# Estrutura de dependência do genoma humano usando modelos com correlação entre indivíduos

## *Uma abordagem combinando modelos mistos generalizados e campos Markovianos*

Francisco José de Almeida Fernandes

Tese apresentada ao
Instituto de Matemática e Estatística
da Universidade de São Paulo
para obtenção do título de
Doutor em Ciências

Programa: Estatística
Orientadora: Profª. Drª. Júlia Maria Pavan Soler

São Paulo
Março de 2023

**Estrutura de dependência do genoma humano usando modelos com correlação entre indivíduos**

***Uma abordagem combinando modelos mistos generalizados e campos Markovianos***

Francisco José de Almeida Fernandes

Esta versão da tese contém as correções e alterações sugeridas pela Comissão Julgadora durante a defesa da versão original do trabalho, realizada em 29 de Março de 2023.

Uma cópia da versão original está disponível no Instituto de Matemática e Estatística da Universidade de São Paulo.

Comissão julgadora:

Profª. Drª. Júlia Maria Pavan Soler – IME-USP
Profª. Drª. Florencia Graciela Leonardi – IME-USP
Profª. Drª. Mariana Rodrigues Motta – UNICAMP
Prof. Dr. Sergio Russo Matioli – IB-USP
Profª. Drª. Suely Ruiz Giolo – UFPR

*DNA - **D**eus **N**os **A**ma*

*A mensagem escrita no nosso genoma*

# Agradecimentos

*O que jamais se questionou não foi provado.*
— Diderot

Agradeço aos meus pais porque mesmo sem terem oportunidade de estudo, sempre me ensinaram o valor dele. Agradeço à minha esposa que com seu otimismo me fez enxergar oportunidades escondidas em cada obstáculo. Agradeço aos meus filhos por se encantarem com minha jornada. Agradeço à minha orientadora, cúmplice, com todas as letras, deste trabalho. Agradeço a todos colegas e professores que me conduziram ao longo de todo este caminho. Agradeço acima de tudo a Deus, por ter permitido todas essas pessoas na minha vida.

# Resumo

Francisco José de Almeida Fernandes. **Estrutura de dependência do genoma humano usando modelos com correlação entre indivíduos: *Uma abordagem combinando modelos mistos generalizados e campos Markovianos*.** Tese (Doutorado). Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2023.

Conhecer a mistura genética herdada e suas implicações, tanto nas características gerais (fenótipos) quanto nas eventuais doenças hereditárias, é fundamental para compreender nossa história ancestral bem como nortear tratamentos médicos. A forma como os blocos de material genético estão estruturados no genoma e como são transmitidos é específico a populações e pode ser analisado através do levantamento de uma estrutura de dependência entre porções cromossômicas. O objetivo deste trabalho é propor uma metodologia estatística para estimar a estrutura de dependência entre marcadores moleculares do genoma humano levando em conta a estrutura dos dados, isto é, se a amostra consiste de indivíduos independentes ou se há relações de parentesco entre eles. Quando a amostra é formada por conjuntos de indivíduos com relação de parentesco (dados de famílias), é mais provável que eles compartilhem entre si grandes porções de material genético. Deste modo, obter regiões de dependência dentro do genoma usando dados de família, impõe um desafio adicional, relativamente ao caso de indivíduos independentes, uma vez que deve-se levar em conta que a dependência genômica pode trazer informação do parentesco entre os indivíduos da amostra. Neste trabalho, utilizamos dados de marcadores moleculares de plataformas SNP-array (do inglês *Single Nucleotide Polimorphism*) que, por sua grande densidade ao longo de todo o genoma, são considerados uma amostragem informativa da variabilidade genética humana. Cada marcador SNP é quantificado de acordo com o número de alelos alvo que carregam, podendo ser 0, 1 ou 2 alelos, descrevendo assim, em cada loco, uma variável aleatória com distribuição Binomial em dois ensaios independentes. O segmento genômico orientado, pode ser representado por uma sequência dessas variáveis aleatórias. A metodologia proposta combina a flexibilidade de Modelos Lineares Generalizados Mistos (MLGM), para acomodar na estimação a dependência familiar entre indivíduos (matriz de parentesco), com a abordagem de campos Markovianos (univariados), para encontrar o contexto (vizinhança) necessário para determinar o estado dos SNPs no genoma. Esta alternativa incorpora as duas dimensões de dependência envolvidas no problema que estamos tratando, isto é, entre indivíduos na amostra e entre marcadores no genoma, coerentemente à realidade biológica. Estabelecendo uma comparação da modelagem via MLGM e sob o modelo linear generalizado (sem considerar a dependência entre os indivíduos), é possível inferir o quanto da estrutura de dependência do genoma deve-se ao efeito de família. Um índice é proposto para quantificar a influência familiar em cada porção genômica. O algoritmo foi implementado na linguagem R e aplicado em estudos de simulação e a dados de famílias brasileiras, permitindo mapear a influência familiar ao longo de cada cromossomo, bem como em algumas regiões gênicas associadas a doenças com componente hereditário. Em particular, a região HLA (do inglês *Human Leukocyte Antigen*) foi caracterizada, em termos dos blocos obtidos, quanto à sua homogeneidade, conservação e influência familiar.

**Palavras-chave:** Dados de família. Modelo Linear Generalizado Misto. Campo Markoviano.

# Abstract

Francisco José de Almeida Fernandes. **Human genomic dependence structure in correlated data: *An approach combining generalized mixed models and Markov random fields***. Thesis (Doctorate). Institute of Mathematics and Statistics, University of São Paulo, São Paulo, 2023.

Knowing the inherited genetic mix and its implications both in complex traits (phenotypes) and in hereditary diseases is essential to the understanding of our ancestral history and in guiding medical treatments. The way that blocks of genetic material are structured in the genome and how they are transmitted can be analyzed by inferring a dependency structure among chromosomal portions. The aim of this work is to propose a statistical methodology to estimate the dependency structure among molecular markers of the human genome, taking into account the structure of the data, that is, whether the sample consists of independent individuals or whether there are kinship relations between them. When the samples consist of sets of individuals with kinship (family data), it is more likely they can share large portions of genetic material. Thus, obtaining regions of dependence inside the genome using family data, imposes an additional challenge, regarding the case of independent individuals, since it must be considered that this dependence may be due to the relationship among individuals in the sample. In this work, we used molecular markers from SNP(Single Nucleotide Polymorphism)-Array platforms, which present high density throughout the entire genome and are considered informative of the human genetic variability. Each SNP marker is quantified according to the number of target alleles they carry (0, 1, or 2), thus describing, in each locus, a random variable with Binomial distribution in two trials. The oriented genomic segment can be represented by a sequence of these random variables. The proposed methodology combines the flexibility of Generalized Linear Mixed Models (GLMM) to accommodate the family dependency among individuals (kinship matrix), with Markov random fields, to find the context (neighborhood) necessary to determine the state of the SNP. This approach incorporates the two dimensions of dependency involved: among individuals and among markers, coherently with biological reality. Compared to Generalized Linear Modeling (without considering the dependence among individuals), it is possible to infer how much of the genome's dependency structure is due to the family effect. An index is proposed to quantify the familial influence on each genomic portion. The algorithm was implemented in the R language and applied in simulation studies and data from Brazilian families, allowing the mapping of family influence along each chromosome, as well as in some gene regions associated with diseases with a hereditary component. Particularly, the HLA region (*Human Leukocyte Antigen*) was characterized, in terms of the blocks obtained, regarding its homogeneity, conservation, and familial influence.

**Keywords:** Family data. Generalized Linear Mixed Models. Markov Random Field.

# Lista de Abreviaturas

| | |
|---|---|
| AIM | *Ancestry Informative Markers*<br>Marcadores de informação ancestral |
| bp | *base pair*<br>pares de base |
| DNA | *deoxyribonucleic acid*<br>ácido desoxirribonucleico - molécula formadora do genoma |
| SNP | *Single Nucleotide Polymorphism*<br>Polimorfismo de nucleotídeo único |
| LD | *Linkage Disequilibrium*<br>Desequilíbrio de Ligação |
| MODY | *Maturity-Onset Diabetes of the Young*<br>diabete diagnosticada na idade madura do jovem (tradução livre) |
| GLM | *Generalized Linear Model*<br>Modelo Linear Generalizado |
| GMM | *Generalized Mixed Model*<br>Modelo Linear Generalizado Misto |
| HLA | *Human Leukocyte Antigen*<br>Antígeno Leucocitário Humano, região específica do genoma<br>contendo genes envolvidos na resposta imunológica (Klein e Sato, 2000) |
| HWE | *Hardy-Weinberg Equilibrium*<br>Equilíbrio de Hardy-Weinberg |
| IIF | Índice de Influência Familiar |
| MCGLM | *Multicovariate Covariance Generalized Linear Models*<br>Modelos Multivariados de Covariância Linear Generalizada |

# Lista de Figuras

# Lista de Tabelas

# Sumário

# Capítulo 1

# Introdução

## 1.1 A Era Genômica

Desde a divulgação da estrutura da molécula do DNA por Francis Crick e James Watson em 1953, os pesquisadores têm se debruçado para identificar e compreender a relação entre as bases nitrogenadas do nosso genoma e as características que nos classificam na espécie humana. Com o advento dos sequenciadores, verdadeiras máquinas de leitura das letrinhas do DNA, tivemos acesso à ordem em que essas bases aparecem no DNA. Na virada do século, o projeto Genoma Humano conseguiu pela primeira vez sequenciar os mais de 3 bilhões de pares de bases divididas em nossos 23 pares de cromossomos[1].

A partir desse marco, iniciou-se a busca por entender, dentro desse alfabeto biológico, onde se encontram as "palavras" e as "frases" que nos tornam iguais e nos tornam diferentes. Marcadores genômicos como o SNP, do inglês *Single Nucleotide Polymorphism* (Polimorfismo de Nucleotídeo Único), foram definidos, abrindo vários horizontes de pesquisa.

Mas, por quê comentar tanto sobre o universo biológico, avanços da genética e questões que fogem tanto do que seria esperado de um escopo de tese de doutorado em Estatística? Os avanços da Ciência, em todas as áreas, muitas vezes (senão todas) acontecem motivados por desafios de compreender melhor a natureza da qual somos feitos ou daquela que nos cerca. Na Genética não tem sido diferente. A enorme quantidade de dados produzida na era genômica obriga o seu tratamento estatístico e a simples alta dimensionalidade dos dados, por si só, já impõe a necessidade de modelos capazes de tratar essa realidade. Muito se tem descoberto sobre o significado e função do código biológico que nos configura, mas ainda estamos explorando, muito provavelmente, apenas a ponta de um enorme iceberg. Assim, antes de avançar no que a modelagem estatística tem evoluído para tratar as características próprias da Genética e antes de expor a nossa proposta nesta tese, cumpre detalhar melhor o contexto biológico no qual este trabalho se insere.

[1] Na verdade, esse sequenciamento ainda não acabou como relatado na reportagem de Marcelo Leite, publicada na Folha de São Paulo em 01/04/2022 (https://www1.folha.uol.com.br/ciencia/2022/04/novo-genoma-e-o-mais-completo-mas-nao-definitivo.shtml).

## 1.2 Recombinação Gênica e a Mistura Genética

É devido, em parte, à ocorrência de recombinação gênica que existe um aumento na variabilidade genética, conferindo igual variação aos descendentes de uma espécie formados a partir dessas células (Andrade e Pinheiro, 2002). A Recombinação Gênica é, em outras palavras, uma ferramenta biológica para nossa diversidade. Combinando o material genético dos pais durante a meiose que produz os gametas, a natureza consegue deixar essa mistura como legado de uma geração para a próxima, na formação de novos indivíduos, criando um fatiamento dinâmico ao longo das gerações, dos pais para os filhos e, consequentemente, nas populações através da sua história.

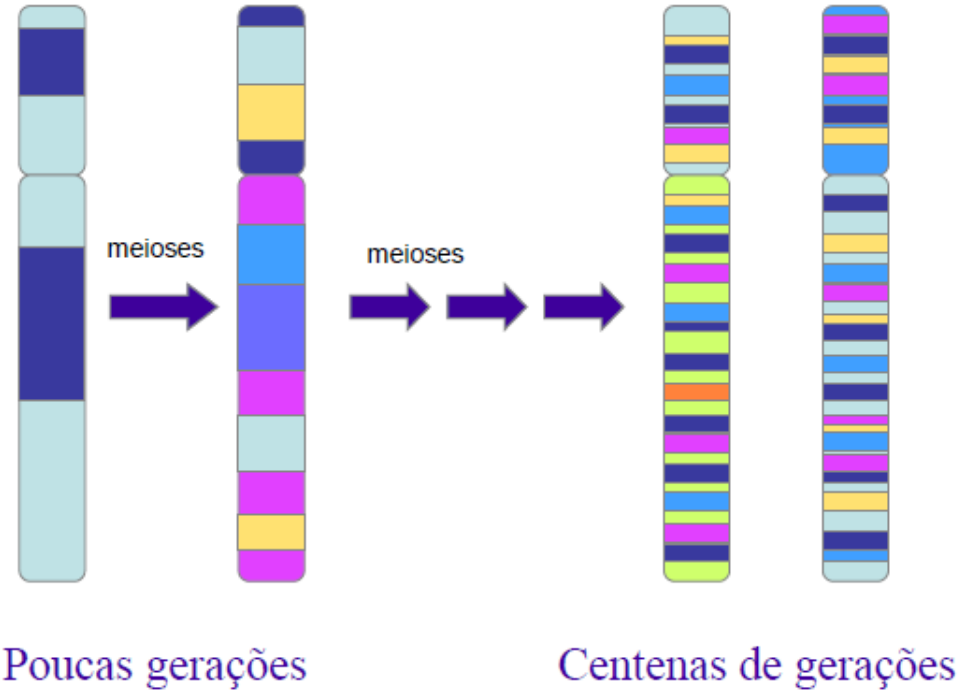



**Figura 1.1:** *Representação do processo de mistura do material genético ao longo das gerações.*

Conforme pode ser visto de forma esquemática na Figura 1.1, o tamanho das porções do genoma que são transmitidas, tende a diminuir conforme cresce o número de gerações. Toda essa divisão, no entanto, não acontece de forma simplesmente aleatória, ao contrário, o mecanismo celular garante a manutenção e estabilidade de segmentos contíguos fundamentais para nossa sobrevivência.

Mas então, como identificar essas porções de material e como caracterizar a forma na qual se modificam ao longo das gerações e nas diferentes populações? Quais delas se mantêm?

## 1.3 Marcadores genômicos e dados genotípicos

Quando falamos sobre estudos na era genômica, aludimos a um elevado domínio de ferramentas da Biologia Molecular, Bioinformática e Estatítisca. Pode-se dizer, em algum

sentido, tratar-se de estudos transdisciplinares voltados a extrair informação biológica valiosa a partir de marcadores genômicos ou moleculares. Marcadores são, em poucas palavras, variações genéticas entre indivíduos. No conjunto dos tipos de marcadores, pode-se citar os CNV's, do inglês *Copy Number Variations* ou variações no número de cópias, microssatélites, entre outros. Descrevê-los foge do escopo deste trabalho, contudo, um tipo em especial, o SNP, do inglês *Single Nucleotide Polymorphism*, será introduzido.

Nosso genoma é organizado em estruturas chamadas cromossomos. Na espécie humana temos 23 pares de cromossomos, cada par contendo um cromossomo originário de cada progenitor, sendo 22 pares denominados autossômicos e um deles chamado de sexual. Os cromossomos em cada par são chamados de homólogos.

O cromossomo por sua vez é constituído de uma molécula bastante estável, chamada de DNA, do inglês *DeoxyriboNucleic Acid*, formada por dois filamentos de açúcar e fosfato torcidos em uma dupla hélice e abrigando quatro tipos de bases nitrogenadas, a saber, timina (T), adenina (A), citosina (C) e guanina (G).

Durante muitos anos, o grande desafio dos estudos do DNA estava em entender como as bases nitrogenadas se ligavam nos dois filamentos, de forma a estabilizar a molécula. De fato, isso foi alvo de grandes disputas científicas até que finalmente, Watson e Crick conseguiram mostrar que as bases se ligavam em pares internamente aos filamentos, formando como que os degraus de uma escada. As ligações entre as bases são tais que se formam apenas dois tipos de pares: uma adenina (A) ligada a uma timina (T) ou uma citosina (C) ligada a uma guanina (G). Em outras palavras, uma determinada posição de um cromossomo pode ter o par A-T ou o par C-G, de tal forma que basta saber uma das bases para saber as duas.

Assim, considerando cada par de cromossomos homólogos, uma posição específica fica totalmente descrita com uma base do primeiro cromossomo e uma base do segundo cromossomo. Por exemplo, se o primeiro tiver as bases A-T e o segundo C-G, pode-se escrever simplesmente AC, ou AG, ou TC, ou ainda TG, sempre usando uma base de cada cromossomo. A esse par de bases dá-se o nome de genótipo e, consequentemente, os dados amostrados descritos dessa forma, de dados genotípicos[2].

Considerando-se cada par de cromossomos homólogos, o SNP, ou polimorfismo de um único nucleotídeo[3], é uma variação entre indivíduos em um único par de bases genotípicas (Figura 1.2). Os SNP's são a mais comum forma de variação no DNA humano (Geraci, 2010) e estão distribuídos ao longo de todo o genoma, sendo objeto de vários estudos, especialmente de associação com doenças. A anemia falciforme, por exemplo, é uma doença causada por um único SNP, uma mutação A → T (Lesk, 2005).

[2] Neste texto, quando nos referirmos a um par de bases no mesmo cromossomo, usaremos o hífen "-"; quando, ao contrário, o par de bases for escrito sem o hífen, refere-se ao genótipo (bases dos dois cromossomos homólogos).

[3] Cada nucleotídeo é uma molécula química composta por uma base nitrogenada (que dá o nome ao nucleotídeo), um grupo fostato e uma pentose (açúcar com 5 carbonos).

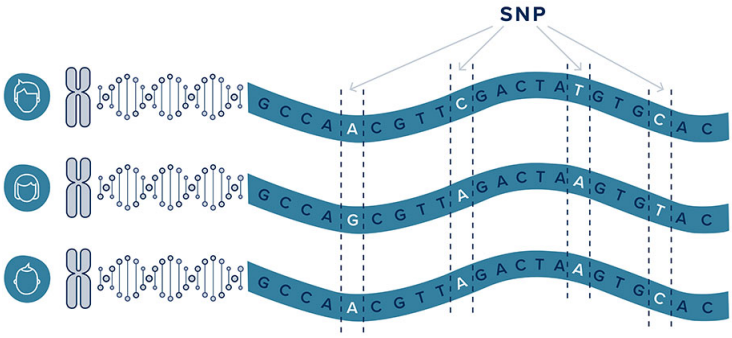



**Figura 1.2:** *Ilustração de marcadores moleculares do tipo SNP. Neste caso são representados 4 SNP's ao longo de uma região genômica com 22 pares de bases, em 3 indivíduos diferentes.*

## 1.4 Blocos de Dependência - Abordagens e Métricas

Conforme contextualizado nas seções anteriores, consideraremos as regiões genômicas contíguas, passadas dos pais para os filhos através dos cromossomos, em outras palavras, porções de material genético que "caminham junto" ao longo das gerações. A determinação dessas regiões com significado biológico (estatisticamente falando, regiões que guardam dependência interna na distribuição de suas bases nitrogenadas), pode ainda ser considerada, no entanto, um problema em aberto na literatura, especialmente quando são considerados alguns aspectos dos dados, como estrutura populacional (gerando amostras estratificadas) e parentesco entre indivíduos (observações correlacionadas) dentro da modelagem.

A caracterização dessas regiões pressupõe então uma métrica dessa dependência, seja por proximidade física (medida, por exemplo, em pares de bases) ou citogenética (medida, por exemplo, em centiMorgan), seja por funcionalidade (composição de genes[4]) ou por outras questões biológicas.

Dentre algumas medidas de dependência, destaca-se o desequilíbrio de ligação (LD do inglês *Linkage Disequilibrium*), uma medida probabilística que comumente toma como base a correlação (normalmente de Pearson) entre marcadores do tipo SNP[5]. O LD comumente aparece quando dois marcadores estão muito próximos na molécula química, de forma que é baixa a probabilidade de haver uma recombinação gênica no intervalo entre eles. Os eventos de recombinação entre dois locos são avaliados segundo um parâmetro $\lambda$, que varia entre 0 e $0,5$. Quanto mais próximo $\lambda$ estiver de 0, mais próximos (em distância citogenética) estão os dois locos e, portanto, menor a probabilidade de haver recombinação gênica nesse intervalo, consequentemente, maior o LD entre esses locos. Apesar de, em um contexto geral, ambas as medidas serem usadas no estudo de associação entre locos genéticos, não há uma função estabelecendo uma relação direta entre LD e recombinação gênica (Lange, 2002).

Vale ressaltar, que é possível encontrarmos dois SNP's em LD mesmo estando distan-

[4] Genes são sequências de nucleotídeos ao longo de uma ou mais regiões do DNA que contêm a codificação química para a elaboração de proteínas (Lesk, 2005).

[5] Existem trabalhos sobre testes de LD, que não necessariamente usam correlação, como, por exemplo, Kim *et al.*, 2008 e Ding *et al.*, 2003.

tes ou até mesmo em cromossomos diferentes; este caso acontece principalmente como resultado de migrações dentro da estrutura populacional, de forma momentânea e tende a desaparecer após algumas gerações (Laird e Lange, 2010). De fato, o LD depende de $\lambda$ e do número de gerações desde que a mutação foi introduzida na população (Jorde, 1995). É possível que eventos como os movimentos migratórios e a miscigenação, que alteram a distribuição dos alelos[6] em uma população, possam quebrar a relação entre LD e eventos de não-recombinação, conduzindo ao LD entre locos distantes (por exemplo, em diferentes cromossomos). Por essa razão, o mapeamento de genes em populações miscigenadas merece cuidados adicionais, como o uso de testes estatísticos protegidos do efeito de resultados falsos positivos, principalmente devido à presença de estratos na amostra. (Fernandes, 2016).

A busca por regiões de marcadores em LD, envolvendo, não apenas pares de locos em desequilíbrio, mas análises de multilocos em LD (blocos de SNPs em regiões genômicas com desequilíbrio na distribuição das probabilidades alélicas ou genotípicas), é a base dos métodos de mapeamento genético, e o nível do desequilíbrio esperado em uma dada região genômica e em uma dada população serve de base para a interpretação dos estudos de associação, especialmente naqueles que visam o entendimento da arquitetura genética de regulação de doenças. Esses blocos podem conter dois ou mais locos em LD, aumentando a complexidade dos padrões e níveis de LD (Fernandes, 2016).

As medidas de desequilíbrio de ligação, considerando análises alélicas e não genotípicas, são largamente utilizadas para montagem de haplótipos[7] (Druet e Georges, 2010), isto é, na reconstrução dos blocos genômicos localizados em um mesmo cromossomo. Mas esta classe de análises está restrita a um primeiro passo de processamento dos dados conhecido como *phasing*, que transforma dados genotípicos em alélicos. Iniciativas de determinação de haplótipos têm envolvido algoritmos baseados em cadeias de Markov ocultas, HMC - *Hidden Markov Chain* (Druet e Georges, 2010) e amostragem de Monte Carlo em cadeias de Markov, MCMC - *Markov Chain Monte Carlo* (Bansal *et al.*, 2008), entre outros (Bourgain *et al.*, 2002). Kim *et al.*, 2008 propôs um modelo baseado em várias cadeias de Markov de múltiplas ordens para particionar o LD de uma região cromossômica multi-loco e a participação de cada possível sub-região no LD total. Além da associação intra blocos, Greenspan e Geiger, 2006 utilizaram cadeias de Markov para mostrar que um modelo levando em conta a dependência entre os blocos de marcadores é mais acurado do que os modelos que assumem que os blocos são independentes. Além da necessidade de conhecimento da fase alélica, estes métodos, em geral, assumem blocos de tamanho conhecido e associam tais blocos a uma doença específica (Fernandes, 2016).

No entanto, quando utilizamos amostras de dados de indivíduos que tenham algum grau de parentesco, chamados simplificadamente por dados de família, não existe uma medida válida do desequilíbrio de ligação, uma vez que vários marcadores acabam por aparecer em suposta alta correlação, dado o compartilhamento de grandes regiões genômicas comuns entre eles.

Uma outra métrica de dependência ou relação entre SNP's é a epistasia, a qual pode

[6] Alelos são formas alternativas de um gene (Andrade e Pinheiro, 2002).

[7] Haplótipos são combinações locais de polimorfismos genéticos, no nível cromossômico, que tendem a ser herdados em conjunto (Lesk, 2005).

ser medida por meio de interação estatística entre dois marcadores que não estejam em LD. Neste caso, a modelagem busca o efeito conjunto (de interação) de dois marcadores em um determinado fenótipo[8]. Tal medida se mostra importante em estudos de associação de regiões genômicas com determinadas doenças, mas diz pouco sobre a estrutura de dependência do genoma, de modo mais geral.

É importante destacar que, no caso em estudo, quanto mais versátil for a estrutura de dependência modelada, melhor deve retratar a realidade biológica, tendo em vista a complexidade do fenômeno físico-químico que forma as moléculas de DNA de cada indivíduo, ao longo das gerações, isso, sem incluir os movimentos migratórios que também afetam essa estrutura. De fato, uma boa parte da origem da dependência que procuramos inferir é, em alguma medida, desconhecida.

Assim, neste trabalho, nossa proposta é explorar a estrutura de dependência do genoma, utilizando como ferramenta probabilística, os campos Markovianos. É um processo estocástico baseado na propriedade Markoviana. Dada uma sequência de variáveis, essa técnica avalia a vizinhança necessária para determinar a distribuição de probabilidades conjunta. Com isso, não é necessário impor nenhuma estrutura prévia de dependência (como a uniforme, autocorrelação, não estruturada, etc.), o que é bastante conveniente quando essa suposta estrutura é desconhecida a priori, ou ainda, composta por várias fontes de dependência, como é o caso no problema aqui exposto.

Quando a inferência da estrutura de dependência do genoma é feita utilizando dados de indivíduos relacionados, neste contexto definidos como indivíduos com alguma relação de parentesco (dados de família), uma dificuldade adicional se impõe, uma vez que, encontrados os blocos, é preciso separar o que é devido a uma relação de dependência genômica genuína presente na população em geral e o que é compartilhamento de grandes regiões genômicas segregadas entre as amostras devido ao grau de parentesco.

O genoma é passado dos pais para os filhos através dos cromossomos, pelos haplótipos. Neste trabalho, o campo Markoviano será aplicado a dados genotípicos, ou seja, considerando cada marcador nos dois cromossomos homólogos. A dependência no nível genotípico, de alguma forma, é decorrente da passagem do material genético entre gerações e assim, reflete a estrutura de dependência do genoma que queremos inferir. Consideramos então, bloco de dependência, ao conjunto de SNP's que guardam relação de dependência entre si, no nível genotípico.

## 1.5 Descrição do Trabalho

Considerando então o genoma amostrado por uma sequência de SNP's em codificação genotípica, este trabalho visa modelar a dependência entre eles e determinar assim blocos de marcadores que refletem as porções genômicas que são transmitidas, através das gerações.

Os SNP's são marcadores distribuídos ao longo de todo o genoma e, por conta da sua alta densidade, são utilizados como amostradores com uma cobertura representativa

[8] Fenótipo é o conjunto das características macroscópicas observadas em um indivíduo (Lesk, 2005).

da variabilidade presente nas populações. O projeto *1000 Genomes*, promovido por um consórcio de países, sequenciou o genoma de 2.504 indivíduos de 26 populações diferentes e reportou, em 2015, o resultado de 84.7 milhões de SNP's catalogados (Consortium, 2015). Um exemplo da concentração de SNP's pode ser vista na Figura 1.3 que mostra a região genômica do gene que codifica a proteína HLA-B27.

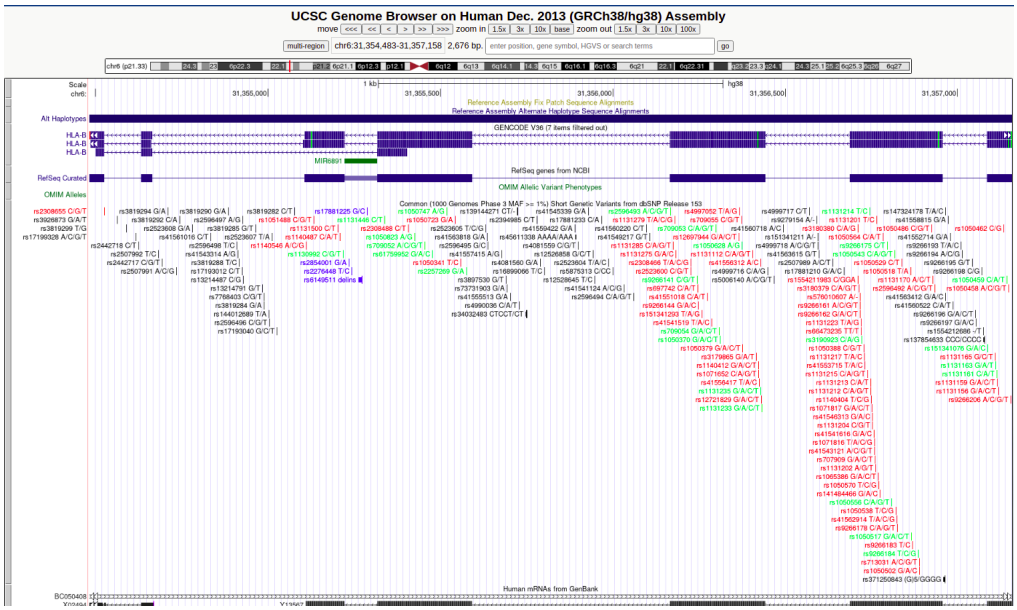



**Figura 1.3:** *SNP's catalogados na região genômica codificadora da proteína HLA-B27. A presença dessa proteína aumenta o risco de desenvolver algumas doenças reumáticas autoimunes.*

Dados genotípicos amostrados dos marcadores do tipo SNP, levam em consideração os dois cromossomos homólogos em uma determinada posição, chamada *locus*. Isso quer dizer que sabemos os dois pares de bases dessa posição mas não temos indicação direta de quais bases estão em um ou outro cromossomo. Um genótipo, considerando um dos pares (o par A-T, por exemplo) como um (1) e o outro (C-G) como zero (0), pode ser então descrito como uma variável aleatória com realização no conjunto $\{0, 1, 2\}$. Na prática, na genotipagem de um *locus*, uma base é adotada como referência, em geral a base (alelo) de menor frequência.

