a progenia e os descendentes. Foram mantidas as mesmas probabilidades de ocorrência das classes genóticas descritas na Tabela 5.1. A estrutura é a mesma adotada no Cenário 1 (Tabela 5.4).

Tabela 5.4- Resultados referentes ao cenário 3

Porcentagem de SNPs significantes	$h_g^2 = 0,1$	$h_g^2 = 0,35$	$h_g^2 = 0,75$
	$\sigma_e^2 = 18, \sigma_g^2 = 2$	$\sigma_e^2 = 13, \sigma_g^2 = 7$	$\sigma_e^2 = 5, \sigma_g^2 = 15$
Estágio 2	98,2%	95,4%	92,8%
Estágio 3	96%	92,6%	90,6%

Os valores da Tabela 5.4 indicam um aumento nos valores percentuais em relação ao Cenário 1 mostrando que para este cenário o modelo proposto por Aulchenko et al. (2007) é mais eficiente no sentido de selecionar SNPs. Este resultado já era esperado pois o resíduo condicional não leva em conta a estrutura de dependência familiar.

Capítulo 6

Considerações Finais

Este trabalho teve dois focos principais: elaborar um material de natureza acadêmica destinado ao ensino de conceitos de Genética e Genômica Quantitativa, tanto a estudantes da área da Biologia Molecular como de Estatística Aplicada, e discutir o procedimento de mapeamento de genes proposto por Aulchenko et al. (2007) em delineamentos com famílias e plataformas de SNPs. Nos dois casos, programas computacionais foram desenvolvidos usando os recursos do *software R* e estão disponíveis aos interessados.

O material didático tem sido aplicado em algumas oportunidades de ensino, como em Cursos de Verão e Oficinas, e tem se mostrado útil como facilitador do ensino dos conceitos e metodologias a que se propôs. Esperamos, contudo, que este primeiro modelo possa ser estendido e aperfeiçoado para melhor atender aos anseios da educação estatística.

Relativamente ao outro foco deste trabalho, o procedimento em multiestágios de Aulchenko et al. (2007) foi apresentado e suas limitações e utilidade foram avaliadas. O procedimento mostrou-se uma boa ferramenta quando o intuito é reduzir o tempo computacional e realizar um primeiro filtro nos SNPs. Contudo, os resultados das simulações apresentadas mostraram situações em que informação é perdida no sentido de identificar SNPs. No procedimento, os ajustes no Estágio 2 consideram apenas resíduos referentes aos erros condicionais quando deveria considerar também o efeito do SNP nos efeitos aleatórios preditos. Deste modo, uma melhoria que pode ser proposta no procedimento de Aulchenko et al. (2007) é no Estágio 1, em que sugere-se calcular

também o valor de $\hat{g}$ (preditor do efeito aleatório) para cada indivíduo. Assim o filtro dos SNPs realizado no Estágio 2 seria feito também usando este termo como variável resposta. Contudo, neste caso como é esperada dependência entre indivíduos relacionados um modelo linear generalizado deve ser adotado. Os SNPs que forem significantes nas análises com os dois componentes serão selecionados para a análise final.

A escolha do modelo proposto por Aulchenko et al. (2007) para estudo se deu principalmente por ser uma das primeiras metodologias publicada sobre mapeamento de genes em dados de famílias e usando informação de marcadores moleculares do tipo SNP. A motivação sendo decorrente da disponibilidade de termos dados deste tipo para análise, envolvendo a população brasileira (Projeto Corações de Baependi, MG, gerenciado pelo Laboratório de Cardiologia e Genética Molecular do InCor). Como uma primeira análise exploratória destes dados pode-se usar o programa que desenvolvemos para executar as análises. Além disso, a modelagem usada por Aulchenko et al. (2007) considera o SNP como um fator fixo e sem informação sobre a estrutura de correlação familiar, caracterização esta que não é bem entendida por profissionais e acadêmicos que atuam na área. Neste sentido, os resultados das simulações mostraram que o procedimento é limitado pois pode ser influenciado pela magnitude do componente poligênico. Conforme a herdabilidade da variável fenotípica aumenta a proporção de SNPs selecionados por este método pode ser pequena, mantendo-se o mesmo valor do efeito fixo (β) do SNP e mesmo tamanho amostral. Ainda, caso o SNP realmente tenha um componente de correlação familiar, a eficiência deste procedimento de busca fica mais prejudicado. Uma alternativa a estas dificuldades é a implementação de uma melhoria no procedimento de Aulchenko et al.

(2007) por adicionar a análise do efeito do SNP também no componente aleatório poligênico.

Apêndice A

Expressões do Capítulo 3

A.1- Obtenção do BLUE e BLUP a partir da equação do modelo misto (3.2)

Considere o modelo

$$Y = X\beta + Z\gamma + \varepsilon \quad (3.2)$$

$$0_{cq \times 1} = 0_{cq \times 1} - I_{cq}\gamma + \eta, \quad (A.1)$$

simultaneamente com $\text{Var}(\varepsilon) = \sigma^2 \mathbf{R}$, $\text{Var}(\eta) = \sigma^2 \mathbf{D}$ e $\text{Cov}(\varepsilon, \eta^T) = 0_{n \times cq}$. Observe que η é um vetor ($cq \times 1$) atuando como “erro” na equação (A.1). Reescrevendo as duas equações (3.2) e (A.1) de forma matricial têm-se,

$$\begin{bmatrix} Y \\ 0 \end{bmatrix} = \begin{bmatrix} X & Z \\ 0 & -I \end{bmatrix} \begin{bmatrix} \beta \\ \gamma \end{bmatrix} + \begin{bmatrix} \varepsilon \\ \eta \end{bmatrix}. \quad (A.2)$$

Pré-multiplicando (A.2) por $\mathbf{R}^{-1/2} \oplus \mathbf{D}^{-1/2} = \begin{bmatrix} \mathbf{R}^{-1/2} & 0 \\ 0 & \mathbf{D}^{-1/2} \end{bmatrix}$, com $\oplus$

representando a soma direta (Searle et al., 1992), obtém-se:

$$\begin{bmatrix} \mathbf{R}^{-1/2} & 0 \\ 0 & \mathbf{D}^{-1/2} \end{bmatrix} \begin{bmatrix} Y \\ 0 \end{bmatrix} = \begin{bmatrix} \mathbf{R}^{-1/2} & 0 \\ 0 & \mathbf{D}^{-1/2} \end{bmatrix} \begin{bmatrix} X & Z \\ 0 & -I \end{bmatrix} \begin{bmatrix} \beta \\ \gamma \end{bmatrix} + \zeta$$

$$\begin{bmatrix} \mathbf{R}^{-1/2} Y \\ 0 \end{bmatrix} = \begin{bmatrix} \mathbf{R}^{-1/2} X & \mathbf{R}^{-1/2} Z \\ 0 & -\mathbf{D}^{-1/2} \end{bmatrix} \begin{bmatrix} \beta \\ \gamma \end{bmatrix} + \zeta$$

$$Y^* = X^* \beta^* + \zeta, \quad (A.3)$$

em que, $Y^* = \begin{bmatrix} \mathbf{R}^{-1/2}Y \\ 0 \end{bmatrix}$, $X^* = \begin{bmatrix} \mathbf{R}^{-1/2}X & \mathbf{R}^{-1/2}Z \\ 0 & -\mathbf{D}^{-1/2} \end{bmatrix}$, $\beta^* = (\beta, \gamma)^T$, tal que $\text{Var}(\zeta) = \sigma^2 \mathbf{I}_{cq+n}$.

Desta forma, (A.3) pode ser considerado como um modelo linear “homocedástico”.

Portanto, o BLUE para β e o BLUP para γ podem ser obtidos por meio da equação (A.3), usando o método de Mínimos Quadrados, sob o qual se obtém

$$(\mathbf{X}^*)^T Y^* = (\mathbf{X}^*)^T \mathbf{X}^* \hat{\beta}^*, \quad (\text{A.4})$$

ou seja,

$$\begin{bmatrix} X^T \mathbf{R}^{-1} Y \\ Z^T \mathbf{R}^{-1} Y \end{bmatrix} = \begin{bmatrix} X^T \mathbf{R}^{-1} X & X^T \mathbf{R}^{-1} Z \\ Z^T \mathbf{R}^{-1} X & Z^T \mathbf{R}^{-1} Z + \mathbf{D}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{\gamma} \end{bmatrix}. \quad (\text{A.5})$$

O BLUE e o BLUP são obtidos resolvendo-se as equações (A.5), que independem da distribuição de γ e ε . Por (A.5), tem-se:

$$X^T \mathbf{R}^{-1} X \hat{\beta} + X^T \mathbf{R}^{-1} Z \hat{\gamma} = X^T \mathbf{R}^{-1} Y \quad (\text{A.6})$$

$$Z^T \mathbf{R}^{-1} X \hat{\beta} + (Z^T \mathbf{R}^{-1} Z + \mathbf{D}^{-1}) \hat{\gamma} = Z^T \mathbf{R}^{-1} Y. \quad (\text{A.7})$$