Há também a codificação haplotípica, na qual utilizamos as bases de cada cromossomo separadamente (codificação alélica). Nessa alternativa, o haplótipo seria codificado no conjunto $\{0, 1\}$ e a amostra de $n$ indivíduos dobra de tamanho (para $2n$), uma vez que usamos os cromossomos homólogos separadamente. Vale ressaltar que, neste caso, quando avaliamos os marcadores, é preciso saber quais estão no mesmo cromossomo, o que implica em um estudo de fase alélica (em inglês, *phasing*), utilizando cromossomos de referência, que, por vezes, pode implicar em alguma incerteza associada (Browning e Browning, 2011). A Tabela 1.1 ilustra um exemplo possíveis de codificações haplotípicas e genotípicas.

Neste trabalho os estudos foram feitos considerando dados genotípicos de SNPs, utilizando campos Markovianos para modelar a estrutura de dependência entre os marcadores. Campos Markovianos permitem modelar a dependência entre variáveis, dentro de um contexto ou vizinhança, sem impor uma estrutura prévia para a matriz de correlação.

**Tabela 1.1:** *Exemplos de codificação genotípica adotando (1) para o par A-T e (0) para o par C-G em cada SNP.*

| Haplótipos | Genótipo | Codificação Haplotípica | Codificação Genotípica |
|---|---|---|---|
| A-T C-G | T C | 1 0 | 1 |
| A-T T-A | T T | 1 1 | 2 |
| A-T A-T | T A | 1 1 | 2 |
| C-G G-C | G G | 0 0 | 0 |
| C-G T-A | G T | 0 1 | 1 |

A Figura 1.4 exemplifica o que buscamos neste estudo. Considerando uma amostra do genoma dada por uma sequência de SNPs, procuramos identificar os blocos de marcadores que refletem os blocos genômicos que caminham juntos ao longo das gerações em uma população. Espera-se que, quando utilizamos indivíduos não relacionados na amostra, estes compartilhem porções menores de material genético, que estão co-segregando na população geral. Em um conjunto de indivíduos da mesma família, no entanto, esses blocos devem maiores e, muitas vezes, co-segregando em núcleos familiares específicos.

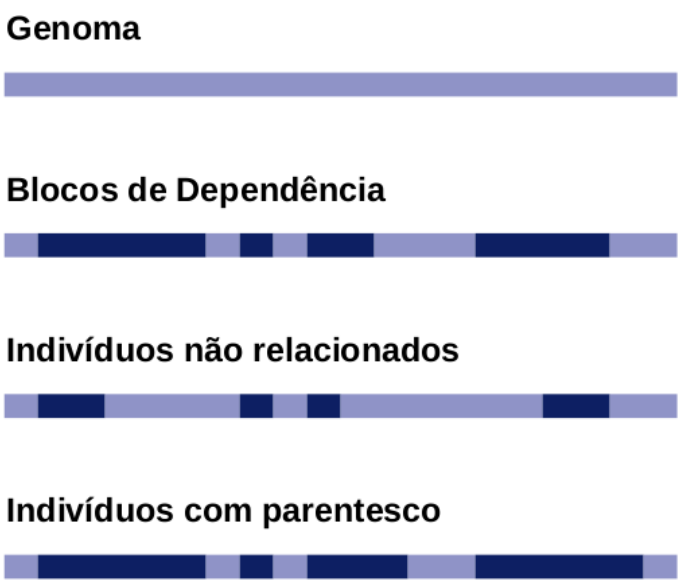



**Figura 1.4:** *Ilustração esquemática mostrando os blocos de marcadores que se deseja obter como estrutura de dependência do genoma humano, comparando seus tamanhos relativos esperados entre indivíduos com e sem parentesco.*

Ambos os delineamentos, indivíduos sem e com relação de parentesco, são importantes em estudos genéticos. Enquanto os primeiros são utilizadas em estudos observacionais e buscam por variantes comuns, os últimos, usando dados de famílias, visam mapear variantes raras, que aparecem com mais frequência em específicos núcleos familiares (Blangero *et al.*, 2013). De fato, o estudo de algumas doenças hereditárias, muitas vezes, somente é possível em famílias afetadas por elas.

O trato de dados de família, contudo, traz uma complexidade adicional à modelagem da estrutura de dependência do genoma, uma vez que o compartilhamento de maiores quanti-

dades de material genético, reflete-se diretamente no tamanho do bloco que caminha junto ao longo das gerações. Assim, ao modelar a dependência entre as variáveis (marcadores) da amostra, deve-se levar em conta a estrutura familiar que está presente nos dados.

Indo além, questiona-se, quanto da estrutura de dependência modelada, é decorrente da influência familiar?

## 1.6 Objetivo

Nosso trabalho tem por objetivo inferir a estrutura de dependência do genoma, considerando-a descrita por blocos de marcadores do tipo SNP que guardam entre si alguma relação de dependência. O propósito principal é identificar os blocos que refletem as regiões genômicas que segregam na população geral, apesar de usarmos dados de família, retirando, portanto, as regiões de dependência explicada pelo parentesco. Tal identificação tem importância tanto em estudos genéticos, como em pesquisas de medicamentos e tratamentos.

Utilizamos dados genotípicos de SNPs em amostras de indivíduos relacionados (dados de família). Optamos pela modelagem através de Campos Markovianos, que permitem modelar a dependência entre as variáveis sem impor nenhuma estrutura prévia de correlação, combinados à ferramenta de modelagem linear generalizada com efeitos aleatórios para tratar a dependência entre indivíduos imposta pela relação de parentesco.

Adicionalmente, é proposta uma abordagem e uma métrica para quantificar a influência familiar em cada bloco de dependência modelado no genoma.

A solução, aplicada em estudos de simulação e a dados de uma amostra da população brasileira, mostrou resultados consistentes com o que é esperado da resposta biológica para regiões específicas do genoma.

## 1.7 Organização do Trabalho

O trabalho está organizado de tal forma que no Capítulo 2 são descritos os conceitos e fundamentação teórica da solução. O Capítulo 3 tem a discussão dos resultados obtidos para dados simulados de indivíduos não relacionados e com relação de parentesco, enquanto que o Capítulo 4 traz algumas análises dos resultados para dados de famílias de uma amostra da população brasileira. Finalmente, no Capítulo 5 estão apresentadas conclusões, discussões e propostas de extensões do trabalho.

# Capítulo 2

# Conceitos

Este capítulo descreve os conceitos e notações considerados no trabalho, bem como a fundamentação teórica da solução proposta.

## 2.1 Conceitos e terminologias

Nesta seção vamos introduzir alguns conceitos e terminologias que serão utilizadas ao longo do texto.

### 2.1.1 Codificação dos SNP's

Conforme descrito no Capítulo 1, este trabalho utiliza dados de marcadores moleculares do tipo SNP, em uma codificação genotípica. Assim, dada uma posição (*locus*) em cromossomos homólogos, o par $(x, y)$ que indica o genótipo, é tal que:

$$(x, y) \in \{\text{A,C,T,G}\}^2,$$

em que, $A = \{\text{A,C,T,G}\}$ é o alfabeto com $|A| = 4$ sendo a cardinalidade de $A$ e os elementos deste conjunto são as bases nitrogenadas, adenina (A), citosina (C), timina (T) e guanina (G). Em outras palavras, uma sequência de SNP's pode ser descrita como uma sequência de variáveis aleatórias com valores assumidos dentro de um alfabeto finito e conhecido. Adicionalmente, uma vez que os pares de bases em cada cromossomo sempre serão do tipo A-T ou C-G, pode-se, sem perda de generalidade, codificar um deles como 0 (zero) e o outro como 1 (um). Dessa forma, o genótipo de um SNP pode, considerando a soma das codificações dos pares de bases, assumir um valor no conjunto ternário $\{0, 1, 2\}$ (ver Tabela 2.1).

**Tabela 2.1:** *Codificação alélica e genotípica das bases nitrogenadas em um determinado locus*

| Bases | Alelos | Genótipos |
|---|---|---|
| $\{\text{A,C,T,G}\}^2$ | $\{0, 1\}^2$ | {0,1,2} |

Portanto, cada SNP pode ser representado como uma variável aleatória com realização no conjunto $\{0, 1, 2\}$. Assim, um SNP em uma posição específica $j$ do genoma, configura

uma variável aleatória definida como:

$$Y_j = y_j; \quad y_j \in \{0, 1, 2\}, \quad j = 1, 2, ..., s,$$

com o conjunto $S$ formado pela sequência de $s$ marcadores, tal que, iniciando na posição 1, temos:

$$S = \{Y_1, Y_2, Y_3, ..., Y_s\} \qquad \text{sendo, } |S| = s \quad \text{a cardinalidade de S,}$$

cujas $s$ realizações serão, por exemplo, do tipo,

$$001121221222 \quad ... \quad 11010202001002200110102 0.$$

### 2.1.2 Janelas e blocos de dependência

Nosso objetivo é inferir uma estrutura que reflita as porções de material genético que são transmitidas através das gerações. Estatisticamente, amostrando o genoma como uma sequência de SNP's, procuramos por conjuntos de marcadores contíguos que guardam entre si alguma relação de dependência.

Assim, é preciso modelar a dependência de cada SNP com os seus vizinhos, tanto aqueles que o precedem, quanto os que o sucedem, em uma sequência finita de marcadores. À dependência de cada SNP para com os seus vizinhos, chamamos, por simplicidade, de "vizinhança do SNP". Aos SNP's que estão incluídos na sua vizinhança, denominamos "janela de dependência do SNP".

Uma vez determinada a vizinhança de cada SNP, um bloco de SNP's pode ser obtido concatenando-se suas respectivas vizinhanças, formando então o "bloco de dependência", ou apenas "bloco". A Figura 2.1 representa esse processo de forma visual. Assim, para cada SNP procuramos a sua vizinhança à esquerda, ou seja, nos marcadores que o precedem, ou ainda, no prefixo (considerando a sequência de SNP's como uma palavra), bem como a sua vizinhança à direita, nos marcadores que o sucedem ou no sufixo.

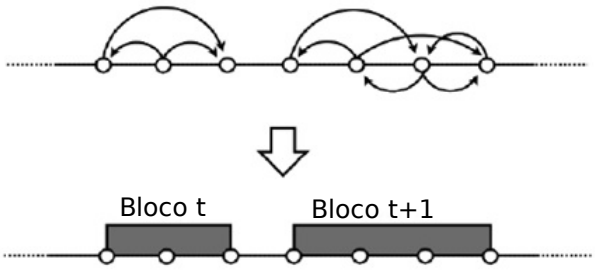



**Figura 2.1:** *Ilustração esquemática do conceito de formação dos blocos de dependência. Cada marcador é representado por um pequeno círculo e sua dependência por arcos a partir dele, a qual chamamos vizinhança do marcador. Formam-se os blocos a partir da sobreposição das vizinhanças de marcadores contíguos.*

## 2.2 Campos Markovianos

Considerando cada SNP como uma variável aleatória, um genoma amostrado por marcadores desse tipo, passa a ser uma sequência dessas variáveis aleatórias.

Reforçando, que o objetivo desse trabalho é inferir a dependência entre essas variáveis, pode-se lançar mão de algumas técnicas estatísticas para tal. A autocorrelação regressiva, muito utilizada em séries temporais ou a análise espacial de dados, usada para estudos epidemiológicos e de agricultura, entre outros, são dois exemplos de modelagem para dependência entre variáveis aleatórias. No primeiro caso, contudo, devido à componente temporal, existe claramente um conceito de passado, presente e futuro na sequência das variáveis, implicando em buscar a dependência apenas da região que antecede a variável em uma determinada posição. A segunda alternativa, deve ser precedida por uma etapa de identificação de padrões de dependência espacial (Monteiro *et al.*, 2004), o que, na verdade, é o que realmente procuramos, ou seja, em outras palavras, o objetivo deste trabalho é pressuposição inicial para a análise espacial.

Cadeias de Markov têm lugar de destaque em modelagem de dados com dependência representados por sistemas estocásticos, em particular, séries discretas de estados, como sequências de DNA, que são frequentemente modeladas por essas cadeias (Avery e Henderson, 1999). Em uma cadeia de Markov, a distribuição condicional de qualquer estado futuro, dados os estados passados, depende apenas do estado presente. É a chamada propriedade Markoviana (Ross *et al.*, 1996). Tal característica pode ser interpretada como perda de memória dentro de um processo estocástico.

Dada uma sequência de estados $Y_j, j = 1, 2, .., s$, a propriedade de Markov pode ser escrita como:

$$P(Y_{j+1} = y_{j+1} \mid Y_1 = y_1, ..., Y_j = y_j) = P(Y_{j+1} = y_{j+1} \mid Y_j = y_j).$$

O conceito de Markov pode ser estendido com as chamadas cadeias de Markov de ordem $k$, nas quais o estado futuro depende dos $k$ estados anteriores, incluindo o atual. Pode-se interpretar esse tipo de cadeia como um processo de Markov de memória $k$:

$$P(Y_{j+1} = y_{j+1} \mid Y_1 = y_1, ..., Y_j = y_j) = P(Y_{j+1} = y_{j+1} \mid Y_{j-k+1} = y_{j-k+1}, ..., Y_j = y_j), \quad k < j.$$

O número de parâmetros para estimação de uma Cadeia de Markov de ordem $K$, no entanto, cresce exponencialmente com $k$. Tal problema pode ser resolvido com o conceito de cadeias de Markov de alcance variável (Bühlmann e Wyner, 1999), na qual, a ordem da cadeia varia com cada estado.

Tais abordagens, a exemplo das séries temporais, ainda são adequadas para tratar dados com uma relação temporal de passado-presente-futuro. No caso específico deste trabalho, o objetivo é, na verdade, encontrar os SNP's no entorno de um determinado SNP de interesse. Assim, dado um SNP na posição $j$, é preciso levar em consideração ambos os lados para determinar sua vizinhança:

De fato, considerando a amostragem do genoma utilizando SNP's como uma sequência

de variáveis aleatórias, pode-se definir um campo aleatório o qual pode ser modelado através do conceito de campos Markovianos, nos quais, ao contrário das cadeias de Markov, a vizinhança de cada estado pode conter estados antes ou depois dele. Assim, a determinação da distribuição de probabilidades de um SNP, deve ser calculada dada sua vizinhança à esquerda e à direita:

$$P(Y_j = y_j \mid Y_1 = y_1, ..., Y_{j-1} = y_{j-1}, Y_{j+1} = y_{j+1}, ..., Y_s = y_s) =$$
$$= P(Y_j = y_j \mid Y_{j-l} = y_{j-l}, ..., Y_{j-1} = y_{j-1}, Y_{j+1} = y_{j+1}, ..., Y_{j+r} = y_{j+r}),$$
$$j - l \geqslant 1, j + r \leqslant s, j = 1, 2, ..., s.$$

Pode-se interpretar os campos Markovianos como uma generalização das cadeias de Markov, nas quais, para predizer um estado, é preciso levar em conta os estados da sua vizinhança (Löcherbach e Orlandi, 2011). A região de vizinhança é denominada "contexto"[1], sendo assim uma extensão multidimensional da noção de cadeias de Markov de alcance variável, introduzidas por Rissanen em seu artigo clássico (Rissanen *et al.*, 1983). Para o problema aqui proposto, o uso de Campos Markovianos considera, portanto, uma versão bidirecional das cadeias de Markov de alcance variável.

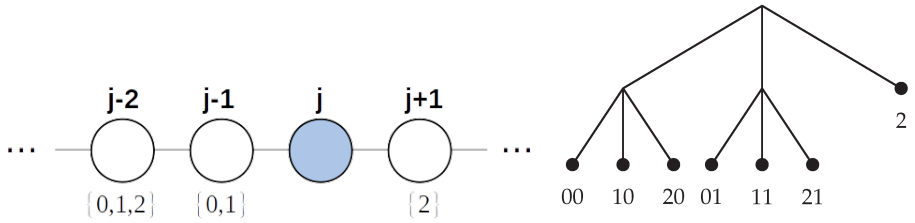



**Figura 2.2:** *À esquerda, uma sequência de SNP's - cada SNP pode ser considerado um estado dentro de uma cadeia de Markov. À direita, a árvore de contexto equivalente à vizinhança necessária para determinar a distribuição das probabilidades condicionais de um determinado estado (SNP j).*

Dado um campo Markoviano, contexto é a vizinhança à esquerda e à direita de um determinado estado, necessária para determinar a sua distribuição de probabilidades conjunta. Transportando para o problema deste trabalho, dado um determinado SNP (estado), os valores assumidos pelos SNP's vizinhos (vizinhança ou contexto), dentro da amostra, configuram a árvore de contexto. A Figura 2.2 ilustra um exemplo de sequência de SNP's com sua respectiva árvore de contexto. No caso da Figura, em particular, o SNP de interesse tem uma vizinhança de 2 SNP's à esquerda e 1 SNP à direita.

## 2.3 Notação

### 2.3.1 Matriz de dados

Conforme referido no Capítulo 1 e também na Seção 2.1.1, cada SNP pode ser interpretado, tomando a codificação genotípica, como uma variável aleatória com realização no

[1] A vizinhança pode também ser definida como uma área, um volume ou até mesmo um espaço índice multidimensional.

conjunto ternário $\{0, 1, 2\}$. O genoma amostrado pelos SNP's será então uma sequência dessas variáveis.

Considerando agora uma amostra de $n$ indivíduos, o conjunto de todos os SNP's da amostra pode ser representado por uma matriz $D_{(n \times s)}$, na qual cada linha $D_{i.} = y^{i}_{1,s}$ contém a codificação dos genótipos do indivíduo $i$ da amostra na sequência avaliada, enquanto que cada coluna $D_{.j} = y^{1,n}_{j}$, configura o conjunto dos valores da variável $Y_j$, em todos os indivíduos da amostra (ver Figura 2.3).

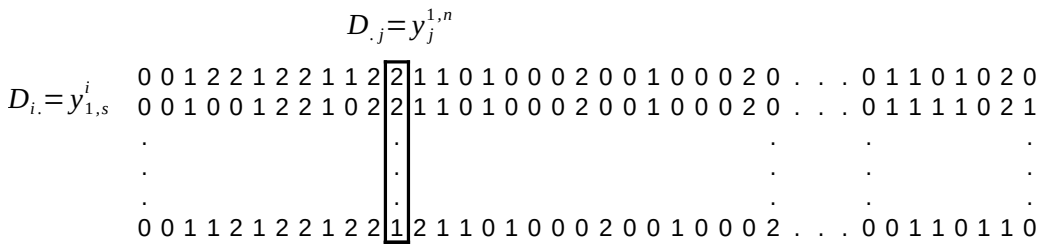



**Figura 2.3:** *Representação da matriz de dados, considerando uma amostra de s SNP's em n indivíduos.*

Importante ressaltar que neste tipo de problema, em geral, os indivíduos são da ordem de centenas enquanto que as variáveis são da ordem de centenas de milhares (dados superdimensionados). Ainda, nesse espaço retangular $n \times s$, a estrutura das colunas define uma matriz quadrada $s \times s$ com as dependências entre as $s$ variáveis e a estrutura das linhas define uma matriz quadrada $n \times n$ contendo as dependências entre os $n$ indivíduos da amostra. O espaço dual de busca de dependência é de fundamental importância para a compreensão da abordagem utilizada neste trabalho, na modelagem de dados de família.

### 2.3.2 Vizinhança ou contexto

Dentro da abordagem de campos Markovianos, o contexto ou vizinhança é o conjunto de SNP's à esquerda e à direita de um determinado SNP de interesse, necessários para caracterizar a sua distribuição de probabilidades conjunta. Vamos definir vizinhança ($V$) como um conjunto de índices,

$$V_j^{l_j, r_j} = \{j - l_j, ..., j - 1, j + 1, ..., j + r_j \mid j, l_j, r_j \in \mathbb{N}, j - l_j \geqslant 1, j + r_j \leqslant s\}, \quad j = 1, 2, ..., s,$$

em que,

$l_j$ é o número de SNPs vizinhos à esquerda da posição $j$,
$r_j$ é o número de SNPs vizinhos à direita da posição $j$,
$s$ é o número total de SNPs da amostra.

A partir dessa definição, os valores de $l_j$ e $r_j$, para $j = 1, ..., s$, determinam o conjunto de dependências entre as variáveis aleatórias que representam os genótipos dos SNPs, em

cada posição $j$ específica da amostra, ou seja, dada a vizinhança $V_j^{l_j,r_j}$, se $j' \notin V_j^{l_j,r_j}$ implica que $Y_j$ é independente de $Y_{j'}$ dados $Y_k$, $k \in V_j^{l_j,r_j}$.

Dessa forma, conseguimos escrever a propriedade Markoviana do campo aleatório da sequência de SNP's como:

$$P(Y_j = y_j \mid Y_1 = y_1, ..., Y_{j-1} = y_{j-1}, Y_{j+1} = y_{j+1}, ..., Y_s = y_s) = P(Y_j = y_j \mid Y_k = y_k, k \in V_j^{l_j,r_j}).$$

Convém ainda formalizar o conceito de prefixo e sufixo dentro da vizinhança. Dado um alfabeto A, uma palavra P de tamanho t fica definida como uma sequência das letras desse alfabeto, ou seja, $P \in A^t$. Considerando uma posição j dentro da palavra P, o conjunto de letras antes dessa posição, é chamado de prefixo, enquanto que o conjunto das letras após, é chamado de sufixo.

$$P = \omega j \rho, \quad \text{sendo: } \omega = \text{prefixo e } \rho = \text{sufixo},$$

e ainda, $|\omega| = l$ e $|\rho| = r$, em que, $l$ e $r$ são as cardinalidades de $\omega$ e $\rho$ , respectivamente.

Estendendo esse conceito para a sequência de SNP's, dado um SNP na posição $j$, denominamos $\omega$ o conjunto de SNP's da vizinhança à sua esquerda e $\rho$, o conjunto de SNP's à direita da vizinhança. À concatenação $(\omega \cdot \rho)$ chamamos simplesmente de SNP's vizinhos do SNP de interesse na posição $j$.

### 2.3.3 Blocos de dependência

Finalmente, uma vez identificada a vizinhança de cada SNP, a montagem dos blocos de dependência que procuramos é feita a partir da concatenação das vizinhanças dos SNP's contíguos, assim como descrito na Seção 2.1.2 e ilustrado na Figura 2.1.

A vizinhança $V_j$ estimada é uma medida da dependência do SNP na posição $j$ em relação aos seus vizinhos à esquerda e à direita, sendo que, $l_j$ e $r_j$, para cada marcador $Y_j$, definem a janela de dependência do SNP na posição $j$. Analogamente, os pares $(l_{j-1}, r_{j-1})$ e $(l_{j+1}, r_{j+1})$ definem as janelas de dependências dos SNP's vizinhos, nas posições $j-1$ e $j+1$, respectivamente. Avaliando a sobreposição das janelas de dependência dos SNP's adjacentes, pode-se definir um bloco $B$ como sendo a sequência de SNP's cujas vizinhanças estão contidas em $B$, em outras palavras, os índices dos marcadores que compõem o bloco.

Formalmente, cada bloco de dependência $B$ é um par de valores, $(a, b)$, com os índices do primeiro e último SNP que estão no bloco:

$$B \in \{(a, b) \mid a, b \in \{1, ..., s\}, a < b\} \quad e \quad \forall j \in B \Rightarrow V_j \subseteq B.$$

## 2.4 Modelagem de Dados Independentes

Consideramos neste caso, uma amostra aleatória de indivíduos (unidades amostrais) independentes, quando estes não guardam entre si qualquer grau de parentesco (como em

geral acontece em estudos genéticos[2]). Apesar do interesse principal deste trabalho estar em dados de família, torna-se importante avaliar a abordagem da dependência genômica utilizando campos Markovianos, primeiramente, em dados independentes (amostra aleatória de unidades amostrais independentes), uma vez que a modelagem para dados de família emprestará parte desses resultados.

Retornando à matriz de dados $D$ introduzida na Seção 2.3.1, lembramos que cada coluna é, na verdade, uma amostra de tamanho $n$ (sendo $n$ o número de indivíduos na amostra) de um determinado marcador genômico. Com a suposição de independência nos dados amostrais, cada marcador é então uma variável aleatória independente e identicamente distribuída. De fato, cada coluna $D_{\cdot j}$ da matriz $D$ pode ser considerada como uma amostra aleatória de tamanho $n$ com distribuição Trinomial, tal que,

$$Y_j \sim Trinomial(n; \phi_j) \quad Y_j^i = y_j^i \in \{0, 1, 2\},$$

em que: $\begin{cases} j = \textit{posição do SNP}, \quad 1 \leqslant j \leqslant s; \quad s = \textit{número total de SNPs}, \\ i = \textit{indivíduo}, \quad 1 \leqslant i \leqslant n; \quad n = \textit{número de indivíduos na amostra}, \\ \phi_j = \textit{é o vetor de parâmetros da Trinomial, contendo as probabilidades} \\ \qquad \textit{de ocorrências dos valores 0, 1 e 2, restritas a somarem 1.} \end{cases}$

Com essas suposições, segue-se a determinação da função de verossimilhança do modelo.

### 2.4.1 Função de Pseudo-Verossimilhança

Assumindo a independência entre os indivíduos da amostra, considerando um determinado marcador $j$, temos que:

$$Y_j^i \perp Y_j^{i'} \quad \text{se} \quad i \neq i'.$$

Assim, a probabilidade conjunta dos indivíduos da amostra, para os dados do marcador $j$, pode ser escrita como um produtório:

$$P(Y_j) = P(Y_j^1, ..., Y_j^n) = \prod_{i=1}^{n} P(Y_j^i = y_j^i); \quad \forall j \quad 1 \leqslant j \leqslant s.$$

Ainda, dado um indivíduo $i$, cada SNP é suposto independente condicionalmente ao restante dos SNP's. Assim, a partir da matriz de dados $D$, considerando $\theta$ um vetor de parâmetros associado à distribuição de probabilidade de $Y_j$, $j = 1, 2, ..., s$, a função de verossimilhança pode ser escrita como:

$$L(\theta \mid D) = \prod_{i=1}^{n} \prod_{j=1}^{s} P(Y_j = y_j^i \mid Y_k = y_k^i, \forall k; k \neq j). \tag{2.1}$$

[2] Em alguns trabalhos pode-se considerar como independentes, os indivíduos acima de um determinado grau de parentesco.

Para determinar a vizinhança de influência de um determinado SNP $j$, precisamos estimar a vizinhança $V_j^{l_j,r_j}$ tal que:

$$P(Y_j = y_j^i \mid Y_k = y_k^i, \forall k; k \neq j) = P(Y_j = y_j^i \mid Y_k = y_k^i, \forall k; k \in V_j^{l_j,r_j}). \tag{2.2}$$

Para estimar o tamanho da vizinhança $V_j^{l_j,r_j}$ centrada no SNP $j$, supomos que os SNPs dessa vizinhança têm influência sobre o SNP $j$, ou seja, no valor assumido pela variável aleatória $Y_j$. Usando então a abordagem de campos Markovianos e introduzindo o conceito de vizinhança, obtemos a função de pseudo-verossimilhança, escrita com base na propriedade Markoviana (Besag, 1975):

$$\tilde{L}(\varphi \mid D) = \prod_{i=1}^{n} \prod_{j=1}^{s} P(Y_j = y_j^i \mid Y_k = y_k^i, \forall k \in V_j^{l_j,r_j}). \tag{2.3}$$

O vetor que parâmetros é $\varphi(\varphi_j), 1 \leqslant j \leqslant s$ e assume-se que a distribuição condicional da variável na posição $j$ é especificada, agora, em termos de um vetor $\varphi_j$, dos parâmetros $(l_j, r_j)$ desconhecidos, sendo $l_j$ a vizinhança à esquerda e $r_j$ a vizinhança à direita da posição $j$. A expressão (2.3) da pseudo-verossimilhança, difere da expressão (2.1), mas ambas convergem para os mesmos valores e os estimadores, $\hat{\varphi}_j = (\hat{l}_j, \hat{r}_j)$, são consistentes (Besag, 1975; Bianchi, 2009).

Aplicando o logaritmo, sem perder generalidade, trocando a ordem dos índices, obtém-se:

$$\ell(\varphi \mid D) = \sum_{j=1}^{s} \sum_{i=1}^{n} \log\left(P(Y_j = y_j^i \mid Y_k = y_k^i, \forall k \in V_j^{l_j,r_j})\right). \tag{2.4}$$

É preciso, portanto, determinar a expressão da probabilidade

$$P(Y_j = y_j^i \mid Y_k = y_k^i, \forall k \in V_j^{l_j,r_j}),$$

a qual pode ser reescrita em termos de prefixos e sufixos, conforme descrito na Seção 2.3.2, na forma:

$$P(Y_j = \delta \mid Y_k = y_k^i, \forall k \in V_j^{l_j,r_j}) = P(Y_j = \delta \mid Y_{j-l} = \omega_1, ..., Y_{j-1} = \omega_l, Y_{j+1} = \tau_1, ..., Y_{j+r} = \tau_r),$$

tal que, tomando $\omega = (\omega_1, ..., \omega_{l_j})$ como uma das realizações de $(Y_{j-l_j}, ..., Y_{j-1})$ , $\tau = (\tau_1, ..., \tau_{r_j})$ como uma das realizações de $Y_{j+1}, ..., Y_{j+r_j}$ e $\delta \in \{0, 1, 2\}$ uma das realizações possíveis para $Y_j$, pode-se escrever simplesmente, $P_j(\delta \mid \omega, \tau)$, em uma notação reduzida[3].

Dado o tamanho da vizinhança, ou seja, a cardinalidade $l_j$ do prefixo $\omega$ e $r_j$ do sufixo $\tau$, teremos diferentes estados dessa vizinhança, considerando aqui como estado, os diferentes valores que as variáveis $(Y_{j-l}, ..., Y_{j-1})$ e $(Y_{j+1}, ..., Y_{j+r})$ assumem na amostra.

[3] Adotou-se $\omega = (\omega_1, ..., \omega_{l_j})$ e $\tau = (\tau_1, ..., \tau_{r_j})$ ao invés do mais rigoroso que seria $\omega_j = (\omega_{(j)1}, ..., \omega_{(j)l_j})$ e $\tau_j = (\tau_{(j)1}, ..., \tau_{(j)r_j})$, respectivamente, para não sobrecarregar a notação.

Definindo então $N_j^D(\omega, \delta, \tau)$ como sendo o número de vezes que o símbolo $\delta$ aparece na posição $j$, com o estado de vizinhança $\omega$ à esquerda e $\tau$ à direita, dentro da amostra $D_{\cdot j}$, pode-se reescrever a equação (2.4) como:

$$\ell(\varphi \mid D) = \sum_{j=1}^{s} \sum_{\omega} \sum_{\tau} \sum_{\delta} \left( \log(P_j(\delta \mid \omega, \tau))^{N_j^D(\omega,\delta,\tau)} \right). \tag{2.5}$$

Maximizar a função em (2.5) nos parâmetros de vizinhança, representa maximizar cada parcela da soma em $j$, isto é, considerando cada SNP individualmente. A *j-ésima* componente da pseudo-verossimilhança em função do estado de vizinhança $(\omega, \tau)$ será,

$$\ell_j(\omega, \tau \mid D) = \sum_{\delta} N_j^D(\omega, \delta, \tau) \log(P_j(\delta \mid \omega, \tau)).$$

Dada a suposição de independência das unidades amostrais, a probabilidade $P_j(\delta \mid \omega, \tau)$ é estimada diretamente da proporção amostral dos diferentes estados de vizinhança. De fato, essa probabilidade fica definida em termos da distribuição Trinomial e a equação (2.5) pode ser reescrita como uma pseudo-verosimilhança empírica:

$$\hat{\ell}(\varphi \mid D) = \sum_{j=1}^{s} \sum_{\omega} \sum_{\tau} \sum_{\delta} \left( N_j^D(\omega, \delta, \tau) \log \left( \frac{N_j^D(\omega, \delta, \tau)}{N_j^D(\omega, \cdot, \tau)} \right) \right), \tag{2.6}$$

em que $N_j^D(\omega, \cdot, \tau)$ corresponde ao número de vezes que o estado de vizinhança $(\omega, \tau)$ aparece na amostra no entorno da posição $j$, para qualquer valor de $Y_j$. Note que a probabilidade $P_j(\delta \mid \omega, \tau)$ da expressão (2.5) foi substituída por sua estimativa $\frac{N_j^D(\omega,\delta,\tau)}{N_j^D(\omega,\cdot,\tau)}$ na expressão (2.6) empírica.

Introduzindo um termo de penalização proposto por Csiszár e Talata, 2006, considerando uma constante $c$ dada por $\frac{(|A|-1)}{2}$, onde $A$ é o alfabeto, no nosso caso, $A = \{0, 1, 2\}$, e um termo proporcional ao tamanho da vizinhança $t(\omega, \tau) = |A|^{|\omega\tau|}$, chegamos à expressão da função de pseudo-verossimilhança empírica para cada SNP:

$$\hat{\ell}_j(l_j, r_j \mid D) = \sum_{\omega} \sum_{\tau} \sum_{\delta} \left( N_j^D(\omega, \delta, \tau) \log \left( \frac{N_j^D(\omega, \delta, \tau)}{N_j^D(\omega, \cdot, \tau)} \right) \right) - \frac{(|A| - 1)}{2} |A|^{|\omega\tau|} \log(n). \tag{2.7}$$

A estimação dos parâmetros do modelo é feita calculando a função dada por (2.7) para cada tamanho de vizinhança de uma posição $j$. Os valores $l_j$ e $r_j$ serão aqueles que representam o tamanho da vizinhança que maximiza a função.

### 2.4.2 Resultados em populações heterogêneas

Considerando uma única população sob estudo, a modelagem da seção anterior foi utilizada por Leonardi, 2007, e devidamente adaptada por Bianchi, 2009. Como resultado desse estudo foi possível estruturar o genoma em blocos de SNP's, nos quais há uma relação de dependência entre os SNP's internos e independência dos SNP's externos aos blocos, constituindo assim uma estruturação para o genoma em regiões dependentes e

independentes.

Adicionalmente, no trabalho de Fernandes, 2016, a mesma abordagem foi utilizada e estendida para populações heterogêneas, em particular, considerando um conjunto de 12 populações com diferentes ancestralidades, sendo 11 populações (ver Tabela 2.2) obtidas do projeto HapMap[4] (uk, 2005) e uma amostra da população brasileira.

**Tabela 2.2:** *Populações disponíveis no projeto HapMap.*

| Rótulo | Origem da População |
|---|---|
| ASW | população do sudoeste dos EUA com ancestralidade africana |
| CEU | residentes de Utah com ancestralidade do norte e ocidente da Europa |
| CHB | chineses em Beijing, China, com ancestralidade da dinastia Han |
| CHD | chineses residentes na área metropolitana de Denver, Colorado |
| GIH | índios Gujarati em Houston, Texas |
| JPT | japoneses em Tokyo, Japão |
| LWK | residentes em Webuye, Quênia com ancestralidade Luhya |
| MEX | residentes em Los Angeles, Califórnia com ancestralidade mexicana |
| MKK | residentes em Kinyawa, Quênia com ancestralidade Maasai |
| TSI | residentes próximos a Florência, Itália com ancestralidade toscana |
| YRI | residentes em Ibadan, Nigéria com ancestralidade Yoruba |

Os resultados obtidos permitiram caracterizar as diferentes populações e, consequentemente, suas ancestralidades, de acordo com os blocos de dependência estimados. Cumpre ressaltar que o termo ancestralidade está sendo colocado aqui sem o devido rigor, apenas para indicar a origem genealógica das populações estudadas.

No trabalho de Fernandes, 2016, as populações foram agrupadas em conjuntos de acordo com a origem geográfica de sua ancestralidade (ver Tabela 2.3).

**Tabela 2.3:** *Agrupamento das populações segundo sua ancestralidade primária*

| Ancestralidade | Populações | Cor Adotada |
|---|---|---|
| Africana | ASW, LWK, YRI, MKK | roxa |
| Ameríndea | MEX, GIH | vermelha |
| Asiática | CHB, CHD, JPT | amarela |
| Européia | TSI, CEU | azul |
| Brasileira | BRZ | verde |

Na análise dos blocos de SNP's encontrados em cada população, tomando o tamanho médio, considerado como a média aritmética do número de SNP's presentes em cada bloco, as populações africanas mostraram os menores valores, enquanto que a população brasileira apresentou o maior tamanho médio entre todas as populações estudadas (ver Figuras 2.4 e 2.5).

[4] O projeto HapMap foi desenvolvido por um consórcio formado por vários países com o objetivo de identificar e catalogar as diferenças e similaridades entre o genoma de seres humanos através do levantamento do mapa dos SNP's em várias populações humanas. Esse projeto foi estendido para o *1000 Genomes*, que consiste de uma catalogação maior das variações do genoma humano.

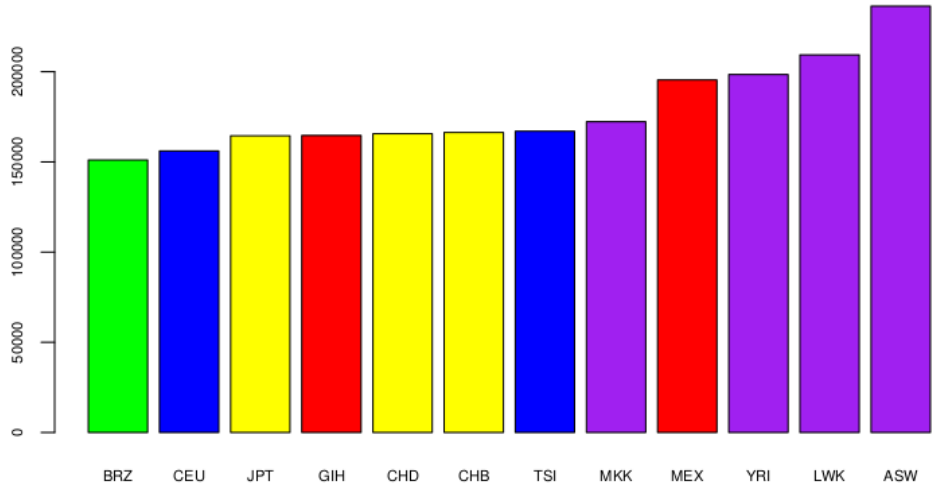



**Figura 2.4:** *Número de blocos por população.*

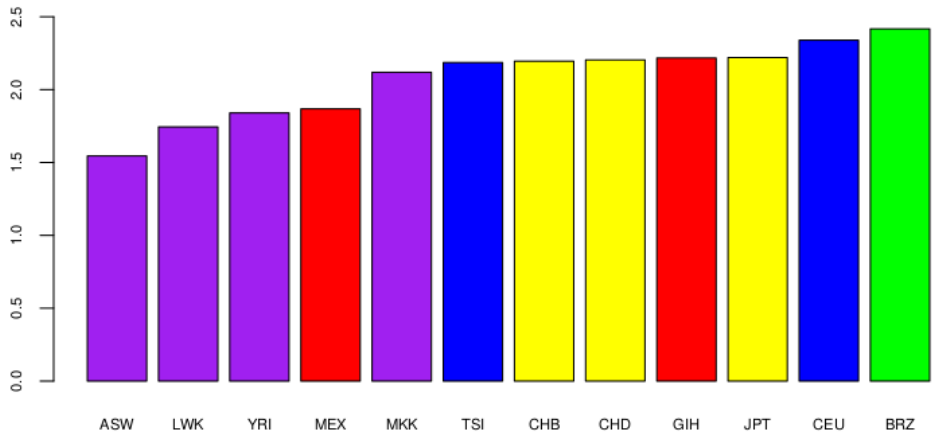



**Figura 2.5:** *Tamanho médio dos blocos por população.*

Tais resultados encontram amparo na literatura. É sabido que o genoma das populações de origem africana tem blocos haplotípicos menores quando comparados com outras populações (Schlesinger, 2010), por serem as mais antigas da história da humanidade. No trabalho de Fernandes, 2016, os blocos encontrados nas populações africanas foram mais numerosos contudo menores, de acordo, portanto, com o que é esperado, uma vez que, devido ao maior número de cruzamentos, há uma redução progressiva no tamanho dos blocos do genoma que são transmitidos aos descendentes, consequência das recombinações gênicas ocorridas. A população brasileira, ao contrário, possui menos blocos e maiores em média, um padrão representativo de porções maiores do genoma ainda presentes, evidenciando sua história recente de miscigenação (Fernandes, 2016).

Os resultados encontrados em Fernandes, 2016 mostram empiricamente que a abordagem proposta de campos Markovianos é capaz de inferir as áreas de material genômico que guardam alguma dependência, coerentemente com a realidade biológica.

## 2.5 Modelagem para Dados de Família

Indivíduos com algum grau de parentesco compartilham porções maiores de material genético entre si, comparados a indivíduos independentes (não parentes). É de se esperar, portanto, que os blocos de dependência sejam maiores quando utilizamos uma amostra com dados de família. A pergunta que se coloca é o quanto dessa dependência pode ser devida à influência familiar e o quanto tem origem em outras fontes.

### 2.5.1 Revisão Bibliográfica

Os delineamentos baseados em famílias têm sido bastante considerados em estudos genéticos. A utilização de indivíduos com a mesma base genética, elimina questões de confundimento de estrutura populacional que são, reconhecidamente, uma fonte de vício em estudos observacionais caso-controle (Laird e Lange, 2010), além de serem fundamentais, por exemplo, para estudos de variantes raras associadas à doenças com um componente hereditário. Por outro lado, impõem a necessidade de tratar as dependências entre indivíduos da amostra na modelagem.

Alguns casos particulares como, estrutura de correlação uniforme para famílias balanceadas e desbalanceadas (Konishi e Rao, 1992) e modelos de resposta multivariada na classe da distribuição Normal ou distribuições mais gerais (Oualkacha *et al.*, 2012; Bonat e Jørgensen, 2016), já encontram soluções na literatura.

Modelos lineares com componentes de variância, por exemplo, o modelo misto poligênico, são muito utilizados para modelar dados de famílias. Blangero *et al.*, 2013 utilizou um modelo linear misto com dois componentes de variância (um para a variabilidade devido a fatores genéticos e outro devido a fatores ambientais) combinado com uma abordagem de decomposição espectral da matriz de relacionamento entre indivíduos para analisar o caso univariado de resposta quantitativa sob normalidade. No modelo linear misto poligênico a matriz de covariância pode ser escrita como:

$$\Omega = 2\Phi\sigma_g^2 + I\sigma_e^2,$$

em que $\sigma_g^2$ e $\sigma_e^2$ são as componentes de variância devido a fatores genéticos e ambientais, respectivamente, $I$ representa a matriz identidade de ordem $n$ e $2\Phi$ é a chamada matriz de parentesco, composta por coeficientes de correlação entre os indivíduos da mesma família, calculados a partir do grau de parentesco ou coeficiente de relacionamento (Andrade e Pinheiro, 2002).

De Andrade *et al.*, 2002 estende essa abordagem para dados longitudinais. Oualkacha *et al.*, 2012 propõe uma extensão do modelo linear misto univariado para o caso multivariado e obtém componentes principais de herdabilidade a partir de uma metodologia adaptada para levar em consideração a estrutura familiar introduzida por Ott e Rabinowitz, 1999. Na formulação do modelo linear misto poligênico multivariado a estrutura de covariância é dada por:

$$\Omega = 2\Phi \otimes \Sigma_g + I \otimes \Sigma_e,$$

em que, $2\Phi$ e $I$ são dados como anteriormente, $\otimes$ é o produto de Kroneker, $\Sigma_g$ e $\Sigma_e$ são matrizes de ordem $s$ com as covariâncias genéticas e ambientais (devido ao termo de erro) entre as $s$ variáveis (fenótipos), respectivamente.

Huggins, 1993 apresenta uma formulação robusta do modelo de componentes de variância em dados de família na classe das distribuições elípticas. Epstein *et al.*, 2009 discutem limitações do uso do modelo linear misto poligênico caso a suposição de normalidade seja violada, entre elas, estimativas viesadas para os parâmetros e erros tipo 1 elevados. Acrescentam ainda que mesmo uma transformação nos dados pode não ser suficiente para eliminar os problemas de inferência e defendem que, se a distribuição for conhecida, modelá-la pode levar a um aumento do poder e estimativas mais eficientes dos parâmetros. Os autores generalizam, então, essa classe de modelos para respostas contínuas dentro da família exponencial.

As Equações de Estimação Generalizadas (Liang e Zeger, 1986) são uma abordagem para tratar dados correlacionados, em especial, em estudos com medidas repetidas. São uma extensão dos modelos de quase-verossimilhança da teoria de Modelos Lineares Generalizados (Wedderburn, 1974; McCullagh, 1983) e conseguem estimativas para os parâmetros de regressão e de correlação, mas dependem de que tanto o modelo quanto a estrutura da matriz de correlação estejam corretamente especificados. São mais utilizados quando o interesse é maior no efeito médio na população em estudo, enquanto que os modelos lineares generalizados mistos, se aplicam mais em situações onde há heterogeneidade entre as unidades amostrais, modelado pelo efeito aleatório (Galdino, 2015).

Em um contexto univariado para respostas contínuas bem como discretas, Wang *et al.*, 2015 apresentam uma formulação do modelo linear generalizado misto para dados de famílias e a implementação desses resultados no programa estatístico **R** (R Development Core Team, 2009) foi feita por Chen e Conomos, 2016. Em 2016, Bonat e Jørgensen (Bonat e Jørgensen, 2016) propuseram uma classe de modelos para tratar respostas multivariadas não-normais, os chamados modelos multivariados de covariância linear generalizada (*Multivariate Covariance Generalized Linear Models* - MCGLM), permitindo o tratamento de estruturas de correlação bastante gerais, como é o caso de dados espaciais e temporais, definidas em termos de uma função de ligação para a matriz de covariância, combinada com a matriz de preditores lineares.

A modelagem proposta por Bonat e Jørgensen (Bonat e Jørgensen, 2016) estende os modelos lineares generalizados (*Generalized Linear Models* - GLM) propostos por Nelder e Wedderburn em 1972, permitindo, além do tratamento de alguns tipos de dados não facilmente tratáveis pelo GLM tradicional (assimétricos, contagem e limitados), a análise de dados de observações não independentes (como é o caso em dados genéticos de famílias), além de respostas multivariadas, inclusive quando de distribuições diferentes.

### 2.5.2 Características do problema proposto

Neste ponto é importante ressaltar que o objetivo deste trabalho é propor uma metodologia estatística para estimar a estrutura de dependência entre marcadores moleculares do genoma humano, levando em conta relações de parentesco entre indivíduos da amostra. Vale enfatizar que quanto maior o grau de parentesco entre indivíduos maior é a proba-

bilidade de compartilharem entre si grandes porções de material genético comum. Mais do que isso, deseja-se inferir o quanto de cada bloco genético compartilhado, pode ter origem nesse parentesco e o quanto corresponde ao que é co-segregado na população em geral.

Para formalização do caso de dados de famílias, vamos fazer uso de alguns resultados da metodologia descrita na Seção 2.4. Tomando a mesma matriz de dados $D$ apresentada na Seção 2.3.1 e considerando agora que ela representa dados de família, não se pode mais assumir que cada coluna é constituída de observações independentes e identicamente distribuídas (lembrando que, cada coluna, representa a amostra de um marcador específico, variável $Y_j$, avaliada em todos os indivíduos da amostra). Neste caso para encontrar a distribuição da variável aleatória que representa cada marcador, será preciso introduzir um princípio importante da genética.

Em 1908, Godfrey Hardy e Wilhelm Weinberg, independentemente, derivaram uma fórmula relacionando a probabilidade genotípica em termos de probabilidades alélicas. Apesar da imposição de várias suposições (população de tamanho infinito, cruzamentos aleatórios, entre outras), mesmo que algumas delas não estejam satisfeitas, o princípio de Hardy-Weinberg postulado por estes autores provê uma boa aproximação das probabilidades genotípicas (Laird e Lange, 2010). Como citado no Capítulo 1, somos constituídos por dois pares de cromossomos, portanto, em cada posição (homóloga) temos o genótipo constituído pelos dois alelos, materno e paterno, transmitidos. Sem perda de generalidade, considerando *loci* dialélicos, pode-se dizer que os valores alélicos possíveis sejam `A` e `a` e ainda, considerando que a probabilidade de se obter o alelo `A` seja $\pi$, as probabilidades genotípicas dos descendentes, após uma geração de cruzamento aleatório será:

$$\begin{cases} P(\text{genótipo } \texttt{AA}) = \pi^2 \\ P(\text{genótipo } \texttt{Aa}) = 2\pi(1-\pi) \\ P(\text{genótipo } \texttt{aa}) = (1-\pi)^2 \end{cases}$$

A lei do equilíbrio de Hardy-Weinberg (HWE - do inglês *Hardy-Weinberg Equilibrium*) diz que essas probabilidades permanecem constantes de geração para geração e explica a base das frequências constantes dos genes como uma aplicação do binômio de Newton (Andrade e Pinheiro, 2002).

Fugas do HWE acontecem por diversas razões, entre as quais, características amostrais (por exemplo, o uso de amostras de indivíduos com um determinado fenótipo de interesse), casamentos consanguíneos ou erros de genotipagem (Laird e Lange, 2010). Em estudos populacionais, no entanto, as probabilidades genotípicas de um determinado *locus*, atingem as proporções de equilíbrio de Hardy-Weinberg depois de uma única geração de cruzamento aleatório, independentemente das frequências genotípicas iniciais (Andrade e Pinheiro, 2002).

A prova da fórmula do equilíbrio de Hardy-Weinberg, apesar de simples, foge do escopo deste trabalho, mas a presença do HWE simplifica substancialmente a teoria e métodos estatísticos (Laird e Lange, 2010) para muitas finalidades de análise de dados genéticos.

Em nosso caso, considerando que a distribuição dos alelos nos *loci* está sob equilíbrio

de Hardy-Weinberg e seguindo a codificação de frequência citada na Seção 2.1.1, para cada marcador em cada indivíduo, pode-se definir a variável $Y_j^i$ como o número de alelos do tipo A avaliado no marcador $j$ para o indivíduo $i$, a qual pode ser, então, modelada por uma distribuição Binomial$(2, \pi)$, assumindo valores no conjunto $\{0, 1, 2\}$ (número de sucessos em duas tentativas independentes), tal que:

$$Y_j^i \sim Binomial(2, \pi_j^i) \quad Y_j^i = y_j^i \in \{0, 1, 2\}.$$

De fato, é possível formular o modelo Trinomial, sob HWE, como o produto de Binomiais, como será mostrado a seguir.

Considerando-se um conjunto de dados genotípicos (no nível do indivíduo), para um determinado *locus* temos a distribuição indicada na Tabela 2.4.

**Tabela 2.4:** *Contagem de genótipos diferentes em um locus para uma amostra de n indivíduos, bem como de suas correspondentes probabilidades*

| | AA | Aa | aa | Total |
|---|---|---|---|---|
| Amostra | $n_{AA}$ | $n_{Aa}$ | $n_{aa}$ | $n$ |
| População | $\pi_{AA}$ | $\pi_{Aa}$ | $\pi_{aa}$ | 1 |

A função de verossimilhança será, a partir de distribuição Trinomial (amostra de indivíduos independentes):

$$L(\underset{\sim}{\pi} \mid \underset{\sim}{n}) = \frac{n!}{n_{AA}!n_{Aa}!n_{aa}!}\pi_{AA}^{n_{AA}}\pi_{Aa}^{n_{Aa}}\pi_{aa}^{n_{aa}},$$

$$L(\underset{\sim}{\pi} \mid \underset{\sim}{n}) \propto \pi_{AA}^{n_{AA}}\pi_{Aa}^{n_{Aa}}\pi_{aa}^{n_{aa}},$$

com $\underset{\sim}{\pi} = (\pi_{AA}, \pi_{Aa}, \pi_{aa})$, tal que $\pi_{AA} + \pi_{Aa} + \pi_{aa} = 1$ e $\underset{\sim}{n} = (n_{AA}, n_{Aa}, n_{aa})$, tal que $n_{AA} + n_{Aa} + n_{aa} = n$.

Assumindo o equilíbrio de Hardy-Weinberg e reescrevendo os dados da Tabela 2.4 no nível cromossômico ou alélico (Tabela 2.5), teremos:

$$\text{EHW} \begin{cases} \pi_{AA} = \pi_A^2 \\ \pi_{Aa} = 2\pi_A\pi_a \\ \pi_{aa} = \pi_a^2 \end{cases} \quad \text{com} \quad 0 \leqslant \pi_A \leqslant 1 \quad \text{e} \quad \pi_a = 1 - \pi_A.$$

**Tabela 2.5:** *Contagem alélica em um locus para uma amostra de n indivíduos, bem como de suas correspondentes probabilidades*

| | A | a | Total |
|---|---|---|---|
| Amostra | $n_A = n_{Aa} + 2n_{AA}$ | $n_a = n_{Aa} + 2n_{aa}$ | $2n$ |
| População | $\pi_A$ | $\pi_a$ | 1 |

Assim:

$$L(\pi_A, \pi_a \mid \underset{\sim}{n}) \propto (\pi_A^2)^{n_{AA}}(2\pi_A\pi_a)^{n_{Aa}}(\pi_a^2)^{n_{aa}} = \pi_A^{2n_{AA}+n_{Aa}} \cdot \pi_a^{2n_{aa}+n_{Aa}} \cdot 2^{n_{Aa}},$$

$$L(\pi_A \mid n, n_A) \propto \pi_A^{n_A}(1-\pi_A)^{(2n-n_A)} = \prod_{i=1}^{n} \pi_A^{y_i}(1-\pi_A)^{2-y_i}; y_i = \begin{cases} 0 \text{ se } aa \\ 1 \text{ se } Aa \\ 2 \text{ se } AA \end{cases}$$

A expressão resultante equivale à distribuição de uma amostra aleatória da variável Binomial com duas tentativas independentes. Deste modo, a distribuição Trinomial assumida para os marcadores no caso de indivíduos independentes, sob HWE, se reduz ao modelo Binomial proposto para cada marcador em cada indivíduo:

$$Y_j^i \sim Trinomial(1; \phi_j^i) \overset{\text{sob HWE}}{\longrightarrow} Y_j^i \sim Binomial(2, \pi_j^i),$$

em que: $\phi_j^i = (P_j(AA), P_j(Aa), P_j(aa))$ e $Y_j^i = y_j^i \in \{0, 1, 2\}$ e $\pi_j^i$ é a probabilidade do alelo de referência do *locus* $j$, no indivíduo $i$.

Quando utilizamos dados de indivíduos com relação de parentesco, pode-se assumir que, variáveis aleatórias (marcadores) de diferentes indivíduos são independentes, se são provenientes de diferentes famílias. Formalizando, sob a notação adotada, em que $i$ representa um determinado indivíduo pertencente a uma determinada família $f$, temos:

$$Y_j^{i,f} \perp Y_j^{i',f'} \text{ se } i \neq i' \text{ e } f \neq f'.$$

Em outras palavras, nossos dados são estratificados de tal forma que cada família é um agrupamento, dentro do qual a dependência (parentesco) deve ser levada em consideração, bem como a independência entre os diferentes agrupamentos.

Para escrever a função de verossimilhança da mesma forma como foi feito na modelagem para dados independentes, é preciso introduzir o conceito de vizinhança nessa específica estrutura da amostra.

Considere uma posição $j$ na sequência de variáveis aleatórias sob estudo. Para uma vizinhança de $Y_j$, ou seja, fixados $l_j$ e $r_j$ em $V_j^{l_j,r_j}$, chamamos de estado da vizinhança uma determinada realização das variáveis nessa vizinhança: $v_{j,l_j,r_j}^u$ com $u \in \mathbb{N}$ sendo o número de diferentes estados encontrados em $V_j^{l_j,r_j}$. Para simplificar a notação, vamos adotar:

$$\xi^u = v_{j,l_j,r_j}^u,$$

como sendo um estado da vizinhança $V_j^{l_j,r_j}$, o qual pode ser escrito em termos de prefixos e sufixos como $\xi^u = (\omega^u, \rho^u)$. Ressalte-se que neste contexto, o índice $j$ fica subentendido na notação simplificada.

Para cada estado, obtemos um subconjunto da amostra $D$, digamos $D_{\xi^u}$, com:

$$D_{\xi^u} = \{D_{i.} \mid i \in I, (Y_{j-l_j}^i, ..., Y_{j-1}^i, Y_{j+1}^i, ..., Y_{j+r_j}^i) = \xi^u\}; \quad I = \{1, ..., n\}.$$

Assim, $D_{\xi^u}$ representa uma subamostra, com as linhas da matriz $D$ correspondentes aos indivíduos que compartilham o mesmo estado de vizinhança. Dessa forma, baseado nos diferentes valores de estado de uma vizinhança, conseguimos uma nova estratificação da amostra, tal que, dentro de cada estrato, a vizinhança da posição $j$ tem o mesmo estado. A Figura 2.6 ilustra a organização dos dados estratificados por famílias e então por estados de uma dada vizinhança. Vale ressaltar, que cada subconjunto $D_{\xi^u}$ possui uma matriz de relacionamento entre indivíduos específica.

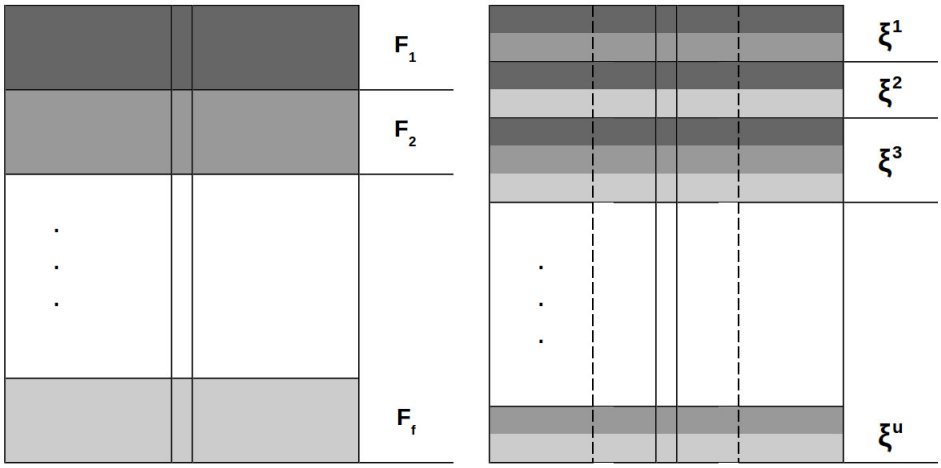



**Figura 2.6:** *Representação esquemática da estratificação da amostra por famílias e por estados de vizinhança.*

Na amostra estratificada por estados de vizinhança, é suposto que as observações são condicionalmente independentes entre diferentes estratos, mas dentro do mesmo estrato a dependência familiar, quando existe, deve ser considerada, ou seja:

$$Y_j^{i,f} \mid \xi^u \quad \perp \quad Y_j^{i',f'} \mid \xi^{u'} \quad \text{para } i \neq i' \text{ e } \xi^u \neq \xi^{u'}, \ \forall \ f \text{ e } f'. \tag{2.8}$$

### Função de Pseudo-Verossimilhança

Essas duas premissas, ou seja, a estratificação da amostra em estados de vizinhança condicionadamente independentes e a distribuição de $Y_j^i$, condicionada nos estados de vizinhança, como sendo uma Binomial em duas tentativas ($(Y_j^i \mid \xi^u) \sim Binomial(2, \pi_j^{i,u})$)[5], permite escrever a função de pseudo-verossimilhança como sendo um produto de binomiais, dado o estado de vizinhança[6]:

$$\mathcal{L}(l_j, r_j \mid D) = \prod_u P_{\xi^u}(Y_j^{l_j, r_j} \mid D_{\xi^u}),$$

[5] Para não sobrecarregar a notação, estamos adotando $\pi_j^{i,u}$ ao invés do mais rigoroso que seria $\pi_j^{i,\xi^u}$, uma vez que o parâmetro da Binomial do SNP $j$, em cada indivíduo $i$, depende do estado de vizinhança $\xi^u$.