Por (A.7),

$$\begin{aligned} (Z^T \mathbf{R}^{-1} Z + \mathbf{D}^{-1}) \hat{\gamma} &= Z^T \mathbf{R}^{-1} Y - Z^T \mathbf{R}^{-1} X \hat{\beta} \\ (Z^T \mathbf{R}^{-1} Z + \mathbf{D}^{-1}) \hat{\gamma} &= Z^T \mathbf{R}^{-1} (Y - X \hat{\beta}) \\ \hat{\gamma} &= (Z^T \mathbf{R}^{-1} Z + \mathbf{D}^{-1})^{-1} Z^T \mathbf{R}^{-1} (Y - X \hat{\beta}) \end{aligned} \quad (\text{A.8})$$

Substituindo (A.8) em (A.6),

$$X^T \mathbf{R}^{-1} X \hat{\beta} + X^T \mathbf{R}^{-1} Z (Z^T \mathbf{R}^{-1} Z + \mathbf{D}^{-1})^{-1} Z^T \mathbf{R}^{-1} (Y - X \hat{\beta}) = X^T \mathbf{R}^{-1} Y$$

e considerando (3.8) e (3.9)

$$\begin{aligned}
X^T R^{-1} X \hat{\beta} + X^T R^{-1} Z C^{-1} Z^T R^{-1} Y - X^T R^{-1} Z C^{-1} Z^T R^{-1} X \hat{\beta} &= X^T R^{-1} Y \\
X^T R^{-1} X \hat{\beta} - X^T R^{-1} Z C^{-1} Z^T R^{-1} X \hat{\beta} &= X^T R^{-1} Y - X^T R^{-1} Z C^{-1} Z^T R^{-1} Y \\
X^T [R^{-1} - R^{-1} Z C^{-1} Z^T R^{-1}] X \hat{\beta} &= X^T [R^{-1} - R^{-1} Z C^{-1} Z^T R^{-1}] Y \\
X^T M X \hat{\beta} &= X^T M Y.
\end{aligned} \tag{A.9}$$

Por conseguinte o BLUE de β é dado por

$$\hat{\beta} = (X^T M X)^{-1} X^T M Y = (X^T V^{-1} X)^{-1} X^T V^{-1} Y, \tag{A.10}$$

pois X é posto coluna completo e M é positiva definida, implicando que $X^T M X$ é não singular.

Substituindo (3.8) em (A.8) obtém-se o BLUP de γ :

$$\hat{\gamma} = C^{-1} Z^T R^{-1} (Y - X \hat{\beta}). \tag{A.11}$$

Observa-se também que

$$Y - X \hat{\beta} = M^{-1} Q Y, \tag{A.12}$$

com

$$Q = M - M X (X^T M X)^{-1} X^T M. \tag{A.13}$$

A matriz Q é simétrica de ordem n , tal que $Q M^{-1} Q = Q$, $Q X = \mathbf{0}$ e o posto de Q é $n - p$. O BLUP e o BLUE satisfazem

$$\mathbf{X}^T \mathbf{R}^{-1} \mathbf{Y} = \mathbf{X}^T \mathbf{R}^{-1} \mathbf{X} \hat{\beta} + \mathbf{X}^T \mathbf{R}^{-1} \mathbf{Z} \hat{\gamma} \quad (\text{A.14})$$

desde que $\mathbf{R}^{-1} - \mathbf{M} = \mathbf{R}^{-1} \mathbf{Z} \mathbf{C}^{-1} \mathbf{Z}^T \mathbf{R}^{-1}$, $\mathbf{X}^T \mathbf{Q} = \mathbf{0}$, usando (A.12) e (3.8) tem-se,

$$\begin{aligned} \mathbf{X}^T \mathbf{R}^{-1} \mathbf{X} \hat{\beta} + \mathbf{X}^T \mathbf{R}^{-1} \mathbf{Z} \hat{\gamma} &= \mathbf{X}^T \mathbf{R}^{-1} \mathbf{X} \hat{\beta} + \mathbf{X}^T \mathbf{R}^{-1} \mathbf{Z} \mathbf{C}^{-1} \mathbf{Z}^T \mathbf{R}^{-1} (\mathbf{Y} - \mathbf{X} \hat{\beta}) \\ &= \mathbf{X}^T \mathbf{R}^{-1} (\mathbf{Y} - \mathbf{M}^{-1} \mathbf{Q} \mathbf{Y}) + \mathbf{X}^T \mathbf{R}^{-1} \mathbf{Z} \mathbf{C}^{-1} \mathbf{Z}^T \mathbf{R}^{-1} \mathbf{M}^{-1} \mathbf{Q} \mathbf{Y} \\ &= \mathbf{X}^T \mathbf{R}^{-1} \mathbf{Y} - \mathbf{X}^T \mathbf{R}^{-1} \mathbf{M}^{-1} \mathbf{Q} \mathbf{Y} + \mathbf{X}^T \mathbf{R}^{-1} \mathbf{Z} \mathbf{C}^{-1} \mathbf{Z}^T \mathbf{R}^{-1} \mathbf{M}^{-1} \mathbf{Q} \mathbf{Y} \\ &= \mathbf{X}^T \mathbf{R}^{-1} \mathbf{Y} - \mathbf{X}^T [\mathbf{R}^{-1} - \mathbf{R}^{-1} \mathbf{Z} \mathbf{C}^{-1} \mathbf{Z}^T \mathbf{R}^{-1}] \mathbf{M}^{-1} \mathbf{Q} \mathbf{Y} \\ &= \mathbf{X}^T \mathbf{R}^{-1} \mathbf{Y} - \mathbf{X}^T \mathbf{M} \mathbf{M}^{-1} \mathbf{Q} \mathbf{Y} \\ &= \mathbf{X}^T \mathbf{R}^{-1} \mathbf{Y} - \mathbf{X}^T \mathbf{Q} \mathbf{Y} \\ &= \mathbf{X}^T \mathbf{R}^{-1} \mathbf{Y} \end{aligned}$$

Uma outra identidade útil é

$$\mathbf{C}^{-1} \mathbf{Z}^T \mathbf{R}^{-1} = \mathbf{D} \mathbf{Z}^T \mathbf{M}, \quad (\text{A.15})$$

na qual usa-se (3.8) e (3.9),

$$\begin{aligned} \mathbf{D} \mathbf{Z}^T \mathbf{M} &= \mathbf{D} \mathbf{Z}^T [\mathbf{R}^{-1} - \mathbf{R}^{-1} \mathbf{Z} \mathbf{C}^{-1} \mathbf{Z}^T \mathbf{R}^{-1}] \\ &= \mathbf{D} \mathbf{Z}^T \mathbf{R}^{-1} - \mathbf{D} \mathbf{Z}^T \mathbf{R}^{-1} \mathbf{Z} \mathbf{C}^{-1} \mathbf{Z}^T \mathbf{R}^{-1} \\ &= \mathbf{D} \mathbf{C} \mathbf{C}^{-1} \mathbf{Z}^T \mathbf{R}^{-1} - \mathbf{D} \mathbf{Z}^T \mathbf{R}^{-1} \mathbf{Z} \mathbf{C}^{-1} \mathbf{Z}^T \mathbf{R}^{-1} \\ &= \mathbf{D} [\mathbf{C} - \mathbf{Z}^T \mathbf{R}^{-1} \mathbf{Z}] \mathbf{C}^{-1} \mathbf{Z}^T \mathbf{R}^{-1} \\ &= \mathbf{D} [\mathbf{D}^{-1} + \mathbf{Z}^T \mathbf{R}^{-1} \mathbf{Z} - \mathbf{Z}^T \mathbf{R}^{-1} \mathbf{Z}] \mathbf{C}^{-1} \mathbf{Z}^T \mathbf{R}^{-1} \\ &= \mathbf{D} \mathbf{D}^{-1} \mathbf{C}^{-1} \mathbf{Z}^T \mathbf{R}^{-1} \\ &= \mathbf{C}^{-1} \mathbf{Z}^T \mathbf{R}^{-1}. \end{aligned}$$

A identidade (A.15) fornece uma fórmula alternativa para o cálculo de $\hat{\gamma}$.

Considerando (A.11), tem-se

$$\begin{aligned}
\hat{\gamma} &= \mathbf{C}^{-1} \mathbf{Z}^T \mathbf{R}^{-1} (\mathbf{Y} - \mathbf{X} \hat{\beta}) \\
&= \mathbf{D} \mathbf{Z}^T \mathbf{M} (\mathbf{M}^{-1} \mathbf{Q} \mathbf{Y}) \\
\hat{\gamma} &= \mathbf{D} \mathbf{Z}^T \mathbf{Q} \mathbf{Y}.
\end{aligned} \tag{A.16}$$

Em suma, o BLUE e o BLUP são dados por (A.10) e (A.16)

$$\begin{aligned}
\hat{\beta} &= (\mathbf{X}^T \mathbf{M} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{M} \mathbf{Y}, \\
\hat{\gamma} &= \mathbf{C}^{-1} \mathbf{Z}^T \mathbf{R}^{-1} (\mathbf{Y} - \mathbf{X} \hat{\beta}) = \mathbf{D} \mathbf{Z}^T \mathbf{Q} \mathbf{Y}.
\end{aligned}$$

Apêndice B

Comandos dos exercícios propostos no capítulo 4

Exercício 1

```
GEQ<- function(m){ # m é o tamanho da sequencia
X = vector()
U = runif((m))
for (i in 1:m){
if (U[i] < 0.4)
X[i] = "A"
if ((U[i] >= 0.4) & (U[i] < 0.5))
X[i] = "C"
if ((U[i] >= 0.5) & (U[i] < 0.9))
X[i] = "T"
if (U[i] >= 0.9)
X[i] = "G" }
return(X)}
```