[6] Adotamos a notação $\mathcal{L}$ para diferenciar da função de pseudo-verossimilhança para o caso de amostras independentes ($\tilde{L}$).

mais especificamente, a função de log-pseudo-verossimilhança, avaliada sob a estratificação da amostra em estados de vizinhança e condicional ao efeito aleatório de família:

$$\ell(l_j, r_j \mid D) = \sum_{u} \sum_{i=i_j^{u_j}}^{n_j^{u_j}} \sum_{\delta=0}^{2} log(P(Y_j^i = \delta \mid (Y_{j-l_j}^i, ..., Y_{j-1}^i, Y_{j+1}^i, ..., Y_{j+r_j}^i) = \xi^u, a_{(i)j} \quad \forall D_{i.} \in D_{\xi^u})), \tag{2.9}$$

em que, $\delta \in \{0, 1, 2\}$ é o conjunto das possíveis realizações para $Y_j^i$ e $\xi^u$ representa cada diferente estado de vizinhança para o SNP na posição $j$, fixando-se $l_j$ e $r_j$. O segundo somatório é aplicado a todos os indivíduos $n_j^{u_j}$ que compartilham o mesmo estado de vizinhança, sendo $i_j^{u_j}$ o primeiro deles. A independência entre os indivíduos que compartilham o mesmo estado de vizinhança é adotada condicionalmente no efeito aleatório de família ($a_{(i)j}$), o qual será definido na Seção 2.5.5.

Assim como no caso de observações independentes (Fernandes, 2016), a verossimilhança precisa sofrer uma penalização conforme aumenta o tamanho da vizinhança. Foi adotada a mesma penalização proposta por Csiszár e Talata, 2006, utilizada para observações independentes, obtendo assim a equação final da função de verossimilhança dada por:

$$\ell(l_j, r_j \mid D) = \sum_{u} \sum_{i=i_j^{u_j}}^{n_j^{u_j}} \sum_{\delta=0}^{2} log(P(Y_j^i = \delta \mid (Y_{j-l_j}^i, ..., Y_{j-1}^i, Y_{j+1}^i, ..., Y_{j+r_j}^i) = \xi^u, a_{(i)j} \quad \forall D_{i.} \in D_{\xi^u})) - \\ - \frac{(|A| - 1)}{2} . |A|^{|\omega^u \rho^u|} \log(n), \tag{2.10}$$

em que $|\omega^u \rho^u| = |j - l_j| + |j + r_j|$ é o tamanho da vizinhança e $n$ é o tamanho amostral, associado à vizinhança considerada.

A Equação (2.10), depende do par $(l_j, r_j)$ mas também de $\pi_j^{i,u}$ (da distribuição Binomial de $Y_j^i$ avaliada condicionalmente na estratificação da amostra em estados de vizinhança e no efeito aleatório de família), o qual será substituído por uma estimativa amostral consistente, conduzindo à log-pseudo-verossimilhança empírica, dada por:

$$\hat{\ell}(l_j, r_j \mid D) = \sum_{u} \sum_{i=i_j^{u_j}}^{n_j^{u_j}} \sum_{\delta=0}^{2} log(\hat{P}(Y_j^i = \delta \mid (Y_{j-l_j}^i, ..., Y_{j-1}^i, Y_{j+1}^i, ..., Y_{j+r_j}^i) = \xi^u, a_{(i)j} \quad \forall D_{i.} \in D_{\xi^u})) - \\ - \frac{(|A| - 1)}{2} . |A|^{|\omega^u \rho^u|} \log(n). \tag{2.11}$$

O par de parâmetros $(\hat{l}_j, \hat{r}_j)$ que maximizam a expressão (2.11) são a vizinhança do SNP na posição $j$:

$$(\hat{l}_j, \hat{r}_j) = \text{argmax}_{l_j, r_j \in S}(\hat{\ell}(l_j, r_j \mid D)).$$

### Espaço de busca por dependências

Na matriz de dados $D$, de dimensão $n \times s$, em cada linha temos uma sequência de variáveis aleatórias com regiões de dependência que queremos encontrar. Como entre as linhas também verificamos uma estrutura de dependência dada pelo parentesco entre os indivíduos, cada SNP de um determinado indivíduo (representado por uma posição específica dentro da matriz $D$), acaba por ter uma região de dependência em duas dimensões, como ilustrado na Figura 2.7.

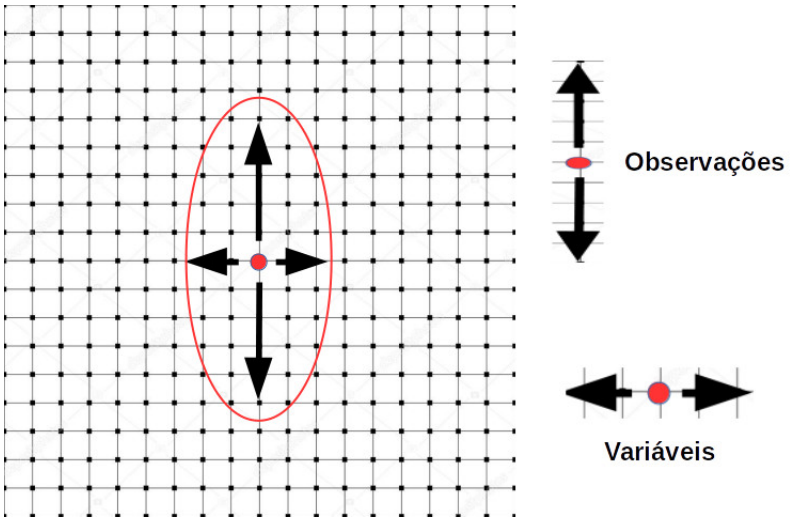



**Figura 2.7:** *Esquema ilustrativo da região de dependência de um SNP na amostra.*

Para tratar esse espaço de busca, poderíamos considerá-lo como um campo Markoviano em duas dimensões. Löcherbach e Orlandi, 2011 propuseram um estimador para o raio da região de dependência de um campo Markoviano bidimensional, o qual seria o menor raio para um círculo contendo todos os pontos necessários para definir o estado do ponto central, uma generalização, portanto, do campo aleatório linear tratado no problema com indivíduos independentes. Essa solução, entretanto, acaba por inferir o mesmo raio para a dependência entre variáveis e entre observações.

Tal suposição não parece razoável dada a natureza bastante distinta entre as duas estruturas de dependência (entre variáveis genéticas e entre indivíduos da amostra) impostas em nosso problema. Sob esta abordagem, intuitivamente, outras formas quadráticas, por exemplo, elipses, que permitem a modelagem mais adaptativa de correlações, parecem ser mais adequadas à estruturação de dados. A Figura 2.7 ilustra, de forma intuitiva, essas regiões de dependência. Adicionalmente, para o problema que estamos investigando, o campo Markoviano é linear e heterogêneo, sendo que as variáveis nesse campo são avaliadas em amostras de famílias.

Considerando o uso de modelos lineares generalizados para ambas as dimensões, uma das alternativas encontradas na literatura e citada na Seção 2.5.1, seria a utilização do

modelo multivariado de covariância linear generalizada (*Multivariate Covariance Generalized Linear Models* - MCGLM) proposto por Bonat e Jørgensen, 2016. A flexibilidade da modelagem MCGLM permite estruturas de dependência específicas tanto entre amostras como entre respostas. A utilização, portanto, apenas do MCGLM, modelando a covariância entre indivíduos por uma estrutura conhecida de parentesco e entre variáveis (genéticas) por uma matriz específica (uniforme, não-estruturada, etc), poderia ser uma solução ao problema de determinação das regiões de dependência, apesar de envolver uma posterior seleção de modelos para determinar a vizinhança do SNP, usando algum teste da esparsidade ou comparação entre as matrizes estimadas para cada vizinhança, o que precisaria ser definido de forma adequada ao problema.

Uma análise mais aprofundada dessa solução, no entanto, mostrou que, sob o MCGLM, a estimação para o primeiro e segundo momento da variável acontece de forma praticamente independente. Em outras palavras, o MCGLM estima bem o fator de escala (segundo momento) mas o preditor de posição (primeiro momento), não sofre influência da covariância. Ainda, sob essa formulação é necessário assumir uma estrutura específica para a matriz de covariância entre as variáveis (neste caso, entre os marcadores SNP). De fato, o que buscamos é uma estimativa associada ao primeiro momento da Binomial, considerando a estrutura de covariância dos dados (parentesco entre os indivíduos) nos diferentes estratos ($\xi^u$) e que permita discriminar a influência familiar nos blocos de SNP's.

### 2.5.3 Solução Combinada

Dadas as características do problema proposto, estão impostos dois desafios: como calcular a probabilidade da Binomial na expressão da pseudo-verossimilhança (Equação (2.10)), uma vez que agora temos observações não independentes e como tratar o espaço bi-dimensional de busca levando em conta as diferentes fontes de (co)variabilidade, entre indivíduos e entre os marcadores.

Isto posto, a alternativa proposta em nosso trabalho é combinar a utilização de Modelos Lineares Generalizados Mistos (GMM - do inglês *Generalized Mixed Models*) para estimar o parâmetro ($\pi_j^{i,u}$) da distribuição Binomial definida para cada variável SNP em cada indivíduo do estrato $\xi^u$, com a abordagem de campos Markovianos (univariados) para encontrar o contexto (vizinhança) necessário para determinar a distribuição de probabilidades condicional de cada SNP, incorporando assim as duas dimensões de dependência formada pelos dados, coerentemente ao que é esperado sob princípios da Biologia. Sob essa estruturação do problema, na Figura 2.8, o GMM modela a direção de dependência indicada, condicionalmente, pelas setas verticais (observações), enquanto que campos Markovianos inferem a direção de dependência indicada pelas setas horizontais (variáveis).

Essa solução combinada permite explorar propriedades das duas abordagens. A modelagem via GMM permite introduzir na análise a estrutura de correlação conhecida entre indivíduos bem como ajustes por covariáveis, se necessário, além de estimar, por máxima verossimilhança, a vizinhança do SNP, utilizando a teoria de campos Markovianos, sem a necessidade de impor uma estrutura específica de covariância entre as variáveis.

Neste sentido, na solução combinada, a estrutura de dependência bivariada é particionada de tal forma que GMM cuida da dependência entre observações e campos Markovianos

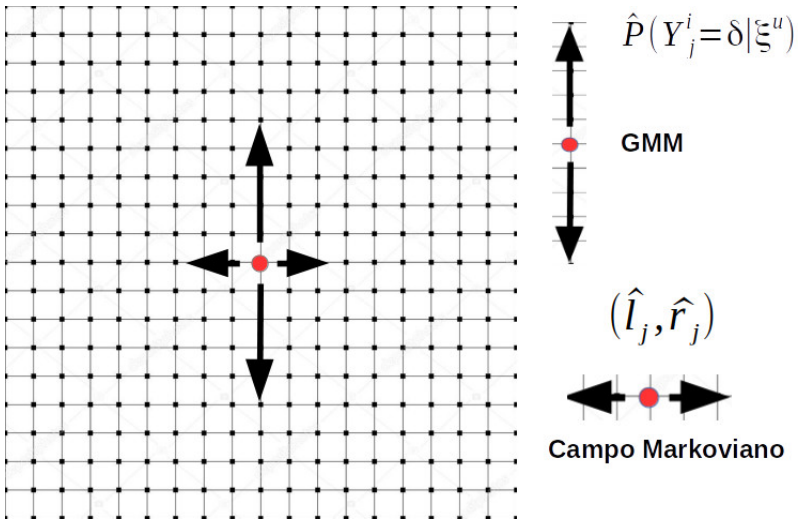



**Figura 2.8:** *Ilustração da solução combinada proposta: modelos lineares generalizados mistos modelando a dependência entre observações e campos Markovianos para inferência da dependência entre marcadores.*

modelam a dependência entre variáveis.

Comparando com a modelagem para indivíduos independentes, a necessidade de usar modelos mistos aparece porque, transformando a distribuição Trinomial em produtos de Binomiais, chegamos ao nível do indivíduo. O problema passa a ser, portanto, escrever um modelo linear generalizado misto, considerando o efeito aleatório de família, para uma resposta Binomial. A variável resposta (genótipo do SNP de cada indivíduo) tem distribuição Binomial com $n = 2$ e o objetivo é estimar o parâmetro $\pi_j^i$ dessa distribuição.

A modelagem GMM foi proposta por Breslow e Clayton, 1993 e representa uma extensão dos Modelos Lineares Generalizados, permitindo a inclusão de efeitos aleatórios no preditor linear, além dos efeitos fixos (Galdino, 2015). A introdução do efeito aleatório permite modelar a estrutura de correlação entre as observações intra-classe, que pertencem ao mesmo agrupamento, possibilitando a análise estatística de dados correlacionados. No modelo de efeito aleatório clássico, as variáveis respostas são condicionalmente independentes, dado o efeito aleatório, dentro do agrupamento. Ainda, sob a premissa de amostragem aleatória, variáveis resposta serão independentes, se pertencerem a diferentes níveis do fator aleatório.

Muitos estudos genéticos envolvendo dados de família, fazem uso do modelo linear misto poligênico com dois componentes de variância, um para a variabilidade devido à fatores genéticos ($\sigma_a^2$), outra devido à fatores ambientais ($\sigma_e^2$), como o trabalho de Blangero *et al.*, 2013, já citado na Seção 2.5.1. Tal modelo, na forma matricial, para a família $f$, é dado por:

$$Y_f = X_f\beta + Z_f\gamma_f + e_f,$$

tal que, $Y_f$ é o vetor de resposta de $n_f$ indivíduos, $n = \sum_f n_f$, $X_f$ é a matriz de delineamento associada a efeitos fixos definidos no vetor $\beta$, $Z_f$ é a matriz de delineamento associada ao vetor de efeitos aleatórios genéticos $\gamma_f$, com $\gamma_f \sim N(0; I_f\sigma_a^2)$, $e_f \sim N(0; I_f\sigma_e^2)$ e $I_f$ a matriz

identidade. Portanto:

$$Cov(Y_f) = \Omega_f = Z_f Z_f' \sigma_a^2 + I_f \sigma_e^2.$$

Na modelagem em estudos genéticos, a matriz de covariância $Z_f Z_f'$ é, em geral, dada por:

$$Z_f Z_f' = 2\Phi_f, \text{ levando a } \Omega_f = 2\Phi_f \sigma_a^2 + I_f \sigma_e^2.$$

Conforme antecipado na Seção 2.5.1, a matriz $2\Phi$ é a matriz de parentesco, a qual modela a correlação entre os indivíduos, digamos, indivíduos $I_1$ e $I_2$ da mesma família $f$, de tal forma que:

$$Corr(Y_{I_1 f}, Y_{I_2 f}) = \frac{\sigma_a^2 2\Phi_{I_1 I_2}}{\sigma_e^2 + \sigma_a^2}.$$

A construção da matriz de parentesco é feita a partir do "coeficiente de relacionamento", $k$ (em alusão ao inglês *kinship*) e do "coeficiente de relação", $r$, entre dois indivíduos ($I_1$ e $I_2$, por exemplo). O primeiro é definido como a probabilidade de que um alelo aleatório de um locus autossômico[7] do indivíduo $I_1$ seja idêntico por descendência[8], ao alelo aleatório no mesmo locus do indivíduo $I_2$.

O "coeficiente de relação", $r$, entre os indivíduos $I_1$ e $I_2$ é definido, então, como sendo a proporção esperada dos alelos idênticos por descendência, ou a correlação genética entre os indivíduos $I_1$ e $I_2$, representado por $r = 2k$ (Andrade e Pinheiro, 2002).

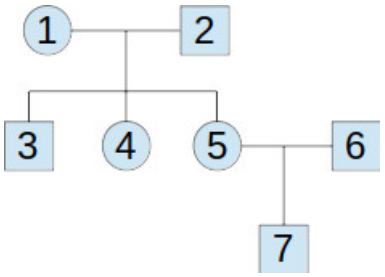
A Tabela 2.6 mostra coeficientes de relacionamento e de relação para alguns casos de parentesco. A Figura 2.9 ilustra o heredograma de uma família com sua respectiva matriz de parentesco.

**Tabela 2.6:** *Exemplos de parentesco e seus respectivos coeficientes de relacionamento (k) e de relação ($r = 2k$).*

| Parentesco | $k$ | $r = 2k$ |
|---|---|---|
| Gêmeos (univitelinos) | 1/2 | 1 |
| Pai/Mãe - filho | 1/4 | 1/2 |
| Irmãos | 1/4 | 1/2 |
| Tio(a) - sobrinho(a) | 1/8 | 1/4 |
| avós - netos | 1/8 | 1/4 |
| Primos (1º grau) | 1/16 | 1/8 |
| sem relacionamento | 0 | 0 |

[7] Autossomo é o cromossomo não sexual. A espécie humana possui 23 pares de cromossomos, sendo 22 autossômicos e 1 sexual.

[8] Definimos "idêntico por descedência" (do inglês, ibd) se dois indivíduos possuem dois alelos idênticos, derivados de uma replicação em um ancestral comum. Definimos "idênticos por estado" (do inglês, ibs) se dois indivíduos possuem dois alelos idênticos, sendo que não compartilham grau de parentesco, isto é, sem considerar a ancestralidade (Andrade e Pinheiro, 2002).




$$2\Phi = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{array} \begin{array}{c} \begin{array}{ccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{array} \\ \begin{bmatrix} 1 & 0 & ½ & ½ & ½ & 0 & ¼ \\ 0 & 1 & ½ & ½ & ½ & 0 & ¼ \\ ½ & ½ & 1 & ½ & ½ & 0 & ¼ \\ ½ & ½ & ½ & 1 & ½ & 0 & ¼ \\ ½ & ½ & ½ & ½ & 1 & 0 & ½ \\ 0 & 0 & 0 & 0 & 0 & 1 & ½ \\ ¼ & ¼ & ¼ & ¼ & ½ & ½ & 1 \end{bmatrix} \end{array}$$

**Figura 2.9:** *Exemplo de uma família hipotética e sua respectiva matriz de parentesco. Cada posição da matriz indica a correlação esperada entre os indivíduos.*

### 2.5.4 Influência familiar

Com a solução combinada proposta na Seção 2.5.3, torna-se possível estimar o parâmetro $\pi_j^{i,u}$ da distribuição Binomial (de cada marcador em cada indivíduo, dado um estado de vizinhança) usando a abordagem de modelos lineares generalizados mistos e a matriz de parentesco ($2\Phi$) para modelar a correlação entre indivíduos da amostra e, consequentemente, calcular a probabilidade da Binomial na expressão (2.11). Usando a mesma estratégia de campos Markovianos para dados independentes, a vizinhança de cada SNP pode ser encontrada através das estimativas dos parâmetros $(\hat{l}_j, \hat{r}_j)$ que maximizam a expressão (2.11). Finalmente, concatenando as vizinhanças encontradas para os SNP's contíguos, determinamos os blocos de dependência que procuramos.

Mais uma vez vale ressaltar que, no caso de amostras com dados de família, é esperado que o material genético compartilhado seja maior que em amostras de indivíduos não relacionados. Nosso interesse é conseguir identificar os blocos que refletem as porções genômicas que co-segregam na população geral, apesar de usar dados de família, em outras palavras, os blocos de dependência que estruturam o genoma sem considerar o efeito de família. Como avaliar o que está sendo desconsiderado na montagem de cada bloco devido à influência familiar?

Cada bloco de dependência será montado a partir da vizinhança de cada SNP, a qual, por sua vez, será estimada pelo campo Markoviano, encontrando o contexto necessário para determinar a distribuição de probabilidades condicional do SNP. O problema pode ser considerado como um problema de predição, no qual procuramos uma regra (contexto) capaz de prever o estado (valor do genótipo) de um SNP. Assim, mantendo a distribuição Binomial, de acordo com a modelagem imposta a partir do EHW, pode-se tomar diferentes alternativas para a estimação de suas probabilidades.

Para inferir então, a influência familiar em cada bloco de dependência, a estratégia proposta é comparar os resultados obtidos utilizando Modelos Lineares Generalizados (GLM - do inglês *Generalized Linear Models*) e GMM. Vamos postular mais detalhadamente o racional que está sendo adotado.

No campo Markoviano, o objetivo é encontrar o contexto (vizinhança) necessário para determinar a distribuição de probabilidades condicional de um determinado estado de interesse (no nosso caso, de um SNP). Sabemos que essa distribuição de probabilidade segue uma Binomial $(2, \pi_j^{i,u})$. Quando o parâmetro $\pi_j^{i,u}$ é estimado a partir do GMM, a dependência entre os indivíduos foi considerada no modelo através do efeito aleatório, portanto, uma parte da distribuição de probabilidade desse SNP já fica predita nesse componente e a estimativa (condicional) que é enviada ao campo Markoviano já está ajustada pela informação familiar. Assim o processamento dos dados no campo Markoviano encontra o contexto já ajustado para o efeito de família.

Ao contrário, usando o modelo GLM, estrategicamente, não se considera que há dependência na amostra, logo, as probabilidades são estimadas (possivelmente com vícios) em tamanhos amostrais não efetivos. Essas estimativas são, então, processadas no campo Markoviano, que precisará de um contexto maior para determinar a distribuição de probabilidades condicional do SNP de interesse. Esse comportamento foi comprovado nas simulações.

Pensando em termos dos estimadores de $\pi_j^{i,u}$, é sabido que, uma vez o modelo GLM desconsiderando a dependência existente entre as observações, este conduz a uma estimativa viesada do parâmetro. Liang e Zeger, 1986 afirmam, contudo, que mesmo nessas condições, o estimador ainda é consistente e razoavelmente eficiente, perdendo a eficiência à medida que a correlação aumenta. Logo, para dados de família, as estimativas pelos modelos GLM e GMM, não são esperadas serem iguais e o "desvio" depende da magnitude da correlação familiar.

O modelo GLM acaba incorporando na média (via função de ligação), o efeito da família e a estimativa viesada desse modelo é interpretada no campo Markoviano, aumentando o contexto necessário. O modelo GMM, por sua vez, é mais verossímil aos dados (de fato, isso se revelou nos resultados de simulação), fornecendo uma estimação consistente para o parâmetro da Binomial e, como consequência, o campo Markoviano estimará, coerentemente, um contexto menor.

A influência familiar está justamente no viés do estimador GLM, cuja magnitude se reflete na diferença de contexto necessária no campo Markoviano, quando comparamos, na solução combinada, os estimadores decorrentes dos modelos GLM e GMM.

Assim, é justamento o "erro" na estimação pelo modelo GLM que nos interessa. Não o erro absoluto, mas o erro relativo, na comparação com o modelo GMM. Em certo sentido, o modelo GLM acaba por funcionar como referência para indicar que, de fato, o modelo GMM consegue tratar o parentesco, mostrando qual seria o contexto necessário se o efeito de família não fosse considerado.

A influência familiar se reflete então na diferença de contexto para a determinação da distribuição de probabilidade condicional do genótipo de cada marcador e, consequentemente, na diferença dos blocos de dependência formados a partir de cada modelagem. Estamos interpretando essa diferença, como a parte da dependência que pode ser explicada pela influência familiar. A Figura 2.10 procura ilustrar, de forma intuitiva, a diferença entre as duas abordagens. Note que, é esperado que a árvore de contexto sob o modelo GMM seja podada em nós finais, mais próximos às folhas (definindo contextos menores), com

base na relevância do efeito de família.

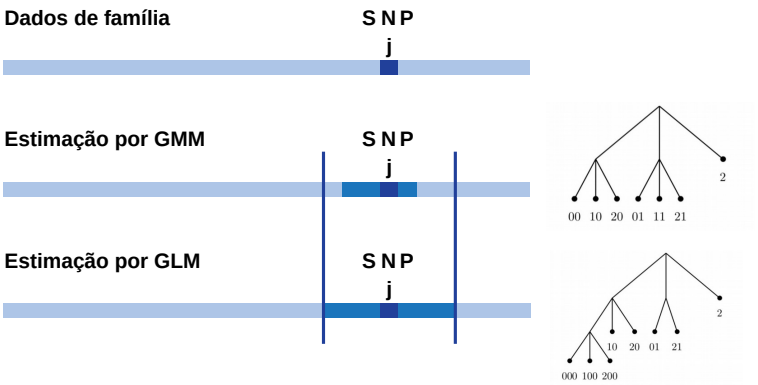



**Figura 2.10:** *Ilustração mostrando a diferença de contextos necessários para determinar a distribuição de probabilidade de um dado SNP, usando a modelagem GLM e GMM. As árvores de contexto são hipotéticas, apenas para representar a diferença entre os dois modelos.*

Isto é, assumimos que existe uma parte do contexto de cada SNP, cuja mensuração, em termos da sua distribuição de probabilidades condicional, deixa de ser necessária no campo Markoviano, quando usamos na Binomial o parâmetro $\pi_j^{i,u}$ estimado pelo GMM. Comparando os resultados dos dois ajustes, portanto, a diferença dos blocos obtidos com cada um dos modelos, pode ser atribuída à parte da dependência que pode ser explicada pela influência familiar (em destaque na Figura 2.11), uma vez que a diferença entre as abordagens é considerar o parentesco na alternativa GMM.

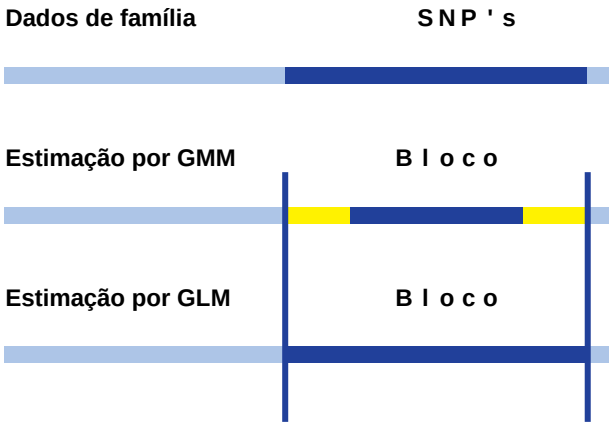



**Figura 2.11:** *Ilustração mostrando, no destaque em amarelo, a dependência que pode ser explicada pela influência familiar, na comparação entre os resultados obtidos usando os modelos GLM e GMM.*

### 2.5.5 Modelagem

#### O efeito da vizinhança

Conforme descrito em 2.5.2, a introdução do conceito de vizinhança impõe a estratificação da amostra em estados de vizinhança. A solução proposta implica em utilizar um modelo de campos Markovianos em que, para cada elemento da sequência, cada ponto do alfabeto $\{0, 1, 2\}$ está definido como um vetor de dados de família, ou seja, observações correlacionadas. Assim, para cada vizinhança considerada no campo Markoviano, temos um conjunto de realizações (contextos). A Figura 2.12 ilustra esse conceito.

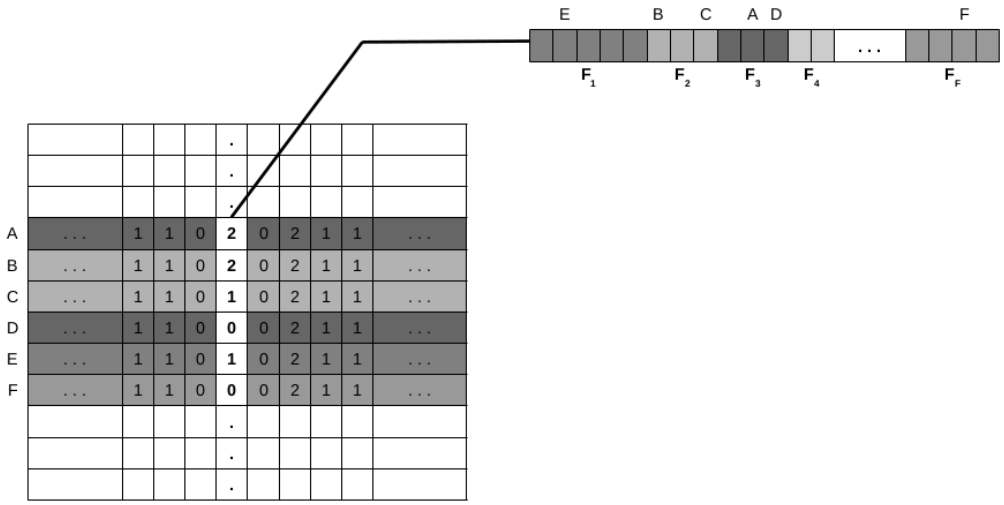



**Figura 2.12:** *Ilustração do modelo dentro de cada contexto específico no campo Markoviano (neste exemplo o contexto seria "110" antes do SNP e "0211" depois dele). As linhas $A, B, ..., F$ seriam os indivíduos e $F_1, F_2, ..., F_F$ as diferentes famílias.*

Assim, o estado de vizinhança, para uma vizinhança fixada, deve ser incluído como um fator dentro da modelagem. Lembrando a notação adotada $\xi^u$ para os estados de vizinhança, com $u \in \mathbb{N}$ sendo o número de diferentes estados, no caso da modelagem GLM (McCullagh, 2019), temos o seguinte modelo estrutural adotado para a função de ligação $g(\cdot)$, no nosso caso, a função logito:

$$g(\pi_j^{i,u} \mid \xi^u) = \alpha_j + X_j^{i,u}\beta_j^u, \tag{2.12}$$

em que, $\alpha_j$ é o intercepto (comum a todas as observações no SNP $j$), $\beta_j^u$ o vetor com dimensão ($u$ x 1) de parâmetros do efeito fixo de estado de vizinhança e $X_j^{i,u}$ corresponde à $i$-ésima linha da matriz $X_j^u$, com dimensão ($n$ x $u$), que é a matriz de planejamento para o efeito fixo do estado de vizinhança, constituída de "$0's$" e "$1's$" sendo que cada indivíduo tem 1 na casela respectiva ao seu estado de vizinhança[9].

Vale ressaltar que a modelagem em (2.12) está associada à variável ${Y^*}_j^i$ que representa a razão entre o número de sucessos e o número de tentativas, de tal forma que $Y_j^i = 2{Y^*}_j^i$, logo,

[9] Mais uma vez, para não sobrecarregar a notação, estamos adotando $u$ no lugar de $\xi^u$, nos parâmetros que dependem dos estados de vizinhança no SNP $j$.

$2Y^{*i}_{\ j} \sim Binomial(2, \pi^{i,u}_j)$. Essa codificação permite modelar o parâmetro $\pi^{i,u}_j$ diretamente do modelo ajustado, uma vez que, $E(Y^{*i}_{\ j}) = \pi^{i,u}_j$.

A modelagem GMM, por sua vez, é usada para estimar o parâmetro $\pi^{i,u}_j$ da distribuição Binomial do SNP $j$ no indivíduo $i$, em cada contexto, levando em consideração o agrupamento familiar, ou seja, usando a família como fator aleatório para modelar a estrutura de correlação. Resta interpretar, portanto, como a vizinhança deverá ser tratada nesse modelo.

Uma alternativa que poderia ser considerada é interpretar a vizinhança como um outro efeito aleatório do modelo, além daquele que cuida da correlação familiar. Neste caso, teríamos um modelo hierárquico (aninhado) de efeitos aleatórios, no qual o efeito de família seria tomado dentro de cada estado da vizinhança.

Essa proposta considera os estados da vizinhança, de cada SNP, como uma amostra aleatória retirada entre todos os estados possíveis. Em termos de modelos mistos, essa é uma suposição razoável, contudo, para efeito do cálculo dos campos Markovianos, introduziria uma mudança importante, incompatível com a definição de contexto, sob a qual cada diferente contexto (vizinhança), deve ser tomado separadamente. Em outras palavras deve-se calcular $\pi^{i,u}_j$ para cada vizinhança. A modelagem considerando a vizinhança como efeito aleatório, acabaria por estimar o valor de $\pi^{i,u}_j$ considerando todos os estados de vizinhança simultaneamente.

Isto posto, na solução combinada, a modelagem GMM precisa, assim como no caso GLM, considerar a vizinhança como efeito fixo no modelo:

$$g(\pi^{i,u}_j \mid Z, \xi^u) = \alpha_j + X^{i,u}_j \beta^u_j + Z^{i,u}_j a_{(i)j}, \tag{2.13}$$

sendo $Z^{i,u}_j$ a $i$-ésima linha da matriz $Z^u_j$, a matriz de delineamento associada ao efeito aleatório de família $a_{(i)j}$, tal que $Z^u_j (Z^u_j)' = 2\Phi^u$, é dada em função da matriz de parentesco que modela a estrutura de correlação entre os indivíduos que compartilham o mesmo estado de vizinhança. O intercepto, $\alpha_j$, a matriz $X^{i,u}_j$ e o vetor de parâmetros $\beta^u_j$, são definidos da mesma forma que no modelo GLM, expressão (2.12).

É importante notar que ambas as modelagens permitem facilmente a inclusão de covariáveis do indivíduo, como gênero, população de origem, etc.

### Matriz de parentesco ajustada

Para completar a solução proposta neste trabalho, ainda será preciso introduzir um (pequeno) ajuste na formulação do modelo GMM. Considerando a suposição de independência condicional dada pela estratificação da amostra em estados de vizinhança, como indicado na expressão (2.8) da Seção 2.5.2, a matriz de parentesco para cada estrato tem que ser reescrita, de tal forma que, indivíduos da mesma família, mas em estados diferentes, sejam tratados (condicionalmente) como não relacionados. A essa matriz chamaremos de matriz de parentesco ajustada. Sua montagem é operacionalmente simples, bastando forçar como zero (0) a correlação entre indivíduos que estão em estados diferentes, mesmo que pertencentes à uma mesma família.

O exemplo abaixo ilustra esse ajuste. Consideremos duas famílias (ver Figura 2.13),

ambas com pais e três filhos e ainda estados de vizinhança dados pela Tabela 2.7. Neste caso, estamos considerando um comprimento de vizinhança com três SNP's à esquerda e dois SNP's à direita.

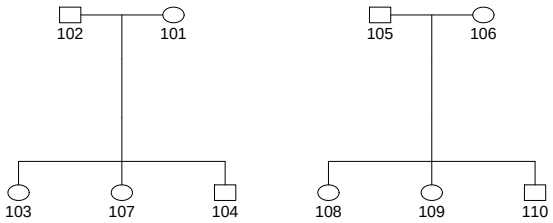



**Figura 2.13:** *Heredogramas de duas famílias hipotéticas com 3 filhos cada.*

**Tabela 2.7:** *Exemplo hipotético considerando os estados de vizinhança de uma amostra de 10 indivíduos e um comprimento de vizinhança de três SNP's à esquerda e dois SNP's à direita. A coluna $j$ identifica o SNP de interesse.*

| | $j-3$ | $j-2$ | $j-1$ | $j$ | $j+1$ | $j+2$ |
|---|---|---|---|---|---|---|
| 101 | 0 | 1 | 2 | 1 | 1 | 0 |
| 102 | 0 | 1 | 2 | 1 | 1 | 0 |
| 103 | 0 | 1 | 2 | 1 | 1 | 0 |
| 104 | 1 | 0 | 1 | 0 | 0 | 1 |
| 105 | 1 | 0 | 1 | 0 | 0 | 1 |
| 106 | 1 | 0 | 1 | 1 | 0 | 1 |
| 107 | 1 | 0 | 1 | 2 | 0 | 1 |
| 108 | 2 | 0 | 2 | 2 | 1 | 0 |
| 109 | 2 | 0 | 2 | 1 | 1 | 0 |
| 110 | 2 | 0 | 2 | 0 | 1 | 0 |

A matriz de parentesco e a matriz de parentesco ajustada para esse caso estão mostradas nas Tabelas 2.8 e 2.9, respectivamente, com destaque para as células que sofreram o ajuste (suposição de independência condicional).

**Tabela 2.8:** *Matriz de parentesco original, considerando as famílias da Figura 2.13.*

| | 101 | 102 | 103 | 104 | 105 | 106 | 107 | 108 | 109 | 110 |
|---|---|---|---|---|---|---|---|---|---|---|
| 101 | 1 | 0 | 1/2 | 1/2 | 0 | 0 | 1/2 | 0 | 0 | 0 |
| 102 | 0 | 1 | 1/2 | 1/2 | 0 | 0 | 1/2 | 0 | 0 | 0 |
| 103 | 1/2 | 1/2 | 1 | 1/4 | 0 | 0 | 1/4 | 0 | 0 | 0 |
| 104 | 1/2 | 1/2 | 1/4 | 1 | 0 | 0 | 1/4 | 0 | 0 | 0 |
| 105 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1/2 | 1/2 | 1/2 |
| 106 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1/2 | 1/2 | 1/2 |
| 107 | 1/2 | 1/2 | 1/4 | 1/4 | 0 | 0 | 1 | 0 | 0 | 0 |
| 108 | 0 | 0 | 0 | 0 | 1/2 | 1/2 | 0 | 1 | 1/4 | 1/4 |
| 109 | 0 | 0 | 0 | 0 | 1/2 | 1/2 | 0 | 1/4 | 1 | 1/4 |
| 110 | 0 | 0 | 0 | 0 | 1/2 | 1/2 | 0 | 1/4 | 1/4 | 1 |

**Tabela 2.9:** *Matriz de parentesco ajustada, considerando as famílias da Figura 2.13.*

| | 101 | 102 | 103 | 104 | 105 | 106 | 107 | 108 | 109 | 110 |
|---|---|---|---|---|---|---|---|---|---|---|
| 101 | 1 | 0 | 1/2 | **0** | 0 | 0 | **0** | 0 | 0 | 0 |
| 102 | 0 | 1 | 1/2 | **0** | 0 | 0 | **0** | 0 | 0 | 0 |
| 103 | 1/2 | 1/2 | 1 | **0** | 0 | 0 | **0** | 0 | 0 | 0 |
| 104 | **0** | **0** | **0** | 1 | 0 | 0 | 1/4 | 0 | 0 | 0 |
| 105 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | **0** | **0** | **0** |
| 106 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | **0** | **0** | **0** |
| 107 | **0** | **0** | **0** | 1/4 | 0 | 0 | 1 | 0 | 0 | 0 |
| 108 | 0 | 0 | 0 | 0 | **0** | **0** | 0 | 1 | 1/4 | 1/4 |
| 109 | 0 | 0 | 0 | 0 | **0** | **0** | 0 | 1/4 | 1 | 1/4 |
| 110 | 0 | 0 | 0 | 0 | **0** | **0** | 0 | 1/4 | 1/4 | 1 |

Chegamos então, por fim, ao modelo GMM, adotado para solução do problema proposto:

$$g(\pi_j^{i,u} \mid Z_A, \xi^u) = \alpha_j + X_j^{i,u}\beta_j^u + Z_{A_j}^{i,u} a_{(i)j}, \tag{2.14}$$

com os componentes do modelo definidos como anteriormente, na expressão (2.13) e sendo $Z_{A_j}^u$ tal que $Z_{A_j}^u (Z_{A_j}^u)' = 2\Phi_A^u$ é a matriz de parentesco que modela a estrutura de correlação entre os indivíduos, ajustada de tal forma que, indivíduos da mesma família, mas em estados de vizinhança diferentes, sejam tratados (condicionalmente) como não relacionados.

### Cálculo da probabilidade Binomial

A estimativa da probabilidade $P(Y_j^i = \delta \mid \xi^u)$ adotada na expressão empírica da função de pseudo-verossimilhança para o caso de dados de família (equação (2.11) apresentada na Seção 2.5.2), será obtida a partir da estimativa do parâmetro $\pi_j^{i,u}$ da distribuição Binomial, usando os modelos GLM e GMM apresentados em (2.12) e (2.14), respectivamente, ou seja, o modelo linear generalizado com efeito fixo de vizinhança e o modelo linear generalizado misto condicional para a média, com efeito fixo de vizinhança e efeito aleatório de família.

Considerando que $(Y_j^i \mid \xi^u)$ tem distribuição Binomial(2,$\pi_j^{i,u}$), $\hat{\pi}_j^{i,u}$ é o estimador de $\pi_j^{i,u}$, tal que:

$$\hat{\pi}_j^{i,u} = \frac{exp(\hat{g}(\pi_j^{i,u} \mid \xi^u))}{1 + exp(\hat{g}(\pi_j^{i,u} \mid \xi^u))},$$

com $g(\pi_j^{i,u} \mid \xi^u)$ definido sob a formulação GLM (expressão (2.12)) ou GMM (expressão (2.14)). Detalhes sobre os recursos utilizados para o ajuste dos modelos GLM e GMM, estão descritos na Seção 3.1.

Estando $l_j$ e $r_j$ fixados para cada loco $j$, definimos a extensão da vizinhança de estudo, seus diferentes estados e, consequentemente, os estratos da amostra. O cálculo das probabilidades será, portanto, sob EHW e para qualquer $i$ em $\xi^u$ no SNP $j$:

$$\begin{aligned}\hat{P}(Y_j^i = 0 \mid \xi^u) &= (1-\hat{\pi}_j^i)^2,\\ \hat{P}(Y_j^i = 1 \mid \xi^u) &= 2(1-\hat{\pi}_j^i)\hat{\pi}_j^i,\\ \hat{P}(Y_j^i = 2 \mid \xi^u) &= (\hat{\pi}_j^i)^2.\end{aligned}$$

Vale ressaltar que, para os casos em que a variância de $Y_j$ for nula para algum tamanho de vizinhança específico[10], a probabilidade da Binomial deve ser atribuída como:

$$\begin{aligned}\text{se } Y_j = 0 \Rightarrow \hat{P}(Y_j^i = 0 \mid \xi^u) &= 1,\\ \text{se } Y_j = 1 \Rightarrow \hat{P}(Y_j^i = 1 \mid \xi^u) &= 1,\\ \text{se } Y_j = 2 \Rightarrow \hat{P}(Y_j^i = 2 \mid \xi^u) &= 1.\end{aligned}$$

A estimativa de $l_j$ e $r_j$ para cada posição $j$ é feita maximizando a função de pseudo-verossimilhança empírica dada por (2.11). O procedimento de sobreposição das vizinhanças encontradas é então usado na construção dos blocos.

### 2.5.6 Índice de Influência Familiar

Reportando novamente à Figura 2.11, a diferença entre os blocos encontrados utilizando a modelagem GMM e GLM, indica a dependência genômica que pode ser explicada pela influência familiar.

Convém, portanto, quantificar, mesmo que descritivamente, essa diferença de modo a permitir análises comparativas entre diferentes porções do genoma, isto é, identificar quais seriam as regiões cromossômicas nas quais a influência familiar é mais importante, ou ainda, caracterizar o comportamento dessa influência ao longo de cada cromossomo.

Com este objetivo, propomos, descritivamente, um índice para quantificar a diferença entre a estrutura de blocos encontrada com cada um dos modelos.

Para melhor compreensão do raciocínio utilizado, tomemos como exemplo as estruturas hipotéticas de uma determinada região genômica, ilustradas na Figura 2.14. Quanto maior a similaridade entre os blocos encontrados sob GLM e GMM, menor deve ser a influência familiar nessa região. Assim, considerando o bloco formado pelos 3 primeiros SNP's, uma vez que as modelagens produziram o mesmo resultado, a influência familiar deveria ser nula. Por outro lado, no bloco formado pelos SNP's 17-18-19, essa influência deveria ser máxima, uma vez que toda a dependência encontrada do bloco GLM, não se repete quando consideramos o parentesco (modelagem GMM).

Intuitivamente, pode-se afirmar que, dado um bloco obtido pela modelagem GLM, quanto mais blocos GMM houver nessa região, maior a influência familiar. Por outro lado, quanto maior forem os blocos GMM dentro dessa mesma região, menor a influência

[10] Esta é uma situação rara que acontece quando temos valores faltantes nos marcadores vizinhos do SNP de interesse e esses indivíduos são justamente aqueles com valores diferentes para $Y_j$, em relação ao restante da amostra. Como consequência, nesse caso, a variância de $Y$ é nula, inviabilizando o cálculo do(s) modelo(s).


**Figura 2.14:** *Representação hipotética das estruturas de blocos de dependência obtidas usando a modelagem GLM (acima) e GMM (abaixo), em uma determinada região genômica. Os SNP's estão numerados e os blocos são indicados pela cor de fundo.*

familiar. Partindo desse raciocínio, pode-se definir dois índices auxiliares:

$$\text{Índice de fragmentação (IFrag)} = \frac{\text{\#blocos GMM}}{\text{\#SNP's}},$$

$$\text{Índice de cobertura (ICob)} = \frac{|\text{ maior bloco GMM }|}{\text{\#SNP's}},$$

em que, "#" representa "número de" e "$|\ x\ |$" o "tamanho de $x$".

Assim, o Índice de Influência Familiar (IIF) deverá ser diretamente proporcional ao IFrag e inversamente proporcional ao ICob.

$$IIF = \frac{IFrag}{ICob} = \frac{\frac{\text{\#blocos GMM}}{\text{\#SNP's}}}{\frac{|\text{maior bloco GMM}|}{\text{\#SNP's}}} = \frac{\text{\#blocos GMM}}{|\text{ maior bloco GMM }|}.$$

Dessa forma, no caso do bloco 1-2-3 da Figura 2.14, teríamos $IIF = 1/3$ e para o bloco 17-18-19, $IIF = 3/1 = 3$. Assim, dado um bloco GMM qualquer, $IIF$ tem limites dados por:

$$\frac{1}{\text{\#SNP's}} \leqslant IIF \leqslant \text{\#SNP's}.$$

Dividindo $IIF$ pelo #SNP's, conseguimos um limite superior unitário, mais conveniente para o nosso propósito:

$$IIF_{norm} = \frac{IIF}{\text{\#SNP's}} \Rightarrow \frac{1}{(\text{\#SNP's})^2} \leqslant IIF_{norm} \leqslant 1.$$

Com essa formulação, $IIF_{norm}$ é "pequeno" quando os blocos GLM e GMM são iguais, tendendo a 0 (desejável) apenas para blocos com muitos SNP's e igual a 1 quando todos os SNP's aparecem independentes na modelagem GMM.

Antes de ajustar os valores extremos, é preciso avaliar o caso especial em que temos apenas um bloco com dois SNP's e o restante dos SNP's independentes, como no bloco 11-12-13-14-15 da Figura 2.14. Neste caso a influência familiar deveria ser muito próxima de 1, de fato, a diferença dele para o caso em que o índice seria unitário, é apenas a presença do bloco duplo de SNP's. No entanto, se analisarmos o limite do $IIF_{norm}$ para esses casos, vemos que:

$$\lim_{\#\text{SNP's}\to\infty}(IIF_{norm}) = \frac{\frac{\#\text{SNP's}-1}{2}}{\#\text{SNP's}} = \frac{\#\text{SNP's} - 1}{2(\#\text{SNP's})} = \frac{1}{2}.$$

Sendo assim, não haveria nenhuma configuração de blocos que produziria um $IIF_{norm}$ no intervalo entre $1/2$ e $1$. Portanto, é mais razoável ajustar o limite superior do índice para $1/2$, evitando assim uma descontinuidade muito grande no índice.

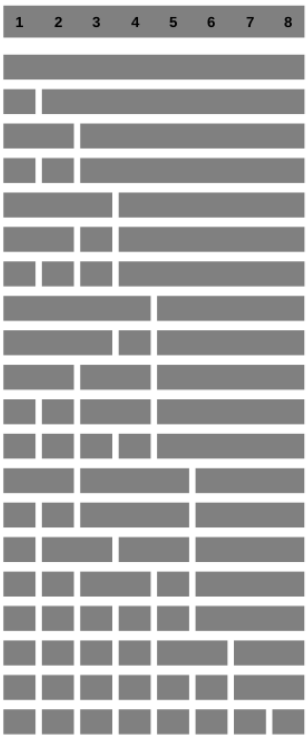
Para o caso dos blocos GLM e GMM serem idênticos, basta impor o índice igual a $0$. Chegamos assim à fórmula final do índice de influência familiar:

$$IIF_{ajustado} = IIF_A = \frac{\#\text{blocos GMM}}{|\text{ maior bloco GMM }|\,(\#\text{SNP's})},$$

considerando que:

$$\left.\begin{array}{ll} \text{quando os blocos GLM e GMM são idênticos} & \Rightarrow \quad IIF_A = 0 \\ \text{quando há apenas SNP's independentes no GMM} & \Rightarrow \quad IIF_A = 0,5 \end{array}\right\} 0 \leqslant IIF_A \leqslant 0,5.$$

Como exemplo, a Figura 2.15 mostra o cálculo do $IIF_A$ para todas as configurações de blocos GMM possíveis quando o bloco GLM possui 8 SNP's.




| # Blocos | Maior Blc | IIF(norm) | IIF_A |
|---|---|---|---|
| 1 | 8 | 0,016 | 0,000 |
| 2 | 7 | 0,036 | 0,036 |
| 2 | 6 | 0,042 | 0,042 |
| 3 | 6 | 0,063 | 0,063 |
| 2 | 5 | 0,050 | 0,050 |
| 3 | 5 | 0,075 | 0,075 |
| 4 | 5 | 0,100 | 0,100 |
| 2 | 4 | 0,063 | 0,063 |
| 3 | 4 | 0,094 | 0,094 |
| 3 | 4 | 0,094 | 0,094 |
| 4 | 4 | 0,125 | 0,125 |
| 5 | 4 | 0,156 | 0,156 |
| 3 | 3 | 0,125 | 0,125 |
| 4 | 3 | 0,167 | 0,167 |
| 4 | 3 | 0,167 | 0,167 |
| 5 | 3 | 0,208 | 0,208 |
| 6 | 3 | 0,250 | 0,250 |
| 6 | 2 | 0,375 | 0,375 |
| 7 | 2 | 0,438 | 0,438 |
| 8 | 1 | 1,000 | 0,500 |

**Figura 2.15:** *Diferentes estruturas de blocos possíveis para um conjunto de 8 SNP's, com os respectivos cálculos do índice de influência familiar.*

# Capítulo 3

# Estudos de Simulação

Neste capítulo, para ilustrar e exemplificar alguns casos e resultados específicos, utilizamos dados simulados pelo programa SimPed (Leal *et al.*, 2005). O SimPed é um programa de simulação de dados haplotípicos e genotípicos considerando estruturas familiares.

O programa permite construir bases de dados de marcadores, em equilíbrio ou desequilíbrio de ligação, o que, no nosso caso é bastante conveniente, uma vez que permite a montagem de blocos de dependência de SNP's, considerando a estrutura familiar, que também é fornecida como dado de entrada ao simulador. Em termos gerais o programa utiliza dois arquivos de entrada, um com a informação das estruturas familiares e outro com os parâmetros de formação dos marcadores. Neste último, quando desejamos gerar SNP's em desequilíbrio de ligação, definimos os haplótiplos desses marcadores e sua probabilidade de ocorrência; por outro lado, para os SNP's em equilíbrio de ligação, a informação é fornecida em termos da probabilidade alélica de cada um. O arquivo de saída fornece os marcadores simulados em codificação haplotípica, a qual convertemos em genotípica para utilização em nossas análises, uma vez que esse é o tipo de dados considerado neste trabalho.

## 3.1 Detalhes do processamento

Conforme descrito no Capítulo 2, a solução combinada proposta neste trabalho depende do estimador $\hat{\pi}_j^{i,u}$ da distribuição Binomial na função de pseudo-verossimilhança empírica. Essa estimativa será obtida de duas formas diferentes, uma utilizando a modelagem GLM e outra a modelagem GMM.

A implementação computacional da solução combinada proposta foi escrita em linguagem **R** (R Core Team, 2020) e o pseudo-código está disponibilizado no Apêndice A. Para ajuste dos modelos GLM utilizamos a função **glm** do pacote **lme4** (Bates *et al.*, 2015), enquanto que os modelos GMM foram ajustados com o uso da função **relmatGlmer** do pacote **lme4qtl** (Ziyatdinov *et al.*, 2018).

O pacote **lme4** é vastamente utilizado para tratar modelos lineares generalizados mistos. O pacote **lme4qtl** é uma extensão do anterior, proposta por Ziyatdinov, A. *et. al.* em 2018, que permite a definição dos efeitos aleatórios através de matrizes de covariância

customizadas. Essa flexibilização tornou possível a utilização das matrizes de parentesco dentro de cada modelo.

É importante ressaltar que, no caso do pacote **lme4qtl** e, consequentemente, na modelagem GMM, ao utilizarmos uma matriz Identidade como matriz de covariância, reduzimos o modelo ao caso de indivíduos independentes.

Em termos de números de modelos a serem ajustados na implementação de nossa metodologia, vale lembrar que, de acordo com a função de pseudo-verossimilhança empírica (2.11) desenvolvida para a solução combinada, a vizinhança de cada SNP será o par $(l_j, r_j)$ que a maximiza, para um dado tamanho de vizinhança. Em outras palavras, considerado um intervalo fixo de marcadores, antes e depois do SNP de interesse, deveremos avaliar todas as possíveis vizinhanças para encontrar qual é a que maximiza a expressão (2.11). A Tabela 3.1 mostra o número de casos para um prefixo de 2 marcadores e um sufixo de 2 marcadores (intervalo de 4 SNP's, sendo 2 antes e 2 depois). Nota-se que, neste exemplo, serão avaliados 9 modelos para determinar a vizinhança de cada SNP.

**Tabela 3.1:** *Indicação dos limites de todas as vizinhanças, considerando um intervalo de 2 marcadores para cada lado do SNP j de interesse.*

| Limite à esquerda ($l_j$) | Limite à direita ($r_j$) | Tamanho vizinhança |
|---|---|---|
| $j-2$ | $j+2$ | 4 marcadores |
| $j-2$ | $j+1$ | 3 marcadores |
| $j-2$ | $j$ | 2 marcadores |
| $j-1$ | $j+2$ | 3 marcadores |
| $j-1$ | $j+1$ | 2 marcadores |
| $j-1$ | $j$ | 1 marcador |
| $j$ | $j+2$ | 2 marcadores |
| $j$ | $j+1$ | 1 marcador |
| $j$ | $j$ | nenhum marcador (sem vizinhança) |

O intervalo de busca é avaliado considerando um determinado número de SNP's tanto antes quanto depois do SNP de interesse. De modo geral, chamando $W$ o tamanho desse intervalo, no caso do exemplo da Tabela 3.1 estaríamos considerando $W = 2$ e o número total de vizinhanças seria $(W+1)^2$.

Considerando um intervalo de SNP's que devem ser processados, a fórmula geral que fornece o número total de diferentes vizinhanças e, por consequência, o número de modelos que deverão ser ajustados, é dada por:

$$\#\text{modelos} = (W+1)^2 \left( \#\text{SNP's} - N \cdot \frac{W}{2} \right), \tag{3.1}$$

em que $N$ é o número de extremidades na amostra considerada. Por exemplo, caso tenhamos um cromossomo com 1000 marcadores e formos processar os 100 primeiros SNP's, teríamos $N = 1$ pois iniciamos em um extremo da amostra. A mesma coisa para o processamento dos últimos 100 marcadores. Caso o objetivo seja processar 100 marcadores de uma região

interna, teríamos $N = 0$ e, finalmente, para processar todos os SNP's, $N$ valeria 2. A evolução da fórmula dada está detalhada no Apêndice B.

Sendo assim, se considerarmos, por exemplo, um cromossomo com 50.000 marcadores e um intervalo de $W = 2$ antes e depois de cada SNP, seriam avaliados 449.982 modelos.

As análises dos dados simulados será dividida em duas etapas. Na primeira delas, Seção 3.2, o objetivo é mostrar como o estimador $\hat{\pi}_j^{i,u}$ se comporta nas duas modelagens, em algumas situações de interesse. Na segunda, Seção 3.3, iremos analisar a formação dos blocos de dependência a partir dos dois modelos.

## 3.2 Estimativas de $\pi_j^{i,u}$

Para compreender as diferenças de estimação para o parâmetro $\pi_j^i$ da distribuição Binomial em cada modelo, foram montados alguns cenários de dados simulados.

### 3.2.1 Cenário 1

A partir de um desenho de padrão para 27 marcadores, repetido duas vezes[1], foram gerados 54 marcadores, utilizando estruturas familiares diversas (84 famílias de vários tamanhos, variando entre 1 e 138 integrantes), totalizando 1676 indivíduos. Foram ajustados os modelos GLM e GMM considerando uma janela de dois SNP's para cada lado do SNP de interesse.

O padrão de dependência adotado neste conjunto de dados simulados é mostrado na Figura 3.1, como informação adicional, uma vez que, nesta etapa, o objetivo é apenas verificar como as estimativas do parâmetro $\pi_j^{i,u}$ se comportam nos diferentes modelos em diferentes cenários, logo, o padrão não interfere nas análises feitas aqui.

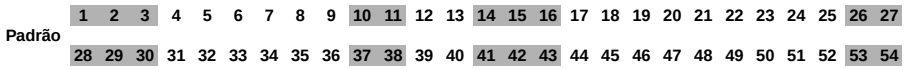



**Figura 3.1:** *Ilustração dos blocos de dependência do padrão imposto para geração dos dados simulados do Cenário 1. Os SNP's estão numerados e os blocos indicados através da cor de fundo (SNP's independentes estão sem fundo).*

Como ilustração, a Tabela 3.2 mostra resultados parciais para o décimo SNP, considerando a vizinhança de um SNP à esquerda e um SNP à direita. O estado de vizinhança ($\xi^u$) é igual a $2 \cdot 0$, ou seja, o SNP vizinho à esquerda tem valor 2 e o SNP vizinho à direita tem o valor 0 no entorno do SNP 10 (SNP de interesse).

Como esperado, o modelo GMM fornece vários valores estimados, uma vez que considera o parentesco no cálculo. Não é possível identificar um comportamento padrão

[1] O programa SimPed permite que um mesmo padrão para um conjunto de SNP's, seja repetido duas ou mais vezes, aumentando assim os marcadores simulados. Claramente, o padrão se repete mas os SNP's simulados são diferentes.