Função que retorna n sequencias em uma matriz com m colunas
É considerada como uma matriz de I indivíduos não relacionados

```
Matrix_Ind <- function ( I, m ){
n = I*2
S = matrix ( NA, n, m )
for ( j in 1:n ){
S[j,] <- GEQ ( m )}
return( S )}
```

```
Dist_Alelos<- function (I, m, y, k){ # Distribuição do Alelo A
Individuos <- Matrix_Ind(I,m)
n <- nrow(Individuos)
v <- vector ()
for (i in 1:k){
Individuos <- Matrix_Ind(I,m)
r <- Individuos[,y]
n <- nrow(Individuos)
freq_rel <- 0
cont <- 0
```

```

for(j in 1:n){
  if (r[j] == "A")
    cont = cont + 1 }
freq_rel = cont/n
v[i]<-freq_rel}
return (v)}

```

```

alelos<-Dist_Alelos(50,5,1,50) # 50 indivíduos, com 5 colunas,escolhi a coluna 1
                                # repeti o processo 50 vezes

```

```

# Gráfico da Distribuição do número de alelos "A"
plot(alelos, xlab="Índice",ylab ="Freq. Relativa",
main= "Distribuição do número de alelos A", cex.main=1, pch=1, ylim=c(0.2,0.6))
abline(0.4,0, lty=3, col=2)

```

```

Dist_Alelos_G<- function (I, m, y, k){  # Distribuição do Alelo G
Individuos <- Matrix_Ind(I,m)
n <- nrow(Individuos)
v <- vector ()
for (i in 1:k){
Individuos <- Matrix_Ind(I,m)
r <- Individuos[,y]
n <- nrow(Individuos)
freq_rel <- 0
cont <- 0
for(j in 1:n){
  if (r[j] == "G")
    cont = cont + 1 }
freq_rel = cont/n
v[i]<-freq_rel}
return (v)}

```

```

alelos<-Dist_Alelos_G(600,5,1,40) # 600 indivíduos, com 5 colunas,escolhi a coluna 1
                                # repeti o processo 40 vezes

```

```

# Gráfico da Distribuição do número de alelos "A"
plot(alelos, xlab="Índice",ylab ="Freq. Relativa",
main= "Distribuição do número de alelos G", cex.main=1, pch=1, ylim=c(0,0.4))
abline(0.1,0, lty=3, col=2)

```

Exercício 2

```

Dist_Alelos_Ind <- function (I, m, y, h){
v <- matrix (NA,h,3)
for (z in 1:h){
# Constroi a matriz com individuos não relacionados

```

```

Banco_Dados <-Matrix_Ind(I,m)
# Codificando a matriz dos individuos
p <- nrow(Banco_Dados)
q <- ncol(Banco_Dados)
Transf<- matrix ( NA, p , q )
for(j in 1:p){
  for(i in 1:q){
    if(Banco_Dados[j,i]=="A" || Banco_Dados[j,i]=="T")
      Transf [j,i]<- 1
    else
      Transf [j,i]<- 0} } # Transforma as letras em zero e um

Codifica <- matrix(NA, (p/2),q) # Soma de duas em duas linhas
for(k in 1:(p/2)){
  t1=(k*2)-1
  t2=k*2
  c<-vector
  a<-Transf[t1,]
  b<-Transf[t2,]
  c<-a+b
  Codifica[k,]<- c } # Retorna uma matriz com com zero, um e dois

r <- Codifica[,y]
t <- nrow(Codifica)

Cont_Cod <- function (y){ # y é o loco (coluna)
  Zero <- 0
  Um <- 0
  Dois <- 0
  for (i in 1:t){
    if (r[i] == 0)
      Zero = Zero + 1
    if (r[i] == 1)
      Um = Um + 1
    if (r[i] == 2)
      Dois = Dois + 1}
  Freq_zero = Zero/t
  Freq_um = Um/t
  Freq_dois = Dois/t
  Freq<-vector
  Freq<-c(Freq_zero, Freq_um, Freq_dois)
  Freq}

v[z,]<- Cont_Cod (Codifica[,y])}
return(v)}

U <-Dist_Alelos_Ind (100,5,1,50)  # A partir de 60 indivíduos com 5 colunas

```

```
# conte a freq de zeros, um e dois na coluna 1
# e repita este processo 50 vezes
```

```
# Gráfico de perfis do numero de genótipos na matrix de familias
perfil<-function(dados){
  n<-ncol(dados)
  m<-nrow(dados)
  for(i in 1:n){
    plot(rep((i-1),m),dados[,i],type='n',xlim=c(0,2),ylim=c(0,1),
      xlab="",ylab="",axes=FALSE)}
    axis(1,0:2)
    axis(2)
    #title(main="Número de genótipos de indivíduos não relacionados",
    cex.main=1,xlab="Alelos",ylab="Frequencia")
    for(i in 1:(n-1)){
      segments(rep((i-1),m),dados[,i],rep(i,m),dados[,i+1],col="black")}}
  perfil(U)
```

```
# Gráfico Simplex
z <- U[, 1]
x <- U[, 2]
y <- U[, 3]
  scatterplot3d(x, y, z, highlight.3d=TRUE, col.axis="blue",
    col.grid="lightblue", main="", pch=20,
    xlab = "P(1)", ylab = "P(2)", zlab = "P(0)")
```

```
##### Exercício 3 #####
```

```
Familia<- function(f, I, m){ # f é a quantidade de familias que serão formadas
  # a partir de uma matriz com i indivíduos e m colunas
```

```
Ind_ao_relacionados <- Matrix_Ind(I,m)
n <- nrow(Ind_ao_relacionados)
m <- ncol(Ind_ao_relacionados)
k <- (nrow ( Ind_ao_relacionados ))/2
```

```
Banco_dados<- matrix(NA, 12*f, m)
```

```
l=1
for(g in 1:f){
```

```
# Indica os genótipo do individuo em uma dada posicao
Posicao <- function (p){
  i<-(p*2)-1
  Ind = matrix ( NA, 2, m )
  for ( j in 1:2 ){
    Ind [ j, ] <- Ind_ao_relacionados [ i, ]
```



```

i <- (p*2)}
return(Ind)}

# Fazendo os casamentos entre os individuos nao relacionados
# Escolhe aleatoriamente dois numeros sem repeticao de uma sequencia
# de tamanho k (n/2), neste caso serão o pai e a mae

vec <- vector
b <- seq ( 1:k )
vec <- sample ( b, 2 )
#print(vec)
Pai <- matrix ( NA, 2 , m )
Mae <- matrix ( NA, 2 , m )
Pai <- Posicao(vec[1])
Mae <- Posicao(vec[2])

# Escolhendo quais gens serão doadas de cada loco
# Escolhe aleatoriamente o número 1 ou 2,
# e retorna a sequencia em um vetor

Letra <- function(m){
X = vector()
U = runif(m)
for (i in 1:m){
if (U[i] <= 0.5)
X[i] = 1
if (U[i] > 0.5)
X[i] = 2 }
return(X)}

# Doação feita pelo Pai
# Forma um vetor com a sequencia de gens escolhido de cada loco

Doacao_Pai<- function(m){
doacao_pai <- matrix (NA, 1, m)
d<-vector
d <- Letra(m)
for (i in 1:m){
doacao_pai [1, i]<- Pai[d[i],i]}
return (doacao_pai)}

#Doação feita pela Mae
# Forma um vetor com a sequencia de gens escolhido de cada loco

Doacao_Mae<- function(m){
doacao_mae <- matrix (NA, 1, m)
v <-vector

```

```

v <- Letra(m)
for (i in 1:m){
doacao_mae [1, i]<- Mae[v[i],i]}
return(doacao_mae)}

# Filha 1
# É formada pelas doações do pai na primeira linha e da mae na segunda linha

Filha_1 <- matrix (NA, 2, m)
Filha_1[1,] <- Doacao_Pai (m)
Filha_1[2,] <- Doacao_Mae (m)

# Filha 2
# É formada pelas doações do pai na primeira linha e da mae na segunda linha
# Mas com uma sequencia de gens diferentes da Filha 1

Filha_2 <- matrix (NA, 2, m)
Filha_2[1,] <- Doacao_Pai (m)
Filha_2[2,] <- Doacao_Mae (m)

# Filha 3
# É formada pelas doações do pai na primeira linha e da mae na segunda linha
# Mas com uma sequencia de gens diferentes da Filha 1 e da Filha 2

Filha_3 <- matrix (NA, 2, m)
Filha_3[1,] <- Doacao_Pai (m)
Filha_3[2,] <- Doacao_Mae (m)

# Filha 4
# É formada pelas doações do pai na primeira linha e da mae na segunda linha
# Mas com uma sequencia de gens diferentes da Filha 1, Filha 2 e da Filha 3

Filha_4 <- matrix (NA, 2, m)
Filha_4[1,] <- Doacao_Pai (m)
Filha_4[2,] <- Doacao_Mae (m)

# Forma uma matriz com os genótipos da familia,
# onde as duas primeiras linha são referentes ao pai, as próximas duas a mae,
# as próximas duas a Filha 1 e assim por diante

Banco_dados[l, ]<- Pai[1,]
Banco_dados[l+1, ]<- Pai[2,]
Banco_dados[l+2, ]<- Mae[1,]
Banco_dados[l+3, ]<- Mae[2,]
Banco_dados[l+4, ]<- Filha_1[1,]
Banco_dados[l+5, ]<- Filha_1[2,]
Banco_dados[l+6, ]<- Filha_2[1,]

```

```

Banco_dados[l+7, ]<- Filha_2[2,]
Banco_dados[l+8, ]<- Filha_3[1,]
Banco_dados[l+9, ]<- Filha_3[2,]
Banco_dados[l+10, ]<- Filha_4[1,]
Banco_dados[l+11, ]<- Filha_4[2,]
g = g+1
l = l+12}
Banco_dados}

Dist_Alelos_Fam <- function (I, m, f, y, h){
v <- matrix (NA,h,3)
for (z in 1:h){

# Constrói o banco de dados com as famílias
Banco_Dados <-Familia(f, I, m)

# Codificando a matriz das famílias
p <- nrow(Banco_Dados)
q <- ncol(Banco_Dados)
Transf<- matrix ( NA, p , q )
for(j in 1:p){
for(i in 1:q){
if(Banco_Dados[j,i]=="A" || Banco_Dados[j,i]=="T")
Transf [j,i]<- 1
else
Transf [j,i]<- 0} } # Transforma as letras em zero e um

Codifica <- matrix(NA, (p/2),q) # Soma de duas em duas linhas
for(k in 1:(p/2)){
t1=(k*2)-1
t2=k*2
c<-vector
a<-Transf[t1,]
b<-Transf[t2,]
c<-a+b
Codifica[k,]<- c} # Retorna uma matriz com com zero, um e dois
r <- Codifica[,y]
t <- nrow(Codifica)
Cont_Cod <- function (y){ # y é o loco (coluna)
Zero <- 0
Um <- 0
Dois <- 0
for (i in 1:t){
if (r[i] == 0)
Zero = Zero + 1
if (r[i] == 1)
Um = Um + 1

```

```

if (r[i] == 2)
Dois = Dois + 1}
Freq_zero = Zero/t
Freq_um = Um/t
Freq_dois = Dois/t
Freq<-vector
Freq<-c(Freq_zero, Freq_um, Freq_dois)
Freq}
v[z,]<- Cont_Cod (Codifica[,y])}
return(v)}

```

```

D <-Dist_Alelos_Fam (50,5,10,1,50)  # A partir de 150 indivíduos não relacionados com
5 colunas (futuros marcadores)

```

```

# forme 10 famílias, escolha a coluna 1
# e repita este processo 50 vezes

```

```

# Gráfico de perfis do numero de genotipos na matrix de familias
perfil<-function(dados){
n<-ncol(dados)
m<-nrow(dados)
for(i in 1:n){
plot(rep((i-1),m),dados[,i],type='n',xlim=c(0,2),ylim=c(0,1),
xlab="",ylab="",axes=FALSE)}
axis(1,0:2)
axis(2)
#title(main="Número de genotipos da matriz de dados de familia",
cex.main=1,xlab="Alelos",ylab="Frequencia")
for(i in 1:(n-1)){
segments(rep((i-1),m),dados[,i],rep(i,m),dados[,i+1],col="black")}}
perfil(D)

```

```

##### Exercício 4 #####

```

```

Modelo_Aditivo <- function (I, m, y){

```

```

Banco_Dados <-Matrix_Ind(I,m)
p <- nrow(Banco_Dados)
q <- ncol(Banco_Dados)
Transf<- matrix ( NA, p , q )
for(j in 1:p){
for(i in 1:q){
if(Banco_Dados[j,i]=="A" || Banco_Dados[j,i]=="T")
Transf [j,i]<- 1
else
Transf [j,i]<- 0} } # Transforma as letras em zero e um
Codifica <- matrix(NA, (p/2),q) # Soma de duas em duas linhas
for(k in 1:(p/2)){
t1=(k*2)-1

```

```

t2=k*2
c<-vector
a<-Transf[t1,]
b<-Transf[t2,]
c<-a+b
Codifica[k,]<- c } # Retorna uma matriz com com zero, um e dois
r <- Codifica[,y]
t <- nrow(Codifica)
v <- matrix (NA,t,1)

for(i in 1:t){
  if (r[i] == 0)
    v[i,]<- rnorm(1,100,5)
  if (r[i] == 1)
    v[i,]<- rnorm(1,120,5)
  if (r[i] == 2)
    v[i,]<- rnorm(1,140,5)}
w<-matrix(NA,t,2)
w[,1]<-r
w[,2]<-v
return(w)}

# Gráfico de perfis para o modelo Aditivo
Aditv <-function(I,m, y, k){
  perf<-matrix(NA,k,3)
  for (i in 1:k){
    Adit<-Modelo_Aditivo(I,m,y)
    Adit<-data.frame(Adit)
    perf[i,1]<-mean(Adit[,2][Adit[,1]=="0"])
    perf[i,2]<-mean(Adit[,2][Adit[,1]=="1"])
    perf[i,3]<-mean(Adit[,2][Adit[,1]=="2"])
  }
  return(perf)}

Dds<-Aditv(60,5,1,50)

perfil<-function(dados){
  n<-ncol(dados)
  m<-nrow(dados)
  for(i in 1:n){
    plot(rep((i-1),m),dados[,i],type='n',xlim=c(0,2),ylim=c(90,150),
    xlab="X",ylab="Y",axes=FALSE)}
    axis(1,0:2)
    axis(2)
    #title(main="Número de genotipos da matriz de dados de familia",
    cex.main=1,xlab="Alelos",ylab="Frequencia")
    for(i in 1:(n-1)){
      segments(rep((i-1),m),dados[,i],rep(i,m),dados[,i+1],col="black") } }

```

```
perfil(Dds)
```

```
#Gráfico dos efeitos aditivos para k simulações
```

```
Aditivo<-function(I, m, y, k){
```

```
Mat_Adit<-matrix(NA,k,1)
```

```
for (j in 1:k){
```

```
  a=0
```

```
  Adit<- Modelo_Aditivo(I,m,y)
```

```
  soma_zero=0
```

```
  soma_um=0
```

```
  soma_dois=0
```

```
  Zero=0
```

```
  Um=0
```

```
  Dois=0
```

```
  cont_zero=0
```

```
  cont_um =0
```

```
  cont_dois =0
```

```
  t<-nrow(Adit)
```

```
  # Calculando o valor aditivo (a)
```

```
  for(i in 1:t){
```

```
    if(Adit[i,1]==0){
```

```
      Zero<-Adit[i,2]
```

```
      cont_zero <-cont_zero +1
```

```
      soma_zero <-soma_zero + Zero }
```

```
    if(Adit[i,1]==1){
```

```
      Um<-Adit[i,2]
```

```
      cont_um <-cont_um +1
```

```
      soma_um <-soma_um + Um }
```

```
    if(Adit[i,1]==2){
```

```
      Dois<-Adit[i,2]
```

```
      cont_dois <-cont_dois +1
```

```
      soma_dois<-soma_dois + Dois } }
```

```
    a<-((soma_dois/cont_dois)-(soma_zero/cont_zero))/2
```

```
    Mat_Adit[j,]<-a }
```

```
  return (Mat_Adit)}
```

```
Aditivo_a<-Aditivo(60,5,1,50) # 60 individuos, 5 colunas,escolhi da coluna 1
```

```
      # repete esse processo 50 vezes
```

```
plot(Aditivo_a[,1], xlab=("Índice"), ylim=c(0,40), ylab="Valor Aditivo")
```

```
abline(20,0, lty=2, col=2)
```

```
##### Exercício 5 #####
```

```
# Dados de famílias - Modelo Poligenico
```

```
# Constroi a matriz phi
```

```
phi<-matrix(NA,6,6)
```

```

phi[1,]<-c(1,0,0.5,0.5,0.5,0.5)
phi[2,]<-c(0,1,0.5,0.5,0.5,0.5)
phi[3,]<-c(0.5,0.5,1,0.5,0.5,0.5)
phi[4,]<-c(0.5,0.5,0.5,1,0.5,0.5)
phi[5,]<-c(0.5,0.5,0.5,0.5,1,0.5)
phi[6,]<-c(0.5,0.5,0.5,0.5,0.5,1)

# Constroi a matriz identidade
ident<-matrix(NA,6,6)
ident[1,]<-c(1,0,0,0,0,0)
ident[2,]<-c(0,1,0,0,0,0)
ident[3,]<-c(0,0,1,0,0,0)
ident[4,]<-c(0,0,0,1,0,0)
ident[5,]<-c(0,0,0,0,1,0)
ident[6,]<-c(0,0,0,0,0,1)

# Constroi a matriz de variancia com herdabilidade 0.7
sigma_g<-0.7
sigma_e<-0.3
sigma_y = sigma_g + sigma_e
herdab<-sigma_g/sigma_y
G<-sigma_g*phi
E<-sigma_e*ident
VAR<-G+E

#Funcao que forma as familias
Poligen<-function(t,mu,Sigma){
  rnormmv <- function(n,mu,Sigma) {
    p <- nrow(Sigma)
    sig.svd <- svd(Sigma)
    Sigmaroot <- sig.svd$u%*%diag(sqrt(sig.svd$d))%*%t(sig.svd$v)
    z <- rnorm(p,0,1)
    x <- t(Sigmaroot%*%z+mu)
    if (n > 1){
      for (ii in 2:n){
        z <- rnorm(p,0,1)
        y <- Sigmaroot%*%z+mu
        x <- rbind(x,t(y))
      } # for ii in 2:n
    } # if $n > 1$
    return(x) # matriz n x p dos dados
  } # rnormmv

#Função que escolhe se os filhos serão homens ou mulheres
H_M<- function(m){ # m é o tamanho da sequencia
  X = vector()
  U = runif(m)