**Tabela 3.2:** *Resultados parciais obtidos para $\hat{\pi}_j^{i,u}$ nos modelos GLM e GMM, no Cenário 1 de dados simulados, para o SNP 10 e estado de vizinhança $\xi^u = 2 \cdot 0$.*

| | Indivíduo | Família | Pai | Mãe | Gênero | SNP 10 | $\xi^u$ | GLM | GMM |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 4918 | 4 | 0 | 0 | 1 | 0 | $2 \cdot 0$ | 0,83051 | 0,70152 |
| 2 | 76807 | 11 | 76505 | 76904 | 2 | 0 | $2 \cdot 0$ | 0,83051 | 0,65192 |
| 3 | 85924 | 36 | 0 | 0 | 2 | 0 | $2 \cdot 0$ | 0,83051 | 0,70152 |
| 4 | 5904 | 5 | 5919 | 5920 | 2 | 1 | $2 \cdot 0$ | 0,83051 | 0,80211 |
| 5 | 7501 | 7 | 7915 | 7916 | 1 | 1 | $2 \cdot 0$ | 0,83051 | 0,80792 |
| 6 | 11912 | 11 | 0 | 0 | 1 | 1 | $2 \cdot 0$ | 0,83051 | 0,80211 |
| 7 | 76803 | 11 | 76505 | 76904 | 1 | 1 | $2 \cdot 0$ | 0,83051 | 0,71603 |
| 8 | 76904 | 11 | 0 | 0 | 2 | 1 | $2 \cdot 0$ | 0,83051 | 0,71603 |
| 9 | 14101 | 14 | 14501 | 14904 | 2 | 1 | $2 \cdot 0$ | 0,83051 | 0,77656 |
| 10 | 14602 | 14 | 14501 | 14904 | 2 | 1 | $2 \cdot 0$ | 0,83051 | 0,77656 |
| 11 | 19919 | 19 | 0 | 0 | 1 | 1 | $2 \cdot 0$ | 0,83051 | 0,80759 |
| 12 | 107905 | 107 | 0 | 0 | 1 | 1 | $2 \cdot 0$ | 0,83051 | 0,80211 |
| 13 | 114502 | 114 | 114920 | 114921 | 2 | 1 | $2 \cdot 0$ | 0,83051 | 0,76690 |
| 14 | 114903 | 114 | 114920 | 114921 | 2 | 1 | $2 \cdot 0$ | 0,83051 | 0,76382 |
| 15 | 114908 | 114 | 114904 | 114903 | 1 | 1 | $2 \cdot 0$ | 0,83051 | 0,78256 |
| 16 | 114909 | 114 | 114904 | 114903 | 2 | 1 | $2 \cdot 0$ | 0,83051 | 0,78256 |
| 17 | 114921 | 114 | 0 | 0 | 2 | 1 | $2 \cdot 0$ | 0,83051 | 0,76690 |
| 18 | 7905 | 7 | 7918 | 7803 | 2 | 2 | $2 \cdot 0$ | 0,83051 | 0,86834 |
| 19 | 12201 | 12 | 0 | 0 | 2 | 2 | $2 \cdot 0$ | 0,83051 | 0,87843 |
| 20 | 16702 | 16 | 16601 | 16902 | 1 | 2 | $2 \cdot 0$ | 0,83051 | 0,89555 |
| 21 | 16703 | 16 | 16601 | 16902 | 1 | 2 | $2 \cdot 0$ | 0,83051 | 0,89555 |
| 22 | 16902 | 16 | 16916 | 16901 | 2 | 2 | $2 \cdot 0$ | 0,83051 | 0,90032 |
| 23 | 16909 | 16 | 16916 | 16901 | 2 | 2 | $2 \cdot 0$ | 0,83051 | 0,89414 |
| 24 | 16919 | 16 | 0 | 0 | 2 | 2 | $2 \cdot 0$ | 0,83051 | 0,88799 |
| 25 | 19910 | 19 | 0 | 0 | 2 | 2 | $2 \cdot 0$ | 0,83051 | 0,88613 |
| 26 | 19911 | 19 | 19912 | 19910 | 1 | 2 | $2 \cdot 0$ | 0,83051 | 0,87587 |
| 27 | 24202 | 24 | 24904 | 24901 | 2 | 2 | $2 \cdot 0$ | 0,83051 | 0,87843 |
| 28 | 27901 | 27 | 0 | 0 | 1 | 2 | $2 \cdot 0$ | 0,83051 | 0,87843 |
| 29 | 32608 | 32 | 32501 | 32502 | 2 | 2 | $2 \cdot 0$ | 0,83051 | 0,87843 |
| 30 | 43906 | 43 | 43801 | 43905 | 2 | 2 | $2 \cdot 0$ | 0,83051 | 0,87843 |

da estimativa em função do nível de parentesco. Temos, por exemplo, irmãos com a mesma estimativa (indivíduos 114908 e 114909) e irmãos com estimativas diferentes (indivíduos 114502 e 114903). Da mesma forma, os indivíduos 19919 e 107905 são ambos pais e apresentam estimativas diferentes, enquanto que os indivíduos 4918 e 85924, pai e mãe, respectivamente, acabam tendo o mesmo valor para o parâmetro estimado.

Como reflexo, vemos que a modelagem GMM confere uma estimativa $\hat{\pi}_j^i$ adaptada à estrutura de covariância imposta pela matriz de parentesco. No caso da modelagem GLM, como estamos no mesmo estado de vizinhança, o valor estimado para todos os indivíduos é igual. De fato, o modelo GLM estima um valor de $\hat{\pi}_j^i$ para cada estado diferente de vizinhança, uma vez que está definido como efeito fixo no modelo. A Tabela 3.3 mostra a estimativa $\hat{\pi}_j^i$ para cada estado diferente da vizinhança de 1 SNP à esquerda e 1 SNP à direita do SNP 10, considerando o modelo GLM. A Figura 3.2 ilustra a diferença entre o $\hat{\pi}_j^i$ estimado com cada um dos modelos, através dos seus histogramas respectivos.

Utilizando a Equação (3.1) para o conjunto de dados simulados, sabendo que a janela de busca considerada foi de dois SNP's para cada lado do SNP de interesse, chegamos a

**Tabela 3.3:** *Estimativas de $\pi_j^{i,u}$ para cada estado de vizinhança do SNP 10 no Cenário 1 de dados simulados.*

| $\xi^u$ | GLM |
|---|---|
| $0 \cdot 0$ | $0,71233$ |
| $0 \cdot 1$ | 0.66588 |
| $0 \cdot 2$ | 0.64096 |
| $1 \cdot 0$ | 0.75391 |
| $1 \cdot 1$ | 0.70948 |
| $1 \cdot 2$ | 0.69459 |
| $2 \cdot 0$ | 0.83051 |
| $2 \cdot 1$ | 0.76159 |
| $2 \cdot 2$ | 0.69536 |

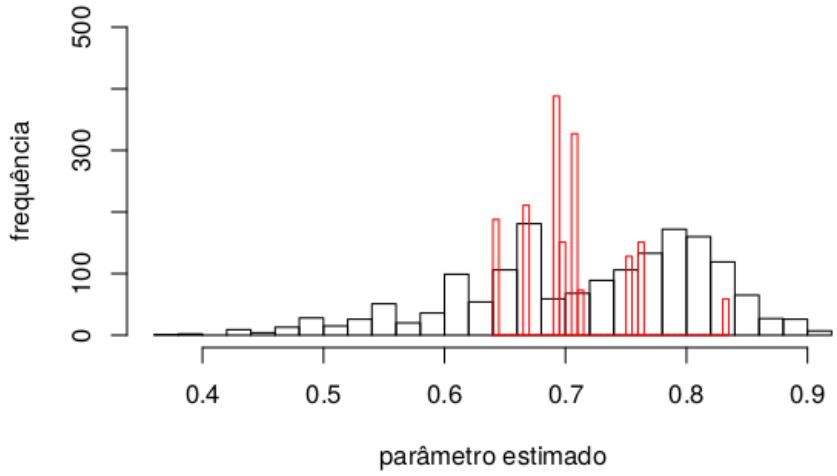



**Figura 3.2:** *Histograma comparativo entre as estimativas $\hat{\pi}_j^{i,u}$ obtidas sob cada um dos modelos, para o SNP 10 de interesse e vizinhança determinada por um SNP de cada lado, utilizando dados simulados do Cenário 1. O histograma em preto reflete as estimativas do modelo GMM enquanto que o histograma em vermelho, do modelo GLM.*

um total de 468 modelos a serem ajustados. Comparando o resultado da função de pseudo-verossimilhança empírica para cada modelo, notamos que em todos os casos o valor para o modelo GMM foi superior ao valor para o modelo GLM.

Assim, estes resultados simulados corroboram com a hipótese de que o modelo GMM é mais verossímil aos dados que o modelo GLM, conforme discutido na Subseção 2.5.4. Consequentemente, a distribuição Binomial fica melhor especificada com o parâmetro $\hat{\pi}_j^i$ estimado pelo modelo GMM, o que exige um contexto menor no campo Markoviano para determinação da distribuição de probabilidade de cada SNP de interesse.

A Tabela 3.4 ilustra o resultado do cálculo das funções de pseudo-verossimilhança dos modelos GLM e GMM, para alguns casos de SNP's de interesse e limites de vizinhança.

**Tabela 3.4:** *Cálculo dos valores das funções de pseudo-verossimilhança empírica para alguns SNP's de interesse e limites de vizinhança, com base nos dados simulados do Cenário 1.*

| | | | pseudo-verossimilhança empírica | |
|---|---|---|---|---|
| $l_j$ | SNP | $r_j$ | modelo GLM | modelo GMM |
| 1 | 1 | 3 | -951,831 | -916,828 |
| 1 | 1 | 2 | -959,780 | -863,645 |
| 1 | 1 | 1 | -1.399,477 | -904,050 |
| 13 | 15 | 17 | -1.148,841 | -1.141,484 |
| 13 | 15 | 16 | -1.158,503 | -1.099,631 |
| 13 | 15 | 15 | -1.212,208 | -1.062,495 |
| 14 | 15 | 17 | -1.165,678 | -1.093,280 |
| 14 | 15 | 16 | -1.168,229 | -1.057,429 |
| 14 | 15 | 15 | -1.212,537 | -1.034,092 |
| 15 | 15 | 17 | -1.414,348 | -1.104,049 |
| 15 | 15 | 16 | -1.418,673 | -1.083,715 |
| 15 | 15 | 15 | -1.564,858 | -1.054,065 |
| 28 | 30 | 32 | -1.404,329 | -1.268,957 |
| 28 | 30 | 31 | -1.418,128 | -1.184,250 |
| 28 | 30 | 30 | -1.427,957 | -1.160,734 |
| 29 | 30 | 32 | -1.418,900 | -1.229,787 |
| 29 | 30 | 31 | -1.428,247 | -1.152,848 |
| 29 | 30 | 30 | -1.435,019 | -1.128,060 |
| 30 | 30 | 32 | -1.526,699 | -1.175,679 |
| 30 | 30 | 31 | -1.529,981 | -1.090,763 |
| 30 | 30 | 30 | -1.530,211 | -1.067,018 |
| 43 | 45 | 47 | -1.545,978 | -1.251,571 |
| 43 | 45 | 46 | -1.557,192 | -1.235,440 |
| 43 | 45 | 45 | -1.568,552 | -1.170,984 |
| 44 | 45 | 47 | -1.558,964 | -1.187,531 |
| 44 | 45 | 46 | -1.563,998 | -1.171,924 |
| 44 | 45 | 45 | -1.571,366 | -1.103,064 |
| 45 | 45 | 47 | -1.568,106 | -1.173,915 |
| 45 | 45 | 46 | -1.570,697 | -1.156,283 |
| 45 | 45 | 45 | -1.573,172 | -1.086,556 |
| 52 | 54 | 54 | -1.465,838 | -1.048,121 |
| 53 | 54 | 54 | -1.468,174 | -990,518 |
| 54 | 54 | 54 | -1.471,550 | -922,495 |

### 3.2.2 Cenário 2

Conforme explicado no Capítulo 2, para execução do algoritmo da solução combinada proposta, o conceito de vizinhança é introduzido através da matriz de parentesco ajustada, a qual impõe que indivíduos da mesma família, porém em estados de vizinhança diferentes, sejam independentes. Assim:

$$Y_j^{i,f} \mid \xi^u \perp Y_j^{i',f'} \mid \xi^{u'} \text{ para } i \neq i' \text{ e } \xi^u \neq \xi^{u'} \ \forall \ f \text{ e } f'.$$

Para verificar essa implicação na estimativa $\hat{\pi}_j^i$, o mesmo conjunto anterior de dados simulados foi utilizado, contudo, com uma alteração: garantindo que indivíduos da mesma família tivessem sempre o mesmo estado de vizinhança. Dessa forma estamos unificando os dois agrupamentos, por família e por estado de vizinhança. Para isso, na implementação, apenas modificamos os SNP's vizinhos de um SNP de interesse, forçando os integrantes da família a pertencerem ao mesmo estado de vizinhança. É claro que esta estruturação dos dados viola pressuposições da simulação dos dados, contudo, é útil para o propósito específico de verificar como a estimativa do parâmetro da Binomial se comporta quando toda a família está no mesmo agrupamento de estado de vizinhança, comparativamente ao caso mais provável, qual seja, a família ter seus integrantes em dois ou mais estados.

A Tabela 3.5 mostra os resultados para o primeiro conjunto de dados simulados (conjunto original utilizado no Cenário 1) e para os dados gerados para o Cenário 2, com a vizinhança alterada, considerando o SNP 8 e uma vizinhança de dois SNP's antes e dois depois. A identificação de cada indivíduo, bem como seus progenitores também estão na Tabela.

Pelos dados, tendo em vista a modelagem GMM, pode-se notar que os irmãos 117911, 117907 e 117904 possuem estimativas $\hat{\pi}_j^i$ diferentes para o Cenário 1 e iguais para o Cenário 2. Os irmãos 117924, 117905 e 117906 têm estimativas diferentes em ambos os cenários. Os irmãos 117603 e 117605 têm estimativas iguais para o Cenário 1 e diferentes para o Cenário 2. Finalmente, os irmãos 117101 e 117602 têm as mesmas estimativas em ambos os cenários. Ainda nos mesmos quatro grupos de irmãos, analisando os estados de vizinhança no Cenário 1, vemos que há casos de mesmo estado de vizinhança e também de estados diferentes. Todos os casos têm o mesmo valor para o SNP de interesse, ou seja, mesmo valor para a variável resposta no modelo.

Deste modo, os resultados vão ao encontro do esperado, ou seja, o modelo misto com o efeito aleatório de família, acaba por estimar o valor do parâmetro de $\pi_j^{i,u}$ considerando o grau de parentesco dentro do agrupamento por estado de vizinhança. Um mesmo indivíduo, pode ter estimativas diferentes, dependendo dos parentes que compartilham o mesmo estado de vizinhança. Em outras palavras, o modelo misto "tira" o efeito da correlação (efeito de família) da estimativa da média e, quando faz isso, acaba "entregando" ao campo Markoviano uma distribuição Binomial mais verossímil aos dados e, consequentemente, o contexto necessário para determinar a distribuição do SNP de interesse é menor.

Vale observar também que, da mesma forma como relatado no item anterior, a modelagem GLM fornece uma estimativa de $\pi_j^{i,u}$ para cada nível de estado de vizinhança, uma vez que esse é fator fixo do modelo. No caso deste cenário, como o estado foi forçado a ser

**Tabela 3.5:** *Comparação entre as estimativas $\hat{\pi}_j^{i,u}$ em ambos modelos, nos Cenários 1 e 2, para o SNP 8, considerando uma vizinhança de dois SNP's antes e dois SNP's depois.*

| | | | | | | Cenário 1 | | | Cenário 2 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | SNP 8 | Indivíduo | Pai | Mãe | Gênero | $\xi^u$ | GLM | GMM | $\xi^u$ | GLM | GMM |
| 1 | 2 | 117911 | 117925 | 117924 | 1 | 00 · 10 | 1,00000 | 1,00000 | 21 · 00 | 0,94074 | 0,97685 |
| 2 | 2 | 117907 | 117925 | 117924 | 1 | 10 · 00 | 0,83333 | 0,90066 | 21 · 00 | 0,94074 | 0,97685 |
| 3 | 2 | 117904 | 117925 | 117924 | 2 | 10 · 01 | 0,92857 | 0,95535 | 21 · 00 | 0,94074 | 0,97685 |
| 4 | 2 | 117909 | 117905 | 117912 | 1 | 10 · 10 | 0,75000 | 0,84343 | 21 · 00 | 0,94074 | 0,95196 |
| 5 | 2 | 117901 | 0 | 0 | 2 | 10 · 11 | 0,96154 | 0,97660 | 21 · 00 | 0,94074 | 0,97285 |
| 6 | 2 | 117924 | 117902 | 117901 | 2 | 10 · 11 | 0,96154 | 0,97660 | 21 · 00 | 0,94074 | 0,97559 |
| 7 | 2 | 117905 | 117902 | 117901 | 1 | 10 · 20 | 1,00000 | 1,00000 | 21 · 00 | 0,94074 | 0,97514 |
| 8 | 2 | 117906 | 117902 | 117901 | 1 | 10 · 21 | 0,83333 | 0,89778 | 21 · 00 | 0,94074 | 0,96652 |
| 9 | 2 | 117910 | 0 | 0 | 1 | 11 · 01 | 0,97368 | 0,98363 | 21 · 00 | 0,94074 | 0,97544 |
| 10 | 1 | 117903 | 0 | 0 | 2 | 11 · 02 | 0,83333 | 0,78511 | 21 · 00 | 0,94074 | 0,85161 |
| 11 | 2 | 117925 | 0 | 0 | 1 | 11 · 11 | 0,89063 | 0,93246 | 21 · 00 | 0,94074 | 0,97613 |
| 12 | 1 | 117703 | 117601 | 117903 | 2 | 11 · 12 | 0,81395 | 0,76466 | 21 · 00 | 0,94074 | 0,73980 |
| 13 | 1 | 117701 | 117906 | 117607 | 1 | 11 · 21 | 0,72222 | 0,68235 | 21 · 00 | 0,94074 | 0,79999 |
| 14 | 2 | 117902 | 0 | 0 | 1 | 20 · 11 | 0,93284 | 0,95814 | 21 · 00 | 0,94074 | 0,97285 |
| 15 | 1 | 117912 | 0 | 0 | 2 | 21 · 10 | 0,91667 | 0,87516 | 21 · 00 | 0,94074 | 0,89338 |
| 16 | 1 | 117706 | 117923 | 117605 | 1 | 21 · 11 | 0,91724 | 0,88257 | 21 · 00 | 0,94074 | 0,75868 |
| 17 | 2 | 117908 | 117910 | 117709 | 1 | 21 · 11 | 0,91724 | 0,94008 | 21 · 00 | 0,94074 | 0,94755 |
| 18 | 1 | 117920 | 0 | 0 | 2 | 21 · 12 | 0,89844 | 0,87222 | 21 · 00 | 0,94074 | 0,73774 |
| 19 | 2 | 117603 | 117921 | 117920 | 1 | 21 · 12 | 0,89844 | 0,91862 | 21 · 00 | 0,94074 | 0,80226 |
| 20 | 2 | 117605 | 117921 | 117920 | 2 | 21 · 12 | 0,89844 | 0,91862 | 21 · 00 | 0,94074 | 0,79029 |
| 21 | 1 | 117923 | 0 | 0 | 1 | 21 · 21 | 0,90769 | 0,86685 | 21 · 00 | 0,94074 | 0,87241 |
| 22 | 1 | 117607 | 117921 | 117920 | 2 | 21 · 22 | 0,93182 | 0,90219 | 21 · 00 | 0,94074 | 0,62731 |
| 23 | 2 | 117601 | 117921 | 117920 | 1 | 21 · 22 | 0,93182 | 0,93908 | 21 · 00 | 0,94074 | 0,77397 |
| 24 | 0 | 117101 | 117921 | 117920 | 2 | 22 · 12 | 0,86250 | 0,56355 | 21 · 00 | 0,94074 | 0,51849 |
| 25 | 0 | 117602 | 117921 | 117920 | 1 | 22 · 12 | 0,86250 | 0,56355 | 21 · 00 | 0,94074 | 0,51849 |
| 26 | 1 | 117921 | 0 | 0 | 1 | 22 · 12 | 0,86250 | 0,66711 | 21 · 00 | 0,94074 | 0,73774 |
| 27 | 2 | 117709 | 117923 | 117605 | 2 | 22 · 12 | 0,86250 | 0,86155 | 21 · 00 | 0,94074 | 0,87121 |

igual, a estimativa pelo modelo GLM, é única.

### 3.2.3 Cenário 3

Um novo conjunto de dados simulados foi gerado utilizando o mesmo padrão de SNP's dos Cenários 1 e 2 (Figura 3.1), contudo, considerando indivíduos sem relação de parentesco, ou seja independentes. Para isso, foi suficiente alterar o arquivo de entrada correspondente, alterando a informação de pai e mãe para 0.

Novamente foram ajustados os modelos GLM e GMM para obter a estimativa de $\pi_j^{i,u}$, neste caso, utilizando a matriz Identidade como a matriz de parentesco nos modelos GMM. O objetivo neste caso foi o de avaliar o comportamento da modelagem GMM para indivíduos independentes, sendo esperado os mesmos resultados da modelagem GLM, uma vez que nesta situação, os dois modelos são equivalentes.

Ajustamos um total de 468 modelos para cada modelagem e as diferenças entre as estimativas $\hat{\pi}_j^i$ ficaram sempre extremamente baixas, como mostrado na Tabela 3.6 resumo, cujas diferenças resultaram apenas de aproximação devido aos métodos numéricos diferentes utilizados em cada pacote.

Esses resultados estão de acordo com o esperado, isto é, de que sob o modelo GMM, ao utilizarmos uma matriz Identidade como matriz de covariância, reduzimos o modelo ao caso de indivíduos independentes. Apenas a título de exemplo, os resultados das estimativas para alguns modelos ajustados considerando o 11° marcador como SNP de interesse e uma vizinhança de um SNP à esquerda e dois SNP's à direita, estão mostrados na

**Tabela 3.6:** *Dados resumo para a diferença de estimativas de $\pi_j^{i,u}$ utilizando os modelos GLM e GMM, considerando dados de indivíduos sem parentesco.*

| mínimo | 1º quartil | média | 3º quartil | máximo |
|---|---|---|---|---|
| -5,2 e-07 | -1,18 e-07 | 9,79 e-09 | 9,91 e-08 | 1,78 e-06 |

Tabela 3.7.

**Tabela 3.7:** *Resultados dos primeiros registros obtidos para o Cenário 3 de dados simulados, considerando o 11º marcador como SNP de interesse e uma vizinhança de um SNP à esquerda e dois SNP's à direita.*

| | ID | V11 | $\xi^u$ | GLM | GMM | Diferença |
|---|---|---|---|---|---|---|
| 1 | 1101 | 1 | 2 · 12 | 0,60714 | 0,60714 | 9,91E-08 |
| 2 | 1201 | 2 | 2 · 02 | 0,62278 | 0,62278 | -1,18E-07 |
| 3 | 1901 | 2 | 1 · 12 | 0,72273 | 0,72273 | 2,24E-07 |
| 4 | 2101 | 2 | 2 · 12 | 0,60714 | 0,60714 | 9,91E-08 |
| 5 | 2201 | 1 | 1 · 12 | 0,72273 | 0,72273 | 2,24E-07 |
| 6 | 2202 | 2 | 0 · 12 | 0,75410 | 0,75410 | 9,95E-08 |
| 7 | 2302 | 2 | 2 · 12 | 0,60714 | 0,60714 | 9,91E-08 |
| 8 | 2303 | 1 | 0 · 12 | 0,75410 | 0,75410 | 9,95E-08 |
| 9 | 2903 | 1 | 1 · 02 | 0,68931 | 0,68931 | 5,52E-08 |
| 10 | 4101 | 2 | 1 · 02 | 0,68931 | 0,68931 | 5,52E-08 |
| 11 | 4201 | 2 | 2 · 02 | 0,62278 | 0,62278 | -1,18E-07 |
| 12 | 4301 | 2 | 1 · 02 | 0,68931 | 0,68931 | 5,52E-08 |
| 13 | 4302 | 2 | 1 · 12 | 0,72273 | 0,72273 | 2,24E-07 |
| 14 | 4303 | 0 | 1 · 12 | 0,72273 | 0,72273 | 2,24E-07 |
| 15 | 4501 | 2 | 1 · 02 | 0,68931 | 0,68931 | 5,52E-08 |
| 16 | 4502 | 0 | 0 · 02 | 0,67647 | 0,67647 | -1,57E-07 |
| 17 | 4506 | 1 | 2 · 22 | 0,58824 | 0,58824 | -5,20E-07 |
| 18 | 4601 | 2 | 1 · 12 | 0,72273 | 0,72273 | 2,24E-07 |
| 19 | 4801 | 1 | 1 · 01 | 0,64706 | 0,64706 | 1,28E-09 |
| 20 | 4802 | 0 | 0 · 11 | 0,70000 | 0,70000 | 7,10E-07 |
| 21 | 4803 | 0 | 1 · 02 | 0,68931 | 0,68931 | 5,52E-08 |
| 22 | 4806 | 2 | 2 · 12 | 0,60714 | 0,60714 | 9,91E-08 |
| 23 | 4901 | 1 | 2 · 12 | 0,60714 | 0,60714 | 9,91E-08 |
| 24 | 4902 | 2 | 2 · 02 | 0,62278 | 0,62278 | -1,18E-07 |
| 25 | 4903 | 0 | 0 · 12 | 0,75410 | 0,75410 | 9,95E-08 |
| 26 | 4904 | 2 | 2 · 12 | 0,60714 | 0,60714 | 9,91E-08 |
| 27 | 4905 | 1 | 2 · 12 | 0,60714 | 0,60714 | 9,91E-08 |
| 28 | 4906 | 1 | 2 · 11 | 0,62766 | 0,62766 | -3,19E-07 |
| 29 | 4907 | 1 | 1 · 12 | 0,72273 | 0,72273 | 2,24E-07 |
| 30 | 4908 | 2 | 0 · 02 | 0,67647 | 0,67647 | -1,57E-07 |

### 3.2.4 Modelos singulares

No ajuste de alguns modelos GMM, é possível receber uma mensagem de aviso de que o modelo ajustado é quase singular. Isto significa, que os parâmetros estão na borda do espaço paramétrico, ou seja, variâncias de uma ou mais combinações lineares dos efeitos aleatórios são próximas a zero (Ziyatdinov *et al.*, 2018). Nessas situações, o valor da estimativa $\hat{\pi}_j^{i,j}$ pelo modelo GMM verificou-se ser igual ao valor estimado para o modelo GLM, para cada estado de vizinhança.

O que ocorre nesses casos, portanto, é que os dados são insuficientes para explicar a complexidade da estrutura de correlação imposta no modelo GMM, ou seja, o efeito aleatório de família não tem significância e o modelo GMM acaba se reduzindo ao modelo GLM. De fato, a variância do efeito aleatório do modelo GMM nestes casos, é sempre nula ou adotada como empiricamente nula[2].

Os modelos em que ocorreram essa situação no Cenário 1 de dados simulados, estão descritas na Tabela 3.8.

**Tabela 3.8:** *Situações no Cenário 1 de dados simulados, em que os dados são insuficientes para explicar a estrutura de covariância do modelo GMM. A coluna $\sigma_a^2$ indica a variância do efeito aleatório.*

| $l_j$ | SNP | $r_j$ | $\sigma_a^2$ |
|---|---|---|---|
| 1 | 2 | 3 | 0 |
| 1 | 2 | 4 | 1,15 e-07 |
| 12 | 14 | 16 | 0 |
| 26 | 28 | 30 | 0 |
| 27 | 29 | 30 | 2,09 e-07 |
| 27 | 29 | 31 | 1,11 e-05 |
| 39 | 41 | 43 | 0 |
| 40 | 41 | 43 | 1,11 e-06 |
| 40 | 42 | 43 | 0 |
| 40 | 42 | 44 | 2,28 e-07 |
| 41 | 42 | 44 | 0 |

O efeito prático disso para o nosso problema é que a estrutura familiar não consegue acrescentar nada para a estimativa de $\pi_j^{i,j}$, naquele SNP, naquele limite de vizinhança e, nesse caso, é totalmente coerente utilizar a modelagem GLM no lugar da GMM. Como os modelos são equivalentes nessas situações, não há nenhuma correção a ser feita.

## 3.3 Blocos de dependência

A Seção anterior foi importante para compreender as diferenças entre as duas modelagens consideradas, GLM e GMM, em diferentes situações, em particular, o impacto nas estimativas do parâmetro da distribuição Binomial dentro da função de pseudo-verossimilhança empírica. A partir de agora vamos analisar os resultados da solução combinada completa, isto é, o resultado final na determinação da estrutura de dependência do genoma.

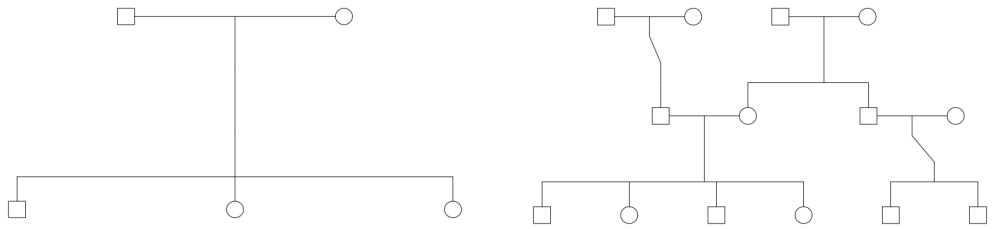
Considerando a distribuição Binomial caracterizada por duas tentativas independentes e o parâmetro $\pi_j^{i,j}$ estimado pelos modelos GLM ou GMM, o campo Markoviano vai determinar qual a vizinhança, ou seja, quantos SNP's à esquerda e à direita do SNP de interesse, maximizam a função de pseudo-verossimilhança empírica definida em (2.11). Concatenando as vizinhanças de cada SNP, conforme descrito na Seção 2.1.2, encontramos os blocos de dependência.