```

```

for (i in 1:m){
  if (U[i] <= 0.5)
    X[i] = 1
  if (U[i] > 0.5)
    X[i] = 2}
return(X)}

t1=1
t2=6
t3=1
t4=2
t5<-t4+4

FAM <- matrix(NA,nrow = (t*6), ncol= 6, dimnames= list (c(),
c("FAMID","ID","FA","MO","SEX","PRESS")))
for(i in 1:t){
#Identifica as famílias
FAM[t1:t2,1]<-rep(i,6)
# Identifica o pai
FAM[t3:t4,3]<-rep(0,2)
FAM[(t4+1):t5,3]<-rep(t3,4)
#Identifica a mae
FAM[t3:t4,4]<-rep(0,2)
FAM[(t4+1):t5,4]<-rep(t4,4)
#Identifica os sexos, onde o sexo do pai e da mae sao fixos e o dos filhos aleatorios
FAM[t3,5]<-1
FAM[t4,5]<-2
FAM[(t4+1):t5,5]<-H_M(4)
FAM[t1:t2,6]<-rmormmv(1,mu,Sigma)
t1<-i*6+1
t2<-t1+5
t3<-t3+6
t4<-t3+1
t5<-t4+4}
FAM[,2]<-seq(1,t*6)
return(FAM)}

# Gerando o banco de dados e construindo a matriz com os valores da herdabilidade
require(kinship)
Pol<-function(h,t,mu,Sigma){# h numero herdabilidade,t numero de familias no banco de
dados
Tran<-matrix(NA,nrow
=h,ncol=3,dimnames=list(c(),c("Sigma_g","Sigma_e","Herdabilidade")))
for(i in 1:h){
bae<-Poligen(t,mu,Sigma)
bae <- data.frame(bae)
attach(bae)

```



```

kmat<-makekinship(FAMID,ID,FA,MO)
fit1<-lmekin((log(PRESS))~1, data=bae, random=~1|ID, varlist=kmat)
Tran[i,1:2]<-fit1$theta
Tran[i,3]<-(Tran[i,1]/(Tran[i,1]+Tran[i,2]))}
return(Tran)}
Herd<-Pol(50,100,130,VAR) # Retorna 50 valores de herdabilidade,
                           # cada valor vindo de um banco de dados com 100 familias

# Gráfico da Herdabilidade
plot(Herd[,3], ylim=c(0,1), ylab= "Herdabilidade")
abline(0.7,0, lty=2, col=2)

#-----
# Construindo os gráficos das correlações intra classe
#-----
require(kinship)

Correlacao<-function(t,mu,Sigma){
bae<-Poligen(t,mu,Sigma)
bae <- data.frame(bae)
attach(bae)
Trans<-matrix(NA,t,4)
for(i in 1:t){
fam1<-subset(bae,FAMID==i)
Trans[i,1]<-max(fam1[,6])
Trans[i,2]<-min(fam1[,6])
Trans[i,3]<-mean(fam1[,6])}
Trans[,4]<-seq(1:t)
return(Trans)}

Corr<-Correlacao(10,130,VAR)
Corr<-data.frame(Corr)
attach(Corr)
plot(Corr[,1]~Corr[,4],pch=19, ylim=c(120,140),xlab="Family",ylab="Y")
seq<-Corr[,4]
min<-Corr[,2]
med<-Corr[,3]
points(seq,min,pch=1,col=12)
points(seq,med,pch=10,col=6)
legend("topright",c("Max","Mean","Min"),pch=c(19,10,1),col=c(1,6,12),inset=0.015,cex=.
8)
abline(130,0,lty=3,col=2)

```

Apêndice C

Comandos da Aplicação – Dados simulados - capítulo 5

```
require(kinship)
require(mvtnorm)

#Funcao que gera alelos com probabilidade p e m marcadores (colunas)
# m é o tamanho da sequencia

GEQ<- function(p,m){
  X = vector()
  U = runif(m)
  a<-p^2
  b<-(p^2)+2*p*(1-p)
  for (i in 1:m){
    if (U[i] < a)
      X[i] = "AA"
    if ( (U[i] >= a ) & ( U[i] < b ) )
      X[i] = "Aa"
    if (U[i] > b)
      X[i] = "aa"
  }
  return(X)
}

# É considerada como uma matriz de I indivíduos não relacionados
Matrix_Ind <- function ( I, p, m ){
  n = I
  S = matrix ( NA, n, m )
  for ( j in 1:n ){
    S[j,] <- GEQ ( p,m )
  }
  return( S )
}

#-----
# Forma f familias (Pai, Mae, F1)
# O pai e a mae são individuos escolhidos aleatoriamente da matriz de
# indivíduos não relacionados
# f é a quantidade de familias que serão formadas
# a partir de uma matriz com i indivíduos e m colunas
# com probabilidade p
```

```

Familia<- function(f, I, p, m){

Ind_nao_relacionados <- Matrix_Ind(I,p,m)
Banco_dados <- matrix(NA, 14*f, m)
l=1
# Escolhendo quais gens serão doadas de cada loco
# Escolhe aleatoriamente o número 1 ou 2,
# e retorna a sequencia em um vetor

Letra <- function(m){
  X = vector()
  U = runif(m)
  for (i in 1:m){
    if (U[i] <= 0.5)
      X[i] = 1
    if (U[i] > 0.5)
      X[i] = 2
  }
  return(X)
}

# Formação do primeiro núcleo familiar
# Fazendo os casamentos entre os individuos nao relacionados
# Escolhe aleatoriamente dois numeros sem repeticao de uma sequencia

for(g in 1:f){
  vec <- vector
  Pai <- vector
  Mae <- vector
  b <- seq ( 1:I )
  vec <- sample ( b, 2 )
  # print(vec)
  Pai <- Ind_nao_relacionados[vec[1],]
  Mae <- Ind_nao_relacionados[vec[2],]

  # Doação feita pelo Pai
  # Forma um vetor com a sequencia de gens escolhido de cada loco

Doacao_Pai<- function(m){
doacao_pai <- matrix (NA, 1, m)
d <-vector
d <- Letra(m)
for(j in 1:m){
  if(Pai[j]=="AA"){
    doacao_pai[1,j] <- 1
  }
}
}

```

```

if(Pai[j]=="aa"){
  doacao_pai[1,j] <- 0
}

```

```

if(Pai[j]=="Aa"){

  if (d[j]==1){
    doacao_pai[1,j]<-1
  }

  if (d[j]==2){
    doacao_pai[1,j]<-0
  }
}
}
return (doacao_pai)
}

```

Doação feita pela Mae

Forma um vetor com a sequencia de gens escolhido de cada loco

```

Doacao_Mae<- function(m){
doacao_mae <- matrix (NA, 1, m)
d <-vector
d <- Letra(m)
for(j in 1:m){
  if(Mae[j]=="AA"){
    doacao_mae[1,j] <- 1
  }
  if(Mae[j]=="aa"){
    doacao_mae[1,j] <- 0
  }
  if(Mae[j]=="Aa"){
    if (d[j]==1){
      doacao_mae[1,j]<-1
    }
    if (d[j]==2){
      doacao_mae[1,j]<-0
    }
  }
}
}
return (doacao_mae)}

```

Filho

É formada pelas doações do pai na primeira linha e da mae na segunda linha

```

F <- matrix (NA, 2, m)

```

```

F[1,] <- Doacao_Pai (m)
F[2,] <- Doacao_Mae (m)
Filho_11<-F[1,]+F[2,]
Filho <- matrix (NA, 1, m)
for(j in 1:m){          # Formação dos genótipos do Filho
  if(Filho_11[j] == 2){
    Filho[1,j] <- "AA"
  }
  if(Filho_11[j] == 0){
    Filho[1,j] <- "aa"
  }
  if(Filho_11[j] == 1){
    Filho[1,j] <- "Aa"
  }
}

# Formação do segundo núcleo familiar
vec <- vector
Pai_2 <- vector
Mae_2 <- vector
b <- seq ( 1:I )
vec <- sample ( b, 2 )
Pai_2 <- Ind_nao_relacionados[vec[1],]
Mae_2 <- Ind_nao_relacionados[vec[2],]

# Doação feita pelo Pai_2
# Forma um vetor com a sequencia de gens escolhido de cada loco
Doacao_Pai_2 <- function(m){
doacao_pai_2 <- matrix (NA, 1, m)
d <-vector
d <- Letra(m)
for(j in 1:m){
  if(Pai_2[j]=="AA"){
doacao_pai_2[1,j] <- 1}
  if(Pai_2[j]=="aa"){
doacao_pai_2[1,j] <- 0}
  if(Pai_2[j]=="Aa"){
    if (d[j]==1){
doacao_pai_2[1,j]<-1 }
    if (d[j]==2){
doacao_pai_2[1,j]<-0}
  }}
return (doacao_pai_2)}

# Doação feita pela Mae_2
# Forma um vetor com a sequencia de gens escolhido de cada loco

```

```

Doacao_Mae_2<- function(m){
doacao_mae_2 <- matrix (NA, 1, m)
d <-vector
d <- Letra(m)
for(j in 1:m){
if(Mae_2[j]=="AA"){
doacao_mae_2[1,j] <- 1 }
if(Mae_2[j]=="aa"){
doacao_mae_2[1,j] <- 0}
if(Mae_2[j]=="Aa"){
if (d[j]==1){
doacao_mae_2[1,j]<-1 }
if (d[j]==2){
doacao_mae_2[1,j]<-0}
}}
return (doacao_mae_2)}

# Filha
# É formada pelas doações do pai na primeira linha e da mae na segunda linha
Fa <- matrix (NA, 2, m)
Fa[1,] <- Doacao_Pai_2 (m)
Fa[2,] <- Doacao_Mae_2 (m)
Filha_12<-Fa[1,]+Fa[2,]
Filha <- matrix (NA, 1, m)
for(j in 1:m){ # Formação dos genótipos do Filho
if(Filha_12[j]==2){
Filha[1,j] <- "AA"}
if(Filha_12[j]==0){
Filha[1,j] <- "aa"}
if(Filha_12[j]==1){
Filha[1,j] <- "Aa"} }

# Fazendo os casamentos entre os Filhos
# Escolhe aleatoriamente dois numeros sem repeticao de uma sequencia
vec <- vector
Filho_Pai <- vector
Filha_Mae <- vector
Filho_Pai <- Filha
Filha_Mae <- Filha

# Doação feita pelo Filho_Pai
# Forma um vetor com a sequencia de gens escolhido de cada loco
Doacao_Filho_Pai<- function(m){
doacao_filho_pai <- matrix (NA, 1, m)
d <-vector
d <- Letra(m)

```

```

for(j in 1:m){
  if(Filho_Pai[j]=="AA"){
    doacao_filho_pai[1,j] <- 1 }
  if(Filho_Pai[j]=="aa"){
    doacao_filho_pai[1,j] <- 0 }
  if(Filho_Pai[j]=="Aa"){
    if (d[j]==1){
      doacao_filho_pai[1,j]<-1 }
    if (d[j]==2){
      doacao_filho_pai[1,j]<-0 }
    }}
  return (doacao_filho_pai)}

```

```

# Doação feita pela Mae
# Forma um vetor com a sequencia de gens escolhido de cada loco
Doacao_Filha_Mae<- function(m){
  doacao_filha_mae <- matrix (NA, 1, m)
  d <-vector
  d <- Letra(m)
  for(j in 1:m){
    if(Filha_Mae[j]=="AA"){
      doacao_filha_mae[1,j] <- 1 }
    if(Filha_Mae[j]=="aa"){
      doacao_filha_mae[1,j] <- 0 }
    if(Filha_Mae[j]=="Aa"){
      if (d[j]==1){
        doacao_filha_mae[1,j]<-1 }
      if (d[j]==2){
        doacao_filha_mae[1,j]<-0 }
      }}
    return (doacao_filha_mae)
  }

```

```

# Filhos
# É formada pelas doações do pai na primeira linha e da mae na segunda linha
Filhos<-matrix(NA,8,m)
for (i in 1:8){
  F <- matrix (NA, 2, m)
  F[1,] <- Doacao_Filho_Pai (m)
  F[2,] <- Doacao_Filha_Mae (m)
  Filho_13<-F[1,]+F[2,]
  Filho1 <- matrix (NA, 1, m)
  for(j in 1:m){ # Formação dos genótipos do Filho
    if(Filho_13[j]==2){
      Filho1[1,j] <- "AA"}
    if(Filho_13[j]==0){
      Filho1[1,j] <- "aa"}
  }
}

```

```

    if(Filho_13[j]==1){
      Filho1[1,j] <- "Aa"}
  }
  Filhos[i,]<-Filho1[1,]
}

```

```

Banco_dados[l, ]<- Pai
Banco_dados[l+1, ]<- Mae
Banco_dados[l+2, ]<- Filho
Banco_dados[l+3, ]<- Pai_2
Banco_dados[l+4, ]<- Mae_2
Banco_dados[l+5, ]<- Filha
Banco_dados[l+6, ]<- Filhos[1,]
Banco_dados[l+7, ]<- Filhos[2,]
Banco_dados[l+8, ]<- Filhos[3,]
Banco_dados[l+9, ]<- Filhos[4,]
Banco_dados[l+10, ]<- Filhos[5,]
Banco_dados[l+11, ]<- Filhos[6,]
Banco_dados[l+12, ]<- Filhos[7,]
Banco_dados[l+13, ]<- Filhos[8,]

```

```

g = g+1
l = l+14
}
Banco_dados
}

```

```

Dados<-NULL
Dados <-Familia (150,200,0.8,1)
# Familia<- function(f, l, p, m)
# f é a quantidade de famílias que serão formadas
# a partir de uma matriz com i indivíduos e m colunas
# com probabilidade p

```

```

# Transforma a matriz de letras para numeros zero, 1 e 2
n<-nrow(Dados)
m<-ncol(Dados)

```

```

Codifica <- matrix(NA, n ,m)
for(k in 1:n){
  for (i in 1:m){
    if(Dados[k,i]== "AA"){
      Codifica[k,i] <- 2
    }
    if(Dados[k,i]== "aa"){
      Codifica[k,i] <- 0
    }
  }
}

```



```

    if(Dados[k,i]=="Aa"){
      Codifica[k,i] <- 1
    }
  }
}

S1<-NULL
gt=500

for(arz in 1:gt)
{
#-----

a <-matrix(NA,2100, ncol= 1,dimnames=list(c(),c("Y1")))

#-----
# Construindo dados para Y1
v1=1
v2=14
for(u in 1:150){
# Constroi a matriz phi
x<-Codifica[v1:v2,1]
y<-Codifica[v1:v2,1]
X<-matrix(NA,14,14)
for(i in 1:13){
for(j in (i+1):14){
if(x[i]==y[j])
X[i,j]=X[j,i]<-1
if(abs(x[i]-y[j])==1)
X[i,j]=X[j,i]<-0.5
if((x[i]==0 && y[j]==2)||(x[i]==2 && y[j]==0))
X[i,j]=X[j,i]<-0 }
i=14
for(j in 2:14)
if(x[i]==y[j])
X[i,j]=X[j,i]<-1
if(abs(x[i]-y[j])==1)
X[i,j]=X[j,i]<-0.5
if((x[i]==0 && y[j]==2)||(x[i]==2 && y[j]==0))
X[i,j]=X[j,i]<-0
diag(X)<-1
# Constroi a matriz fhi
fhi<-matrix(NA,14,14)
fhi[1,]<-c(1,0,0.5,0,0,0,0.25,0.25,0.25,0.25,0.25,0.25,0.25,0.25)
fhi[2,]<-c(0,1,0.5,0,0,0,0.25,0.25,0.25,0.25,0.25,0.25,0.25,0.25)
fhi[3,]<-c(0.5,0.5,1,0,0,0,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5)

```

```

fhi[4,]<-c(0,0,0,1,0,0.5,0.25,0.25,0.25,0.25,0.25,0.25,0.25,0.25)
fhi[5,]<-c(0,0,0,0,1,0.5,0.25,0.25,0.25,0.25,0.25,0.25,0.25,0.25)
fhi[6,]<-c(0,0,0,0.5,0.5,1,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5)
fhi[7,]<-c(0.25,0.25,0.5,0.25,0.25,0.5,1,0.5,0.5,0.5,0.5,0.5,0.5,0.5)
fhi[8,]<-c(0.25,0.25,0.5,0.25,0.25,0.5,0.5,1,0.5,0.5,0.5,0.5,0.5,0.5)
fhi[9,]<-c(0.25,0.25,0.5,0.25,0.25,0.5,0.5,0.5,1,0.5,0.5,0.5,0.5,0.5)
fhi[10,]<-c(0.25,0.25,0.5,0.25,0.25,0.5,0.5,0.5,0.5,1,0.5,0.5,0.5,0.5)
fhi[11,]<-c(0.25,0.25,0.5,0.25,0.25,0.5,0.5,0.5,0.5,0.5,1,0.5,0.5,0.5)
fhi[12,]<-c(0.25,0.25,0.5,0.25,0.25,0.5,0.5,0.5,0.5,0.5,0.5,1,0.5,0.5)
fhi[13,]<-c(0.25,0.25,0.5,0.25,0.25,0.5,0.5,0.5,0.5,0.5,0.5,0.5,1,0.5)
fhi[14,]<-c(0.25,0.25,0.5,0.25,0.25,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,1)

```

```

# Constroi a matriz identidade

```

```

ident<-matrix(NA,14,14)
ident[1,]<-c(1,0,0,0,0,0,0,0,0,0,0,0,0,0)
ident[2,]<-c(0,1,0,0,0,0,0,0,0,0,0,0,0,0)
ident[3,]<-c(0,0,1,0,0,0,0,0,0,0,0,0,0,0)
ident[4,]<-c(0,0,0,1,0,0,0,0,0,0,0,0,0,0)
ident[5,]<-c(0,0,0,0,1,0,0,0,0,0,0,0,0,0)
ident[6,]<-c(0,0,0,0,0,1,0,0,0,0,0,0,0,0)
ident[7,]<-c(0,0,0,0,0,0,1,0,0,0,0,0,0,0)
ident[8,]<-c(0,0,0,0,0,0,0,1,0,0,0,0,0,0)
ident[9,]<-c(0,0,0,0,0,0,0,0,1,0,0,0,0,0)
ident[10,]<-c(0,0,0,0,0,0,0,0,0,1,0,0,0,0)
ident[11,]<-c(0,0,0,0,0,0,0,0,0,0,1,0,0,0)
ident[12,]<-c(0,0,0,0,0,0,0,0,0,0,0,1,0,0)
ident[13,]<-c(0,0,0,0,0,0,0,0,0,0,0,0,1,0)
ident[14,]<-c(0,0,0,0,0,0,0,0,0,0,0,0,0,1)