[2] Nos casos dos modelos não singulares, a variância do efeito aleatório foi sempre maior que zero.

Assim, mesmo que seja mencionado no texto daqui em diante, por abuso de linguagem, "os resultados do modelo GLM ou GMM", o que deve ser entendido é que "os blocos de dependência obtidos, depois da concatenação da vizinhança dos SNP's, quando essa vizinhança foi estimada usando a distribuição Binomial formulada sob cada um dos modelos, GLM e GMM".

Para avaliarmos a determinação dos blocos de dependência por nossa solução combinada, foram simulados dados utilizando alguns padrões, usando estruturas familiares diferentes e, para alguns casos, foram geradas também 5 réplicas de amostras. As réplicas são possíveis indicando ao programa SimPed diferentes sementes para a geração de números (pseudo) aleatórios. Esclarecemos que, o elevado tempo computacional exigido para ajustar todos os modelos envolvidos na solução proposta, inviabiliza a execução de um grande número de réplicas e uso de um número maior de marcadores, que seriam necessárias para um estudo de simulação eficiente.

Além da estrutura familiar utilizada nas simulações da Seção anterior, com 84 famílias de vários tamanhos, variando entre 1 e 138 integrantes, totalizando 1676 indivíduos, foram usadas duas estruturas simplificadas: uma com famílias de 5 indivíduos, sendo 2 pais e 3 filhos, a qual chamamos de "nuclear" e outra com 3 gerações e 14 membros (Wang e Thompson, 2019), identificada pela sigla "3GP". Foram geradas 335 famílias nucleares, totalizando 1675 indivíduos, e 120 famílias de 3 gerações, totalizando 1680 indivíduos. As estruturas familiares citadas estão mostradas na Figura 3.3.



**(a)** *Estrutura das famílias nucleares* **(b)** *Estrutura das famílias com 3 gerações*

**Figura 3.3:** *Estruturas familiares utilizadas nas simulações. À esquerda a estrutura da família nuclear e à direita a família com 3 gerações.*

É importante esclarecer que os dados são simulados considerando a dependência entre os SNP's, bem como o relacionamento familiar, ao mesmo tempo. Uma vez os dados tendo sido gerados, não é possível identificar o quanto da dependência é devida ao desequilíbrio de ligação e quanto é devida ao parentesco. Uma alternativa para contornar esse problema, através dos parâmetros de entrada do programa SimPed, é discutida no último cenário, que também é aquele que permite uma melhor avaliação dos resultados dos blocos de dependência e da diferença entre os dois modelos.

### 3.3.1 Cenário 4

Neste cenário de simulação, o padrão de geração dos dados utilizado consistiu de um conjunto de 30 marcadores, todos independentes. Foram utilizadas as estruturas familiares

nuclear e de 3 gerações (3GP), em 5 réplicas para cada uma. Os resultados estão mostrados na Figura 3.4.

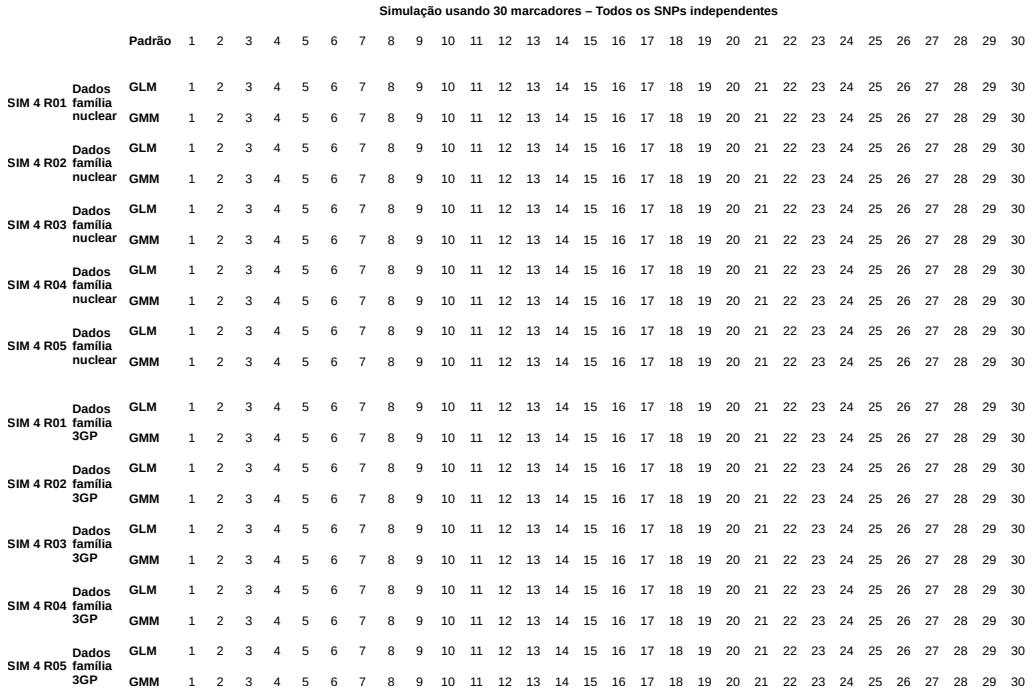



**Figura 3.4:** *Estrutura de dependência obtida para o conjunto de dados simulados com 30 marcadores, todos independentes - Cenário 4.*

O objetivo central desta simulação, era garantir que os modelos não indicariam dependência onde ela não existisse. Pelo que se nota dos resultados, a solução combinada com ambos os modelos, conseguiu capturar perfeitamente o padrão imposto na simulação, para as duas estruturas familiares em todas as réplicas, não identificando nenhum tipo de dependência entre os marcadores.

### 3.3.2 Cenário 5

Neste estudo de simulação, o primeiro padrão de dados foi gerado definindo um grupo de 15 marcadores, repetido duas vezes, totalizando 30 SNP's. O objetivo principal neste caso foi o de verificar a recuperação dos blocos de dependência em cada um dos modelos, tanto com dados familiares como considerando indivíduos independentes. A estrutura familiar foi a mesma utilizada na Seção anterior, totalizando 1676 indivíduos divididos em 84 famílias, em cada uma de 5 réplicas.

O padrão de dados definido neste estudo e os resultados obtidos estão mostrados na Figura 3.5.

Analisando a Figura 3.5, o primeiro ponto importante a verificar é que a modelagem GLM consegue capturar perfeitamente o padrão de dependência imposto na simulação dos dados, tanto para os dados gerados com estrutura familiar como considerando indivíduos independentes.

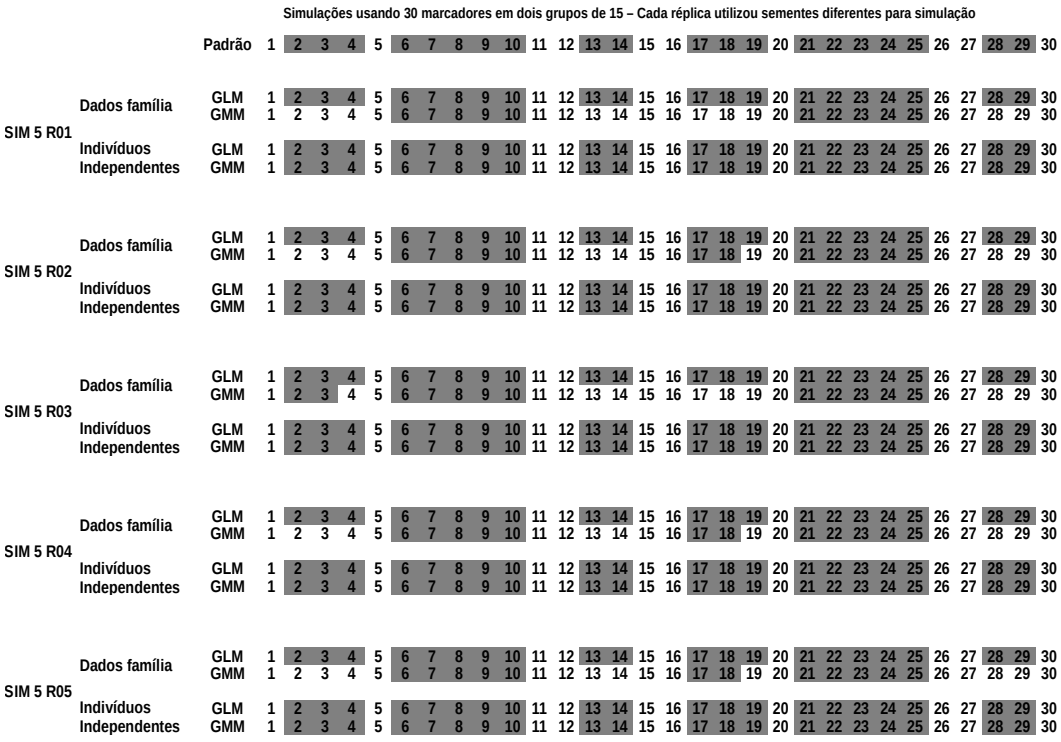



**Figura 3.5:** *Resultados do Cenário 5 de simulação, contendo 30 marcadores em dois grupos de 15. Os blocos estão identificados pela cor de fundo cinza, SNP's independentes aparecem sem cor de fundo. Para cada réplica os dados foram simulados considerando estrutura familiar e também como indivíduos independentes.*

A modelagem GMM, quando os dados são gerados para indivíduos independentes, produz o resultado exatamente igual à modelagem GLM, o que está de acordo com o esperado, tendo em vista que os modelos, GLM e GMM, são equivalentes nessa situação. Esse resultado da simulação é muito importante pois mostra que, na ausência de relacionamento entre os indivíduos, a solução combinada proposta recupera exatamente a estrutura de dependência dos marcadores.

A diferença entre as duas modelagens aparece quando os dados são gerados com estrutura familiar. É importante lembrar que, nestes casos, os dados são simulados considerando tanto a dependência do padrão de SNP's definido, como a relação de parentesco, simultaneamente. Não é possível, a partir dos dados gerados na simulação, identificar o quanto da dependência em cada bloco é devido ao efeito de família. De fato, a solução com a modelagem GLM continua capaz de recuperar os blocos de dependência do padrão definido, mas é incapaz de separar o quanto dessa dependência foi produzida pela relação familiar. A modelagem GMM, por outro lado, recupera blocos diferentes de dependência em alguns casos. A interpretação é que, nestes casos, a relação de dependência entre os dados gerados foi provocada, em parte, pela relação de parentesco informada e, identificada quando utilizamos o modelo GMM.

Assim, a dependência do bloco de SNP's 13-14, por exemplo, pode ser explicada pela relação de parentesco entre os indivíduos da amostra, em todas as 5 réplicas consideradas. O bloco de SNP's 2-3-4, por outro lado, ainda tem parte da sua dependência na réplica 3,

entre os SNP's 2 e 3, que não pode ser explicada pela estrutura familiar dos dados. Situação idêntica acontece com o bloco de SNP's 17-18-19 que, na segunda, quarta e quinta réplica, mostra os SNP's 17 e 18 ainda com dependência. A diferença dos resultados entre as réplicas não deve ser tomada como uma inconsistência, ao contrário, seriam reflexo da influência do parentesco nos dados simulados, que não é a mesma nas diferentes réplicas.

Os blocos maiores, 6-7-8-9-10 e 21-22-23-24-25 foram recuperados integralmente pelas duas modelagens. Uma discussão mais aprofundada sobre esses casos será discutida no próximo cenário.

### 3.3.3 Cenário 6

Avaliando o resultado do cenário anterior, notamos que parece haver uma tendência da modelagem GMM em não recuperar blocos pequenos (leia-se 2 SPN's), mas recuperar blocos maiores, no conjunto de dados. Para avaliar essa hipótese, foi gerado um novo padrão de dados simulados, com 50 marcadores, desta vez forçando blocos menores e maiores de SNP's.

Neste cenário, a forma adotada foi alterar os parâmetros possíveis de entrada do simulador de dados de forma a forçar uma dependência provocada pelo parentesco nos blocos maiores e diminui-la (ou eliminá-la) nos blocos menores. Assim, a modelagem GMM deveria produzir o mesmo resultado que a GLM em blocos pequenos de dois ou três SNP's, enquanto que blocos maiores apareceriam apenas na modelagem GLM. A alternativa encontrada para montar esse cenário de simulação foi alterar a probabilidade de cada um dos haplótipos, para cada bloco de dependência do padrão simulado.

Lembrando o racional utilizado para elaborar o Índice de Influência Familiar ($IIF_A$), apresentado na Seção 2.5.6, quanto mais iguais forem os blocos obtidos pela modelagem GLM e GMM, em uma determinada porção genômica, menor o $IIF_A$, isto é, menor a influência familiar nessa região. Por outro lado, quanto mais distintas as estruturas obtidas pelos dois modelos, maior a influência familiar. A Figura 3.6 mostra de forma esquemática esse raciocínio, para dois casos extremos.

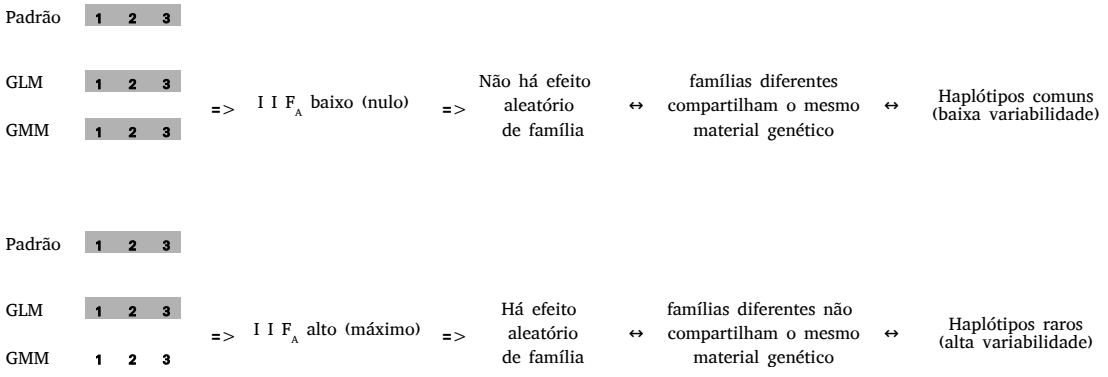



**Figura 3.6:** *Possível relação entre o resultado da modelagem da solução proposta e a frequência dos haplótipos em uma determinada região genômica. As setas bidirecionais são apenas indicação e não implicação.*

Sob o ponto de vista biológico, quanto mais frequente (comum) for um determinado

haplótipo, mais famílias conterão esse material genético, ou seja, a interpretação é que essa porção genômica segrega na população de modo geral, independentemente da agregação familiar. Haplótipos comuns na população geral, portanto, são compartilhados, com igual probabilidade, por indivíduos parentes ou sem parentesco. Haplótipos raros na população, ao contrário, são, com maior probabilidade, compartilhados entre parentes mas não entre indivíduos não relacionados. Assim, haplótipos comuns devem ter baixa variabilidade na população em geral, enquanto que haplótipos raros ou que têm alta variabilidade, devem ser específicos às famílias.

Em termos da modelagem estatística aqui proposta, o modelo GMM consegue adicionar informação importante para a estimativa do parâmetro da Binomial, quando a dependência tem origem familiar, ou seja, quando o efeito aleatório de família é significante. Quando isso acontece, dizemos que há efeito de família no modelo proposto. Assim, quando diferentes famílias compartilham o mesmo haplótipo (comum), o efeito do parentesco se torna não significante e, consequentemente, o modelo GMM tem um resultado próximo ou igual à modelagem GLM.

Se, ao contrário, um determinado haplótipo tiver várias categorias de resposta (raro), o efeito de família é mais importante, uma vez que o efeito do parentesco pode se refletir no desvio em relação à média geral. No contexto biológico, a presença de vários haplótipos diferentes na população, indicam que, dependendo da família, um haplótipo pode ser mais frequente que outro. Importante notar que o que chamamos aqui de manifestação de um haplótipo é, em alguma medida, a representação de um determinado estado de vizinhança dentro do modelo considerado.

As Figuras 3.7 e 3.8 mostram as simulações processadas para as estruturas familiares nuclear e três gerações, respectivamente. Para cada uma delas foram feitas duas configurações de distribuição dos haplótipos. Em cada Figura estão mostrados os padrões da simulação e os resultados obtidos em cada modelagem. Abaixo de cada teste, os haplótipos introduzidos no simulador, com suas respectivas probabilidades.

Assim, alterando a variabilidade dos haplótipos, é possível interferir na forma como a influência familiar se reflete nos dados simulados e, consequentemente, modificar o resultado, em particular da modelagem GMM, segundo o raciocínio exposto acima.

Avaliando, por exemplo, a simulação com estrutura de família nuclear na Figura 3.7, o bloco 2-3-4 não aparece na modelagem GMM no primeiro teste, contudo, diminuindo a variabilidade dos haplótipos, o resultado do modelo GMM se iguala ao GLM, ou seja, interpretamos que não há mais influência familiar. No bloco 36 a 43, que é recuperado parcialmente no primeiro teste, ao permitirmos uma maior variabilidade dos haplótipos, o efeito aleatório de família passa a ser significante e, consequentemente, a modelagem GMM mostra que a dependência entre os SNP's pode, agora, ser explicada pelo efeito de família.

Alterando a distribuição dos haplótipos, acabamos por interferir também no equilíbrio de ligação, uma vez que a dependência entre os marcadores é dada justamente pela informação da frequência dos haplótipos no arquivo de entrada do simulador. Por esse motivo, alguns blocos de SNP's acabam não sendo reconhecidos nem mesmo pela modelagem GLM, como por exemplo os blocos 11-12, 17-18 e 44-45 no primeiro teste, ainda na Figura 3.7.

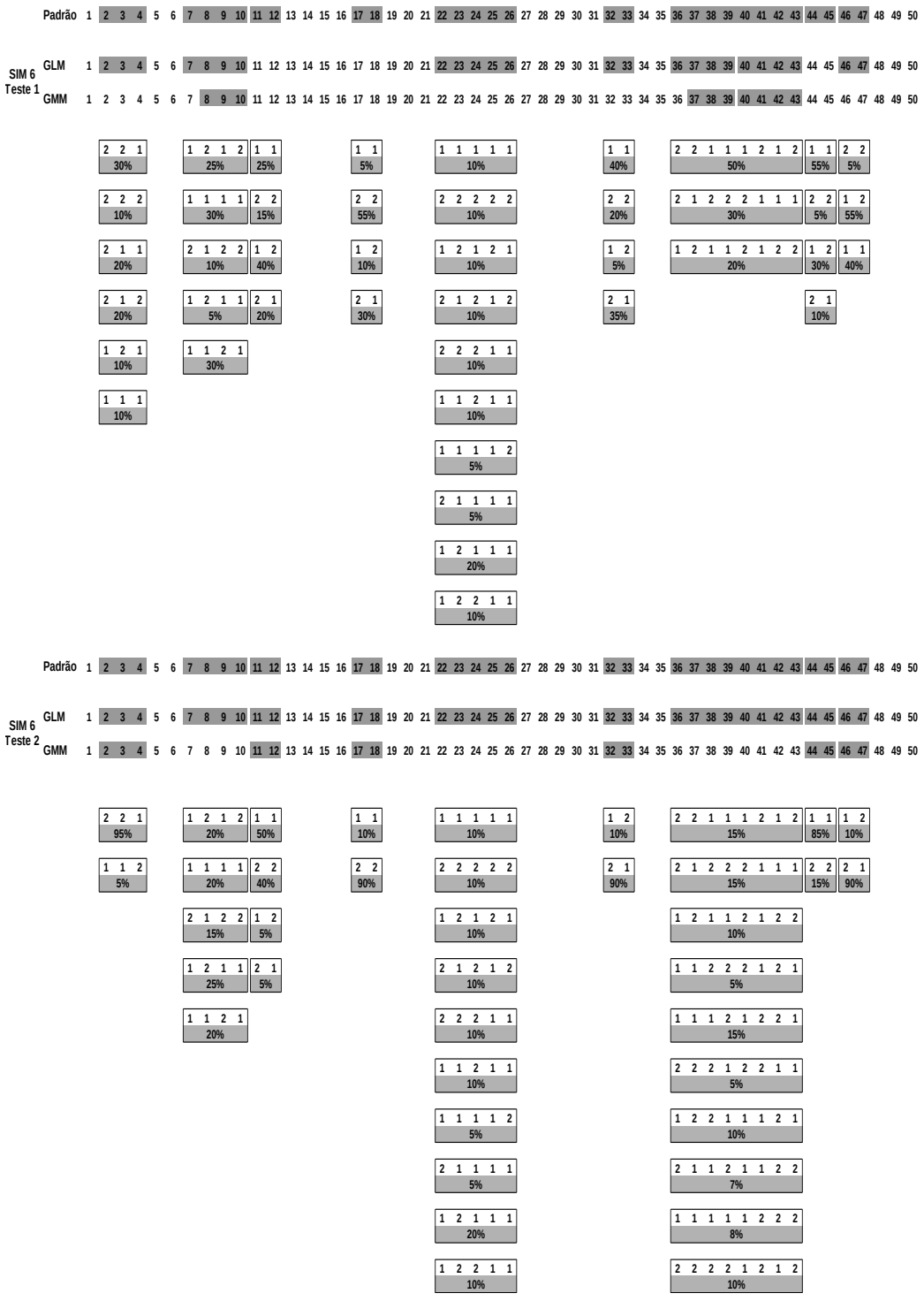



**Figura 3.7:** *Padrão de dependência gerado para 50 marcadores, alterando a distribuição de frequência dos haplótipos, para estrutura familiar nuclear. São mostrados duas simulações, para cada uma delas, a distribuição dos haplótipos está indicada abaixo do resultado.*

Ao ajustarmos as frequências dos haplótipos, no entanto, o resultado do segundo teste correspondeu ao esperado.

Utilizando a estrutura familiar de três gerações (Figura 3.8), o bloco 2-3-4 não é recuperado pela modelagem GMM, mesmo concentrando os haplótipos. Por outro lado, os blocos 17-18, 32-33, 44-45 e 46-47, passam a aparecer em ambas as modelagens quando as categorias de haplótipos são reduzidas, isto é, tornando os haplótipos comuns, indicando que agora a dependência entre os SNP's não pode mais ser explicada pela influência familiar.

Outro ponto importante a considerar, baseado nos resultados simulados apresentados, é que quanto menor o bloco, maior deve ser a concentração dos seus haplótipos, ou seja, mais comum deve ser na população geral, para que o efeito de família não apareça.

Estes resultados revelados por simulação são bastante relevantes, uma vez que refletem o raciocínio desenvolvido no Capítulo 2, qual seja, a modelagem GMM consegue tratar a dependência de origem familiar no conjunto de dados dos marcadores, retirando o efeito do parentesco da estimativa da média (desvio em relação à média). Assim, o resultado obtido para a estrutura de dependência utilizando a solução combinada de campo Markoviano, com o modelo misto utilizando a família como efeito aleatório, consegue refletir as porções genômicas que co-segregam em geral na população, mesmo utilizando indivíduos com parentesco.

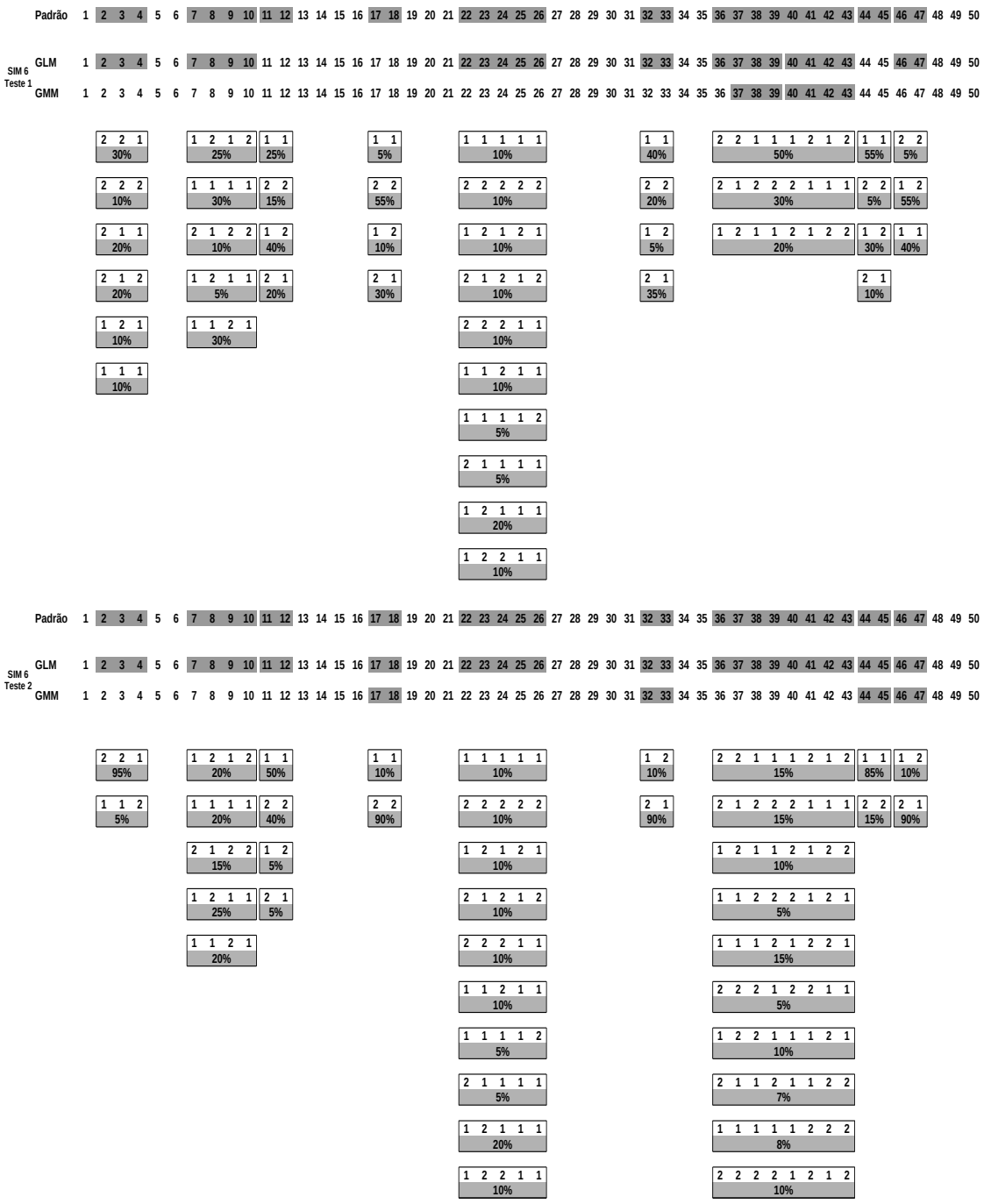



**Figura 3.8:** *Padrão de dependência gerado para 50 marcadores, alterando a distribuição de frequência dos haplótipos, para estrutura familiar com três gerações. São mostrados duas simulações, para cada uma delas, a distribuição dos haplótipos está indicada abaixo do resultado.*

# Capítulo 4

# Aplicação

Neste capítulo mostramos alguns resultados obtidos utilizando a solução proposta em um conjunto de dados reais de famílias, amostradas da população brasileira.

## 4.1 Descrição dos Dados

A base de dados é composta por 2.318 indivíduos, pertencentes a 93 famílias (variando entre 1 e 203 integrantes, com mediana de 17 membros), amostradas na cidade de Baependi, localizada ao sul de Minas Gerais (ver Figura 4.1). Os dados genéticos foram coletados utilizando a plataforma Affymetrix 6.0, totalizando 809.228 marcadores genotipados, em equilíbrio de Hardy-Weinberg, distribuídos ao longo dos 22 cromossomos autossômicos, conforme indicado na Tabela 4.1. Para busca da vizinhança de cada SNP, foi considerado um intervalo de 2 SNP's à esquerda e 2 SNP's à direita do SNP de interesse.

**Figura 4.1:** *Mapa indicando a localização geográfica da cidade de Baependi, na qual foram amostrados os dados processados neste trabalho.*

### 4.1.1 Detalhes do processamento

No processamento dos dados dos genótipos das famílias de Baependi via nossa solução combinada, ocorreram casos de não convergência do algoritmo no ajuste do modelo GLM (identificados por mensagens de aviso na saída dos ajustes). Essas situações aconteceram quando o valor do SNP de referência não variava dentro de cada estado de vizinhança.

**Tabela 4.1:** *Número de marcadores genotipados em cada cromossomo.*

| Cromossomo | Número de marcadores | Cromossomo | Número de marcadores |
|---|---|---|---|
| 1 | 66251 | 12 | 39728 |
| 2 | 68585 | 13 | 31938 |
| 3 | 56867 | 14 | 26066 |
| 4 | 52124 | 15 | 24294 |
| 5 | 52613 | 16 | 25708 |
| 6 | 52360 | 17 | 19364 |
| 7 | 43956 | 18 | 24662 |
| 8 | 45503 | 19 | 11017 |
| 9 | 38487 | 20 | 21229 |
| 10 | 44745 | 21 | 11795 |
| 11 | 41320 | 22 | 10616 |

Como exemplo de um desses casos, a Tabela 4.2 mostra os dados para o SNP 35674 no cromossomo 1.

**Tabela 4.2:** *Dados para o SNP 35674 do cromossomo 1 na amostra de dados das famílias de Baependi.*

| Estado de vizinhança $\xi^u$ | Valor do SNP 35674 |
|---|---|
| 00 | sempre 0 |
| 01 | sempre 1 |
| 02 | sempre 2 |
| 10 | sempre 0 |
| 11 | sempre 1 |
| 20 | sempre 0 |

Após avaliação detalhada da distribuição dos estados na vizinhança desses SNP's com problemas, as estimativas de $\pi_j^i$ obtidas, apesar da não convergência, estavam de acordo com o esperado, para cada estado de vizinhança, isto é, a menos de arredondamento, eram iguais a 0, $\frac{1}{2}$ e 1, quando o genótipo do SNP era 0, 1 ou 2, respectivamente. Portanto, as soluções nesses casos foram aceitas para a solução proposta.