```

```

# Constroi a matriz de variancia

```

```

sigma_g <-1
sigma_e <-1
sigma_snp <-18
G<-(sigma_g*fhi)+(sigma_snp*X)
E<-sigma_e*ident
VAR<-G+E
a[v1:v2,1]<-rmvnorm(1, mean = rep(120,14), sigma = VAR)
v1=v2+1
v2=v1+13 }

```

```

#-----#-----#

```

```

#Funcao que forma as familias

```

```

Parte1<-function(t){
t1=1
t2=14
t3=1

```

```

t4=2
t5=4
t6=5
t7=3
t8=6
t9=7
t10=14

FAM <-matrix(NA,nrow = (t*14), ncol=4)
# ,dimnames=list(c(),c("FAMID","ID","FA","MO"))
for(i in 1:t){
#Identifica as famílias
FAM[t1:t2,1]<-rep(i,14)
FAM[t3:t4,3]<-rep(0,2) # Identifica os avos
FAM[t5:t6,3]<-rep(0,2) # Identifica os avos_2
FAM[t7,3]<-t1
FAM[t8,3]<-t5
FAM[t9:t10,3]<-rep(t7,8)
FAM[t3:t4,4]<-rep(0,2) # Identifica os avos
FAM[t5:t6,4]<-rep(0,2) # Identifica os avos_2
FAM[t7,4]<-t4
FAM[t8,4]<-t6
FAM[t9:t10,4]<-rep(t8,8)
t1<-i*14+1
t2<-t1+13
t3<-t3+14
t4<-t4+14
t5<-t5+14
t6<-t6+14
t7<-t7+14
t8<-t8+14
t9<-t9+14
t10<-t10+14}
FAM[,2]<-seq(1,t*14)
return(FAM)}
Fam<-Parte1(150)

#Unindo os dados em uma única tabela
Dados_new <-matrix(NA,2100, ncol=
6,dimnames=list(c(),c("FAMID","ID","FA","MO","Y1","SNP1"))))
Dados_new[,1]<-Fam[,1]
Dados_new[,2]<-Fam[,2]
Dados_new[,3]<-Fam[,3]
Dados_new[,4]<-Fam[,4]
Dados_new[,5]<-a[,1]
for (i in 1:1){
Dados_new[,i+5]<-Codifica[,i]

```

```

}
bae <- as.data.frame(Dados_new)
attach(bae)

#-----
# Implementação de Multiestágios
#-----

# ***** Estágio 1 *****

kmat<-makekinship(FAMID,ID,FA,MO)

fit1<-lmeKin(Y1~1, data=bae, random=~1|ID, varlist=kmatrix)
Tran<-matrix(NA,nrow
=1,ncol=3,dimnames=list(c(),c("Sigma_g","Sigma_e","Herdabilidade")))
i=1
Tran[i,1:2]<-fit1$theta
Tran[i,3]<-(Tran[i,1]/(Tran[i,1]+Tran[i,2]))
Delta<-as.matrix((2*kmatrix)*fit1$theta[1])
dim(Delta)
Sigma<-as.matrix(fit1$theta[2]*diag(2100))
dim(Sigma)
sig_y <-(fit1$theta[1])+(fit1$theta[2])
sig_y
var(fit1$residuals)
D<-as.matrix(Delta/sig_y)
dim(D)
R<-as.matrix(Sigma/sig_y)
dim(R)
Z<-diag(2100)
dim(Z)
Zt<-t(Z)
dim(Zt)
R_1<-solve(R)
dim(R_1)
V<-as.matrix(sig_y*((Z%*%D%*%Zt)+R))
dim(V)
V_1 <- solve(V)
M <-as.matrix(sig_y*V_1)
X<-matrix(rep(1,2100),ncol=1, byrow=TRUE)
dim(X)
Xt<-t(X)
Y<-Y1
beta <-as.matrix(solve(Xt%*%M%*%X))%*%Xt%*%M%*%Y
gama<-solve((Zt%*%R_1%*%Z)+solve(D))%*%Zt%*%R_1%*%(Y-X*beta[1,1])
var(gama)
Tran

```

```

EC <- Y - X%%beta[1,1] - Z%%gama # Erros Condicionais
EA <- Z%%gama # Efeitos Aleatórios
EM <- EA + EC # Erros Marginais

# ***** Estágio 2 *****
# Escolhendo os SNPs significativos
hh<-NULL
ss<-NULL
HH<-NULL
tt<-NULL
p<-NULL

hh<-matrix(NA,1,2,dimnames=list(c(),c("Marcadores","p-valor")))
NM<- vector
NM<-1:1 # numero de marcadores
hh[,1]<- NM

for(i in 1:1)
{
  hh[i,2]<-summary(lm(EC~bae[,i+5]))$coefficients[2,4]
} # Matriz com os p-valores dos SNPs
ss<-which(hh[,2]<= 0.05) # Seleciona os SNPs significativos
cat("SNP1=", ss, "\n")
cat("p1=", hh[,2][which(hh[,2]<= 0.05)], "\n")

# ***** Estágio 3 *****
if(length(ss)== 0)
{
  S1<-rbind(S1, c(0,0,0));
}
else
{
  HH <-matrix(NA,length(ss),2,dimnames=list(c(),c("Marcadores","p-valor")));
  HH[,1]<- ss;
  for(g in levels(as.factor(ss)))
  {
    #HH[w,2]<-lmekin(Y1~bae[, (as.numeric(g)+5)],data=bae, random=~1|ID,
varlist=kmat)$ctable[2,4]
    fit<-NULL;
    fit<-lmekin(Y1~bae[, (as.numeric(g)+5)],data=bae, random=~1|ID,
varlist=kmat)$ctable[2,4];
    p<-rbind(p, fit);
    #cat("g=", g, "-", "p=", p, "capiroto", as.numeric(g)+5, "\n");
  }

  tt<- which(p<= 0.05)

```

```
    if(length(tt)>0)
    {
        S1<-rbind(S1, c(length(ss),length(tt),tt[1]))
    }
    else
    {
        S1<-rbind(S1, c(length(ss),0,0))
    }
}
}
S1
```

Referências Bibliográficas

- 1- Almasy, L and Blangero, J. (1998). Multipoint Quantitative-trait Linkage Analysis in General Pedigrees. *Am J Hum Gen.* **62**:1198-211.
- 2- Almasy, L. and Blangero, J. (2008). Human QTL Linkage Mapping. *Genetic*. doi: 10.1007/s10709-008-9305-3
- 3- Amabis, J. M. and Martho, G. R. (1997). *Fundamentos de Biologia*, Editora Moderna, São Paulo- SP.
- 4- Amin, N., Cornelia, S., Van Duijn, C. M., Aulchenko, Y. S. (2007). A Genomic Background Based Method for Association Analysis in Related Individuals. *PLoS One*. **2** (12), e1274.
- 5- Aulchenko, Y. S., Koning, D., and Haley, C. (2007). Genomewide Rapid Association Using Mixed Model and Regression: A fast and Simple Method for Genomewide Pedigree-based Quantitative Trait Loci Association Analysis. *Genetics* **177**:577-585.
- 6- Barbujani, G., Magagni, A., Minch, E. and Cavalli-Sforza, L. L. (1997). An Apportionment of Human DNA Diversity. *Proc Natl Acad Sci USA* **94**: 4516-4519.
- 7- Bearzoti, E. (1997). *Simulação de Recorrente Assistida por Marcadores Moleculares em Espécies Autógamas*. Tese de Doutorado, ESALQ-USP, Piracicaba-SP.
- 8- Bearzoti, E. (2000). Análise de QTL no Melhoramento de Plantas. 2ª jornada em genética e melhoramento de plantas. FUNAPE, 63-209.
- 9- Birren, B., Green, E. D., Hieter, P., Klapholz, S., Myers, R. M., Riethman, H., Roskams, J. (1999). *Genome Analysis- A Laboratory Manual, Mapping Genomes*, vol. 4, CSHL Press.
- 10- Blangero, J. (2004). Localization and Identification of Human Quantitative Trait Loci: King Harvest has Surely Come. *Current Opinion in Genetics & Development* **14**:233-240.
- 11- Callegari- Jacques, S. M., Grattapaglia, D., Salzano, F. M., Salamoni, S. P., Crossetti, S. G., Ferreira, M. E., Hutz, M. H. (2003). Historical Genetics: Spatiotemporal Analysis of the formation of the Brazilian Population. *Am J Hum Biol* **15**:824-834.

- 12- Cardon, L. R. and Abecasis, G. R. (2000). Some Properties of Variance Components Model for Fine- Mapping Quantitative Trait Loci. *Behavior Genetics*, Vol. 30, N° 3.
- 13- Carlson, L. E., Speca, M., Patel, K. D., Goodey, E. (2003). Mindfulness-Based Stress Reduction in Relation to Quality of Life, Mood, Symptoms of Stress, and Immune Parameters in Breast and Prostate Cancer Outpatients. *Psychosomatic Medicine*. **65**: 571-581.
- 14- Carter, T. F. and Falconer, D. S. (1951). Stocks for detecting linkage in the mouse and the theory of their designs. *Journal Genet.* **50**: 307-323.
- 15- Cavalli- Sforza, L. L., Menozzi, P., Piazza, A. (1994). *History and Geography of Human Genes*. Princeton. Princeton University Press. 413p.
- 16- Chakravarti, A. (1999). Population genetics- making sense out of sequence, *Nat Genet* **21** (Supl): 56-60.
- 17- Collins, F. S., Brooks, L. D. and Chakravarti, A. (1998). A DNA Polymorphism Discovery Resource for Research on Human Genetic Variation. *Genome Res* **8**: 1229-1231.
- 18- Crow, H. and Franklin, J. (1970). *An Introduction to Population Genetics Theory*. Ed. Harper & Row, New York.
- 19 - de Andrade, M., Amos, C. I., Thiel T. J. (1999). Methods to Estimate Genetic Components of Variance for Quantitative Traits in Family Studies. *Genetic Epidemiology* **17**: 64 – 76.
- 20- Demidenko, E. (2004). *Mixed Models, Theory and Applications*. Wiley- Interscience.
- 21- Duarte, N. E. (2007). *Análise Multivariada no mapeamento genético de traços quantitativos*. Dissertação (mestrado).
- 22- Falconer, D. S., Hill, W. G. and Mackay, T. F. C. (1996). *Introduction to Quantitative Genetics*, 4° edition, Pearson Prentice Hall.
- 23- Farah, S. B. (1997). *DNA segredos e mistérios*. São Paulo: Sarvier.
- 24- Ferreira, M. E. & Grattapaglia, D. (1995). *Introdução ao uso de marcadores RAPD e RFLP em análise genética*. Brasília. EMBRAPA/CENARGEN. 220p.
- 25- Fisher R. A. (1934). The effects of methods of ascertainment upon the estimation of frequencies. *Ann. Eugenics* **6**:13-25.
- 26- Foulkes, A. S. (2009). *Applied Statistical Genetics with R, for Population-based Association Studies*. Springer.

- 27- Futuyma, D. J. (1998). *Evolutionary Biology*. 3^a ed. Sinauer, Sunderland, MA. 751pp.
- 28- Giampaoli, V., Singer, J. (2009). Likelihood ratio tests for variance components in linear mixed models. *Journal of Statistical Planning and Inference*, v.139, p. 1435-1448.
- 29- Giolo, S. R., Pereira, A. C., de Andrade, M., Oliveira, C. M., Krieger, J. E. and Soler, J. M. P. (2009). Genetic analysis of age-at-onset for cardiovascular risk factors in a Brazilian family study. *Human Heredity* **68**:131-139.
- 30- Graybill, F. A. (1976). *Theory and application of the linear model*. Duxbury, North State, Massachusetts.
- 31- Griffiths, A. J. F., Wessler, S. R., Lewontin R. C. and Carroll, S. B. (2008). *Introdução à Genética*, 9^o edição, Guanabara Koogan.
- 32- Haines J. L., Pericak-Vance M. A. (1998). *Approaches to gene Mapping in Complex Human Disease*, Wiley-Liss.
- 33- Haldane, J. B. S. (1919). The combination of linkage values, and the calculation of distance between the loci of linked factors. *Journal of Genetics*. **8**: 299-309.
- 34- Harville, D. A. (1976). Extension of the Gauss-Markov Theorem to Include the Estimation of Random Effects. *The Annals of Statistics* **4**, 384-395.
- 35- Harville, D. A. (1990). BLUP (Best Linear Unbiased Prediction) and Beyond. In *Advances in Statistical Methods for Genetic Improvement of Livestock*. (D. Gianola and K. Hammond, eds.) 239-276. Springer, New York.
- 36- Henderson, C. R. (1975). Best Linear Unbiased Estimation and Prediction under a Selection Model. *Biometrics*, Vol. 31, No. 2, pp 423-447.
- 37- Hilden-Minton, J. A. (1995). *Multilevel Diagnostics for Mixed and Hierarchical Linear Models*. PhD Thesis. University of California, Los Angeles.
- 38- Hodges, J. S. (1998). Some Algebra and Geometry for Hierarchical Models, applied to diagnostics (with discussion). *Journal of the Royal Statistical Society B* **60**: 197-536.
- 39- Karlin, S. (1984). *Theoretical aspects of genetic map function in recombination processes*. In A. Chakravarti (Ed.) & Human population genetics. The Pittsburgh Symposium, 209-228.
- 40- Kosambi, D. D. (1944). The estimation of map distances from recombination values. *Annals of Eugenics*. **12**: 172-175.
- 41- Lange, K. (1997). *Mathematical and statistical methods for genetic analysis*. New York: Springer.

- 42- Lee, P. S. and Lee K. H. (2000). Genomic Analysis. *Current Opinion in Biotechnology* **11**: 171-175.
- 43- Li, C. C. and Mantel, N. (1968). A simple method of estimating the segregation ratio under complete ascertainment. *Am. J. Hum. Genet.* **20**:61-81.
- 44- Liu, B. H. (1998). *Statistical Genomics: Linkage, Mapping, and QTL Analysis*. Boca Raton FL: CRC Press.
- 45- Lynch, M. and Walsh, B. (1998). *Genetics and Analysis of Quantitative Traits*. Massachusetts: Sinauer Sunderland.
- 46- McLean, A. R., William, L. S., Stroup, W. W. (1991). A unified approach to mixed linear models. *The American Statistician* **45**: 559 – 578.
- 47- Morton, N. E. (1959). Genetic tests under incomplete ascertainment. *Am. J. Hum. Genet.* **11**:1-16.
- 48- Neter, J., Wasserman, W., Kutner, M. H. (1985). *Applied linear statistical models regression, analysis of variance and experimental designs*. Richard D. Irwing, Inc.
- 49- Nobre, J. S. (2004). *Métodos de Diagnósticos para Modelos Lineares Mistos*. Dissertação apresentada ao IME-USP.
- 50- Oliveira, C. M., Pereira, A. C., de Andrade, M., Soler, J. M. P. and Krieger, J. E. (2008). Heritability of Cardiovascular Risk Factors in a Brazilian Population: Baependi Heart Study. *BMC Medical Genetics* **9**: 32.
- 51- Ott, J. (1991). *Analysis of human genetic linkage*, Revised Edition. Johns Hopkins University Press, Baltimore.
- 52- Pedrosa, M. A. F. (2006). *Composição Genética de Quatro Populações Remanescentes de Quilombos do Brasil com Base em Microsatélites e Marcadores de Ancestralidade*. Brasília.
- 53- Pritchard, J. K. and Cox, N. J. (2002). The allelic architecture of human disease genes: common disease - common variant...or not? *Hum Mol Genet*, **11**:2417-2423.
- 54- Rao, D. C., Morton, N. E., Lindsten, J., Hulten, M. and Yee, S. A. (1977). A mapping function for man. *Hum. Hered.* **27**: 99-104.
- 55- Robinson, G. K. (1991). That BLUP is a Good Thing: The Estimation of Random Effects (with discussion). *Statistical Science* **6**: 15-51.
- 56- Russel, M. W., Huse, D. M., Drowns S., Hamel, E. C., Hartz, S. C. (1998). Direct medical costs of coronary artery disease in the United States. *Am J Cardiol* **81** - 1110-5.

- 57- Sacco, R. L., Sabala, E. A., Rundek, T., Juo, S. H., Huang, J. S., DiTullio, M., Homma, S., Almonte, K., Lithgow, C. G., Boden-Albala, B. (2007). Design of a family study among high-risk Caribbean Hispanics: the Northern Manhattan Family Study. *Ethn Dis, Spring* **17**: 351-7.
- 58- Searle, S. R., Casella, G., and McCulloch, C. M. (1992). *Variance Components*. New York: Wiley.
- 59- Self, S. G., Liang, K. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio test under nonstandard conditions. *J Amer. Stat. Assoc.* **82**, 605-610.
- 60- Sham, P. (1998). *Statistics in human genetics*. London: Arnold Applications of Statistics.
- 61- Storey, J. D. and Tibshirani R. (2001). Estimating False Discovery Rates Under Dependence, with Applications to DNA Microarrays. *Technical Report 2001-28*, Department of Statistics, Stanford University.
- 62- Stephens, J. C., Schneider, J. A., Tanguay, D. A., Choi, J., Acharya, T. and Stanley, S. E. (2001). Haplotype variation and linkage disequilibrium in 313 human genes. *Science* **293**: 489-493.
- 63- Stoneking, M. (2001). Single nucleotide polymorphisms: From the evolutionary past... *Nature* **409**: 821-822.
- 64- Sturtevant, A. H. (2001). *A History of Genetics*. Cold Spring Harbor Laboratory Press and Eletronic Scholarly Publishing Project: <http://www.esp.org/books/sturt/history/>.
- 65- Thompson, R. (1969). Iterative estimation of variance components for non-orthogonal data. *Biometrics*: 26: 767-73.
- 66- Wolfinger, R. D. (1993). Covariance structure selection in general mixed models. *Comunication in Statistics- Simulation*, v. 22, n. 4, p. 1079-1106.
- 67- Ziegler, A.; Konig I. R. (2008). *A Statistical Approach to Genetic Epidemiology*. Wiley-VCH.