Em relação aos modelos GMM, também ocorreram alguns casos em que não foi possível o ajuste do modelo, seja por não convergência ou por problemas no cálculo numérico[1]. Nossa pesquisa indicou que há algumas razões para esse tipo de problema acontecer mas, em geral, deve-se a um super ajuste do modelo. Nestes casos, os modelos GMM não foram considerados, assim como os modelos GLM respectivos (para o mesmo SNP e tamanho de vizinhança). A Tabela 4.3 mostra, para cada cromossomo processado, o número de modelos avaliados (que pode ser verificado a partir da fórmula (3.1)) e o número de modelos retirados por conta da não convergência.

Em termos dos valores da função de pseudo-verossimilhança, ocorreram alguns casos em que o valor para o modelo GLM foi superior àquele calculado para o modelo GMM,

[1] Nessas situações a mensagem de erro indicava que a matriz de $V^tV$ não era positiva definida.

**Tabela 4.3:** *Modelos avaliados e considerados para cada cromossomo processado.*

| Cromossomo | # SNP's | # modelos | # modelos retirados | # modelos considerados |
|---|---|---|---|---|
| 1 | 66251 | 596241 | 18091 (3,03%) | 578150 |
| 2 | 68585 | 617247 | 15566 (2,52%) | 601681 |
| 3 | 56867 | 511785 | 12523 (2,45%) | 499262 |
| 4 | 52124 | 469098 | 12255 (2,61%) | 456843 |
| 5 | 52613 | 473499 | 12254 (2,59%) | 461245 |
| 6 | 52360 | 471222 | 13088 (2,78%) | 458134 |

o que está em desacordo com o esperado, uma vez que o modelo GMM, considerando o parentesco entre os indivíduos, seria sempre mais verossímil aos dados.

Analisando esses casos, contudo, a diferença, apesar de ser a favor da modelagem GLM, era extremamente pequena, decorrente, na verdade, da aproximação numérica de cada algoritmo. Os valores das estimativas de $\pi_j^i$ foram as mesmas, a menos de um delta, nos dois modelos nessas situações, ou seja, são exemplos de casos em que o modelo GMM é equivalente ao GLM, conforme discutido no Capítulo 3.

Na Tabela 4.4 temos um exemplo da diferença entre as estimativas de $\pi_j^i$ obtidas com cada um dos modelos, mostrando a sua ordem de grandeza. A Tabela 4.5 mostra o número de casos em cada cromossomo processado, bem como o valor máximo da diferença entre as funções de pseudo-verossimilhança.

**Tabela 4.4:** *Exemplo da diferença entre as estimativas de $\pi_j^{i,u}$ obtidas em cada um dos modelos, para um caso em que a pseudo-verossimilhança empírica do modelo GLM foi superior à do modelo GMM. O valores referem-se ao SNP 5 do cromossomo 1, com vizinhança de dois SNP's à esquerda e nenhum à direita.*

| $\xi^u$ | $\pi_j^{i,u}$ (GLM) | $\pi_j^{i,u}$ (GMM) | Diferença (delta) |
|---|---|---|---|
| 00 | 0,010980 | 0,010980 | -1,06060E-12 |
| 01 | 0,035714 | 0,035714 | -4,99600E-15 |
| 10 | 0,451005 | 0,451005 | 1,404432E-14 |
| 11 | 0,384615 | 0,384615 | 6,400436E-14 |
| 20 | 0,841837 | 0,841837 | 9,703349E-14 |

**Tabela 4.5:** *Número de casos (em cada cromossomo processado) em que a diferença entre o valor da pseudo-verossimilhança empírica do modelo GLM e do modelo GMM foi positiva e o valor máximo dessa diferença.*

| Cromossomo | Modelos com diferença positiva | maior diferença |
|---|---|---|
| 1 | 7213 | 4,252E-11 |
| 2 | 8092 | 3,482E-9 |
| 3 | 6838 | 5,173E-11 |
| 4 | 6428 | 4,303E-9 |
| 5 | 6197 | 1,999E-9 |
| 6 | 5760 | 2,948E-9 |

Depois do processamento e montagem dos blocos de dependência utilizando as duas modelagens, foi efetuado o cálculo do Índice de Influência Familiar (IIF) descrito na Seção 2.5.6 do Capítulo 2. Foram observados alguns valores que fugiam dos limites considerados na definição do índice, no caso, valores iguais a zero ou valores entre 1/2 e 1.

Conforme discutido na Seção 2.5.6, antes do ajuste dos limites inferior e superior, o IIF deveria respeitar os seguintes intervalos:

$$0 < IIF < 0,5 \quad \text{ou} \quad IIF = 1.$$

Uma verificação desses casos mostrou a existência de alguns blocos fora do padrão esperado pelo raciocínio proposto para a influência familiar (Seção 2.5.4). A Figura 4.2 ilustra um desses exemplos e facilita a compreensão desses casos.

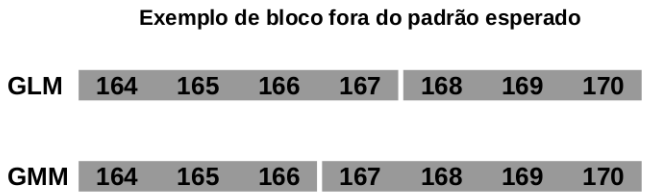



**Figura 4.2:** *Exemplo de caso de bloco fora do padrão esperado na comparação dos resultados entre as modelagens GLM e GMM. A região em destaque é entre os SNP's 164 e 170 no cromossomo 1.*

Como se nota na Figura 4.2, um bloco obtido pela modelagem GMM acabou por ultrapassar o limite do bloco respectivo a partir da modelagem GLM, nessa região genômica específica (entre os SNP's 164 e 170 do cromossomo 1). No caso deste exemplo, o SNP 167 que pertence a um bloco como resultado do modelo GLM, acaba por pertencer a outro bloco quando o modelo usado é o GMM.

Essas ocorrências são devidas a uma diferença pontual na vizinhança estimada de um ou mais SNP's da região. No caso do exemplo em questão, as vizinhanças estimadas de cada SNP, considerando cada um dos modelos, está descrita na Tabela 4.6. A Figura 4.3 ilustra a formação dos blocos baseado nas vizinhanças estimadas de cada modelo.

**Tabela 4.6:** *Vizinhanças estimadas para os SNP's 164 a 170 no cromossomo 1, para cada um dos modelos.*

| | Modelo GLM | | Modelo GMM | |
|---|---|---|---|---|
| SNP | $l_j$ | $r_j$ | $l_j$ | $r_j$ |
| 164 | 164 | 166 | 164 | 166 |
| 165 | 164 | 166 | 164 | 166 |
| 166 | 164 | 167 | 166 | 166 |
| 167 | 166 | 167 | 167 | 168 |
| 168 | 168 | 169 | 168 | 169 |
| 169 | 168 | 170 | 168 | 170 |
| 170 | 169 | 170 | 169 | 170 |

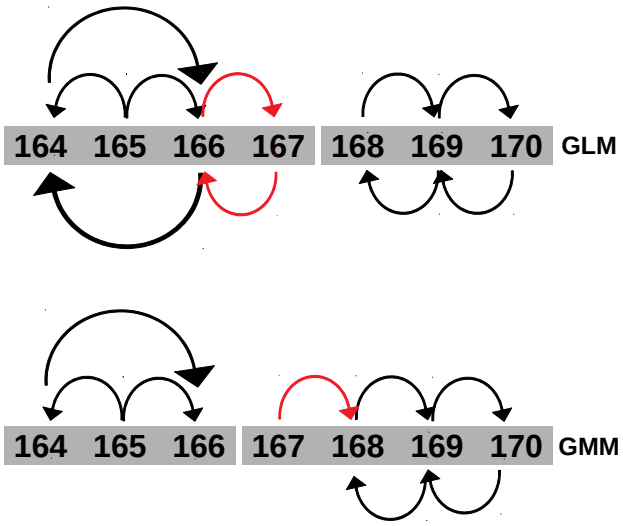



**Figura 4.3:** *Montagem dos blocos dos SNP's 164 a 170 no cromossomo 1. As setas indicam as vizinhanças estimadas para cada marcador, em cada modelo.*

A setas destacadas em vermelho na Figura 4.3 indicam as vizinhanças que provocaram a diferença na montagem dos blocos.

Esse comportamento empírico não esperado (considerando o que foi discutido na Seção 2.5.4), ocorreu em um número restritos de casos dentre o total de ajustes considerados. A Tabela 4.7 mostra o número de casos de blocos fora do padrão, identificados para cada cromossomo, acompanhados do percentual em relação ao número de blocos totais na modelagem GLM. Como se observa, o número dessas ocorrências é relativamente pequeno.

**Tabela 4.7:** *Número de blocos considerados fora do padrão esperado, em cada cromossomo processado e sua quantidade relativa aos blocos encontrados com a modelagem GLM.*

| Cromossomo | # Blocos GLM | # Blocos fora do padrão (%) |
|---|---|---|
| 1 | 11186 | 530 (4,74%) |
| 2 | 11261 | 566 (5,03%) |
| 3 | 9307 | 491 (5,28%) |
| 4 | 8642 | 477 (5,52%) |
| 5 | 8685 | 471 (5,42%) |
| 6 | 8548 | 407 (4,76%) |

Tendo em vista os resultados obtidos com os dados disponíveis nessa aplicação, blocos fora do padrão esperado ocorrem em situações muito específicas, contudo, ao nosso ver, não invalidam a solução. Finalmente, para efeito do cálculo do Índice de Influência Familiar (IIF), tais blocos não foram considerados.

## 4.2 Comparação dos cromossomos processados

Selecionamos os cromossomos com mais de 50.000 marcadores amostrados, assim, foram processados os primeiros 6 cromossomos que acabam sendo os maiores também em termos de tamanho molecular. Em particular, do ponto de vista aplicado, um objetivo adicional de nosso trabalho é avaliar os resultados na região HLA (do inglês *Human Leukocyte Antigens*), uma região específica do nosso genoma, localizada no cromossomo 6, responsável por nossa resposta imunológica. Os resultados dessa região serão tratados em uma subseção específica.

Destacamos na Tabela 4.8, uma análise descritiva dos dados. Com exceção do cromossomo 1, em todos os outros foram encontrados mais blocos na modelagem GMM do que na modelagem GLM. Por outro lado, conforme esperado, o tamanho dos blocos é maior quando usamos a modelagem GLM. Os valores médios são bem próximos entre os cromossomos, variando nos intervalos $(5,84; 5,93)$ e $(3,87; 3,94)$ para os modelos GLM e GMM, respectivamente. A mediana também confirma os blocos maiores para o modelo GLM, com valor igual a 5, enquanto que no caso do modelo GMM, o valor é 3, em todos os cromossomos.

**Tabela 4.8:** *Estatísticas descritivas do processamento de cada cromossomo, com as duas modelagens propostas.*

| | | | Tamanho dos Blocos | | | |
|---|---|---|---|---|---|---|
| Cromossomo | Modelagem | Blocos encontrados | mínimo | máximo | média | mediana |
| 1 | GLM | 11186 | 2 | 41 | 5,69 | 5 |
| | GMM | 11138 | 2 | 24 | 3,87 | 3 |
| 2 | GLM | 11261 | 2 | 54 | 5,88 | 5 |
| | GMM | 11567 | 2 | 23 | 3,93 | 3 |
| 3 | GLM | 9307 | 2 | 44 | 5,91 | 5 |
| | GMM | 9619 | 2 | 23 | 3,94 | 3 |
| 4 | GLM | 8642 | 2 | 38 | 5,84 | 5 |
| | GMM | 8958 | 2 | 22 | 3,94 | 3 |
| 5 | GLM | 8685 | 2 | 37 | 5,86 | 5 |
| | GMM | 8941 | 2 | 22 | 3,93 | 3 |
| 6 | GLM | 8548 | 2 | 35 | 5,93 | 5 |
| | GMM | 8933 | 2 | 27 | 3,90 | 3 |

As Figuras 4.4 a 4.9 mostram porções genômicas com aproximadamente 70 SNP's, escolhidas de forma aleatória em cada um dos cromossomos processados, para ilustrar os blocos de dependência obtidos com cada modelagem.

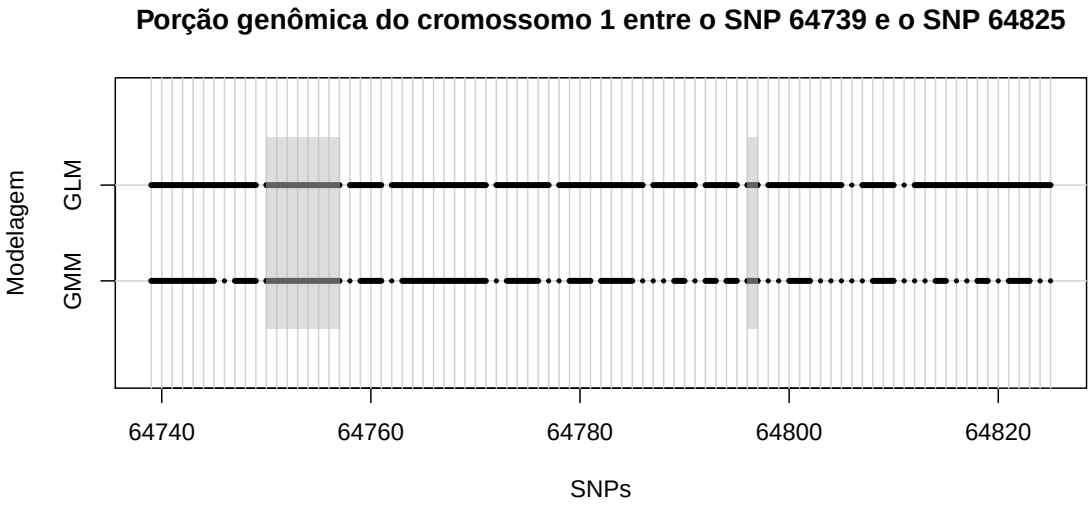



**Figura 4.4:** *Porção genômica do cromossomo 1, mostrando os blocos de dependência obtidos com as modelagens GLM e GMM.*

No cromossomo 1 (Figura 4.4), vemos em destaque duas regiões (entre os SNP's 64750 e 64757 e também entre os SNP's 64796 e 64797), nas quais foram encontradas uma mesma estrutura de dependência em ambos os modelos. São exemplos de regiões com Índice de Influência Familiar igual a zero, ou seja, a dependência entre os SNP's nessas regiões não pode ser explicada pela relação de parentesco dos indivíduos da amostra. No Capítulo 3, foram feitos testes de simulação para verificar se a modelagem GMM seria capaz de reconhecer blocos pequenos de SNP's. Em particular, a região delimitada pelos SNP's 64796 e 64797, com apenas dois SNP's é um caso prático que corrobora com os resultados obtidos com dados simulados.

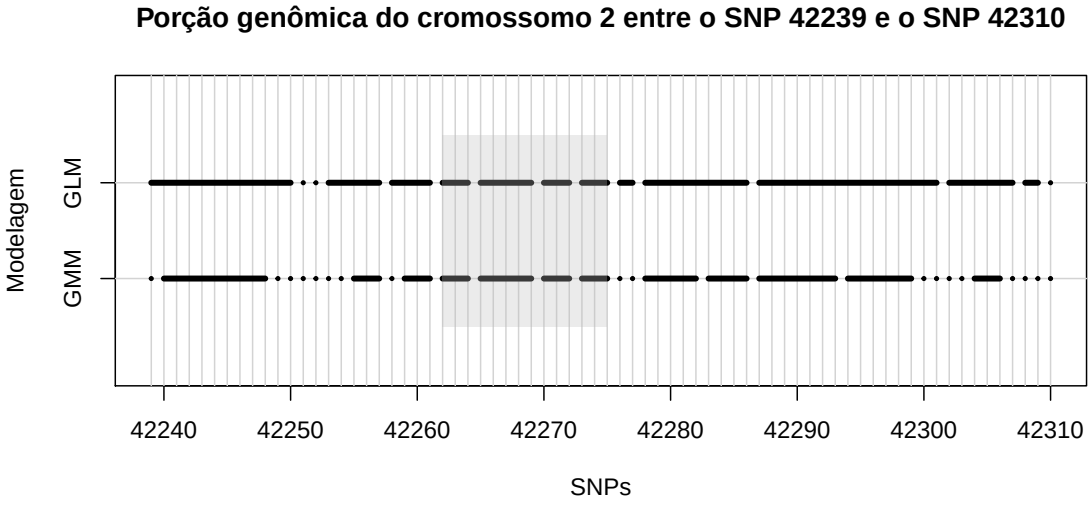



**Figura 4.5:** *Porção genômica do cromossomo 2, mostrando os blocos de dependência obtidos com as modelagens GLM e GMM.*

No cromossomo 2 (Figura 4.5) vemos uma região entre os SNP's 42262 e 42275 com a mesma estrutura de dependência obtida pelas duas modelagens, composta por 4 blocos. É mais um exemplo de uma região na qual a dependência não pode ser explicada pela relação de parentesco, neste caso, formada por mais de um bloco de dependência.

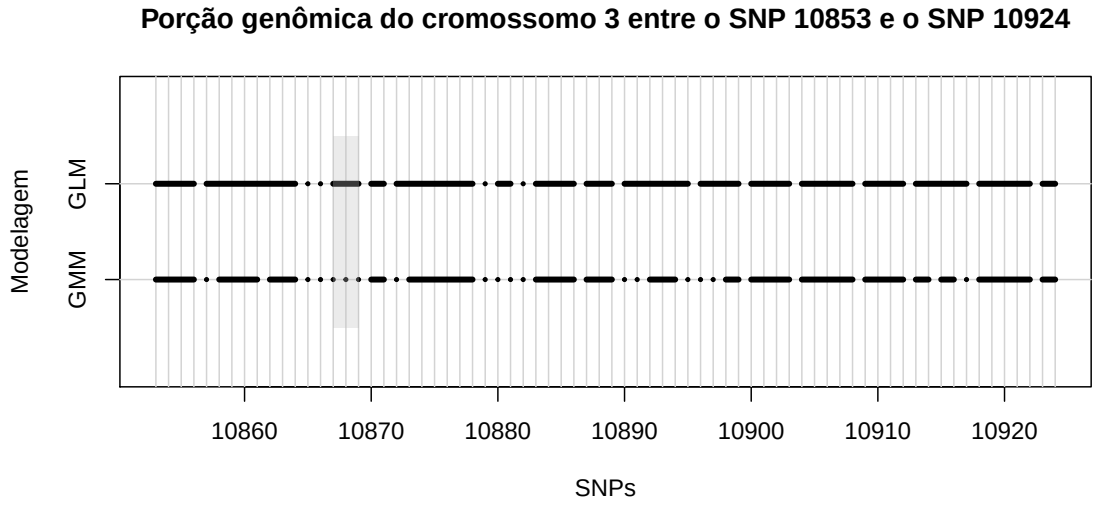



**Figura 4.6:** *Porção genômica do cromossomo 3, mostrando os blocos de dependência obtidos com as modelagens GLM e GMM.*

A Figura 4.6 destaca uma região de 3 SNP's, que formam um bloco quando modelados pelo GLM e são independentes na modelagem GMM. É um exemplo de Índice de Influência Familiar máximo, no qual toda a dependência pode ser explicada pela relação familiar.

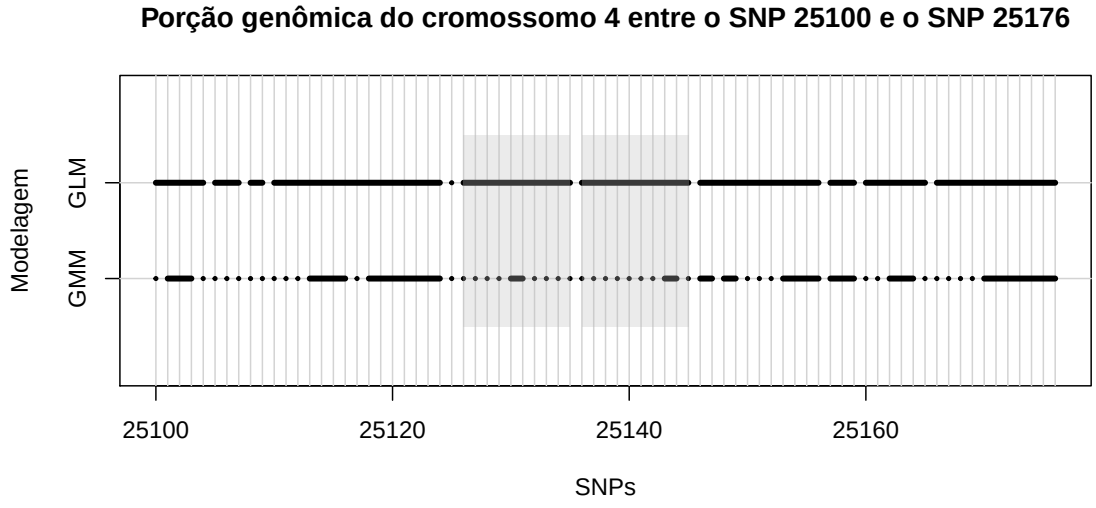



**Figura 4.7:** *Porção genômica do cromossomo 4, mostrando os blocos de dependência obtidos com as modelagens GLM e GMM.*

Na Figura 4.7, temos duas regiões em destaque que são exemplo do caso utilizado para definir o limite superior do IIF, discutido na Seção 2.5.6, no Capítulo 2. Caso os blocos com dois SNP's (entre 25130-25131 na primeira região e entre 25143-25144 na segunda região), fossem também de SNP's independentes, o IIF seria máximo nessas porções. Contudo, o cálculo do Índice de Influência Familiar nestes casos é igual a $0,45$ em ambos os casos. Por maior que seja o bloco de dependência nesses casos, o valor do IIF nunca ultrapassará (na verdade nunca nem mesmo chegará) a $1/2$, justificando assim o limite superior imposto no índice proposto.

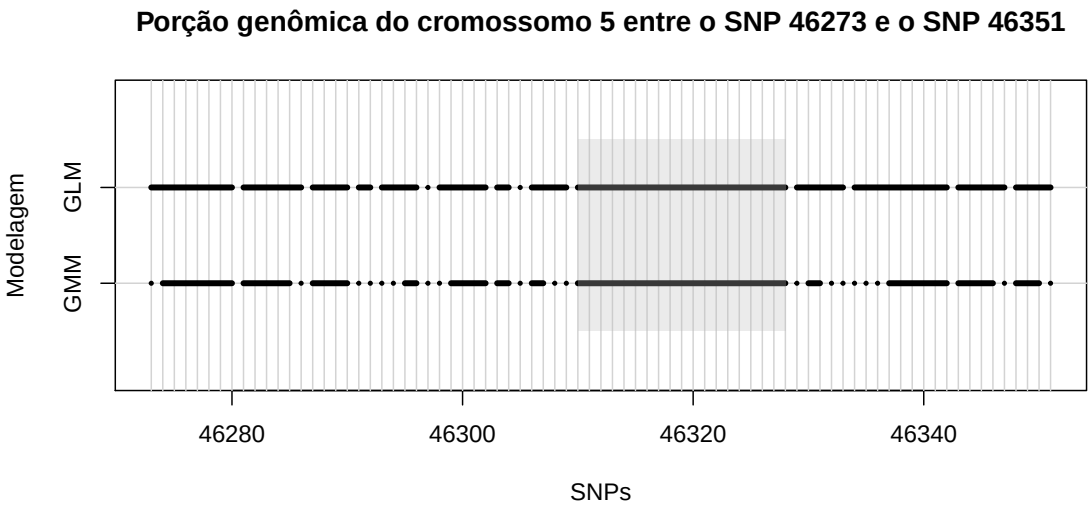



**Figura 4.8:** *Porção genômica do cromossomo 5, mostrando os blocos de dependência obtidos com as modelagens GLM e GMM.*

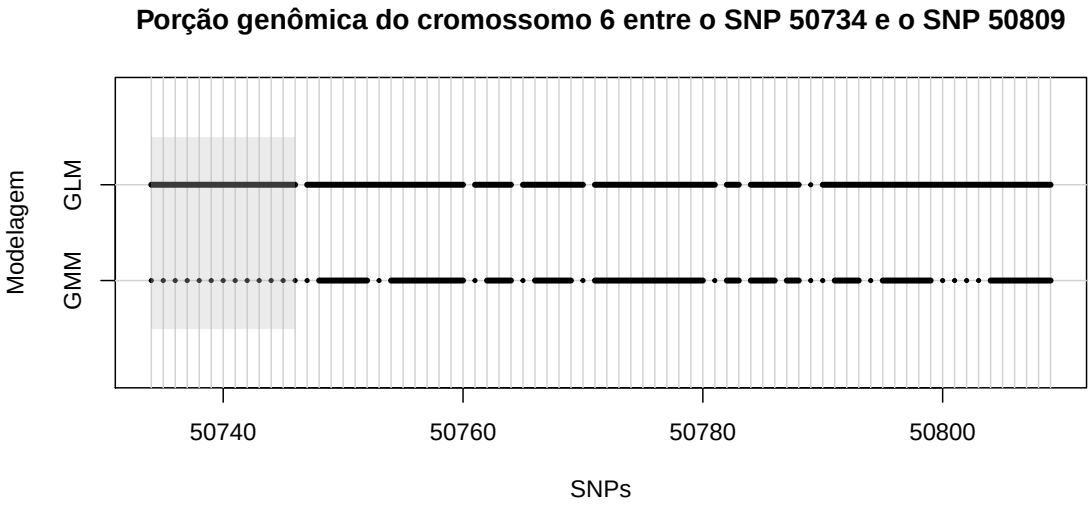



**Figura 4.9:** *Porção genômica do cromossomo 6, mostrando os blocos de dependência obtidos com as modelagens GLM e GMM.*

Finalmente, as Figuras 4.8 e 4.9 mostram dois exemplos de blocos relativamente grandes na modelagem GLM (19 SNP's no cromossomo 5 e 13 SNP's no cromossomo 6), o primeiro deles com IIF igual ao mínimo (0) enquanto que o segundo com IIF máximo (0, 5).

Como seria inviável mostrar todos os blocos de dependência encontrados em cada cromossomo, optamos por mostrar o comportamento do Índice de Influência Familiar (IIF), por cromossomo, através de um gráfico de média móvel de 500 SNP's, mostrados nas Figuras 4.10 a 4.15.

Uma comparação entre os gráficos nos revela que o comportamento do IIF varia bastante entre os cromossomos e não parece haver alguma tendência comum para o índice, seja nas extremidades ou na porção interior de cada um deles. Algumas regiões se destacam em cada cromossomo e serão discutidas brevemente.

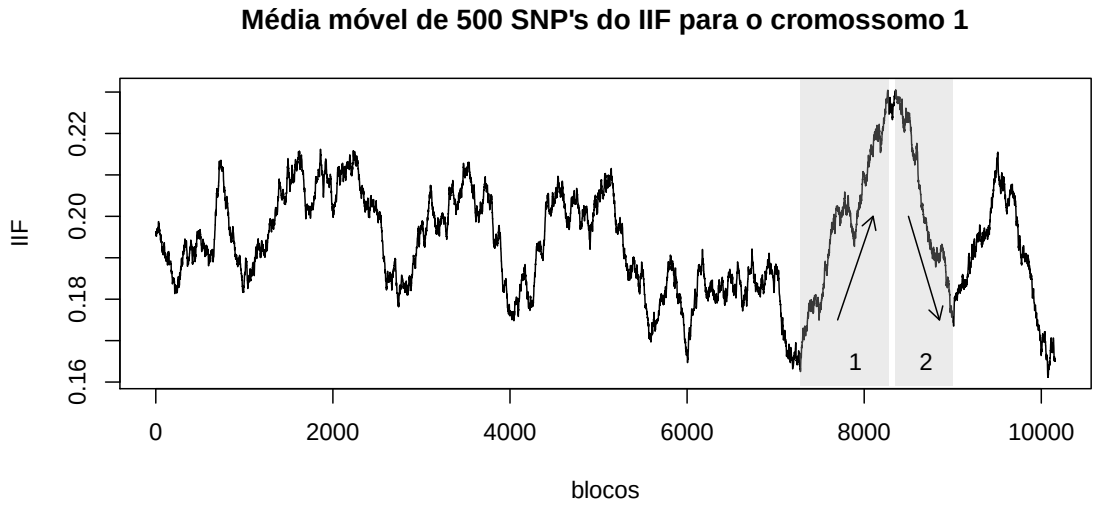



**Figura 4.10:** *Gráfico de média móvel de 500 SNP's do Índice de Influência Familiar - IIF - para os dados processados do cromossomo 1.*

No cromossomo 1 (Figura 4.10), notamos que as duas regiões em destaque mostram um comportamento diferente do observado no restante do cromossomo. Na primeira delas, identificada com o número 1, há um aumento importante do IIF, seguido por uma queda, também relevante, do índice na região 2.

O destaque da Figura 4.11, nos mostra uma região na qual o IIF se mantém maior do que no restante do cromossomo 2, indicando uma porção genômica com mais influência familiar na sua estrutura de dependência.

Na Figura 4.12, ao contrário, conseguimos identificar uma região na qual o IIF é menor que o verificado no restante do cromossomo 3, consequentemente, mostrando uma região na qual a estrutura de dependência tem menos influência familiar.

O cromossomo 4 (Figura 4.13), parece manter o IIF dentro de uma faixa constante de variação, com algumas porções que se distinguem desse comportamento.

No cromossomo 5 nota-se uma queda gradual do índice. Na área em destaque na

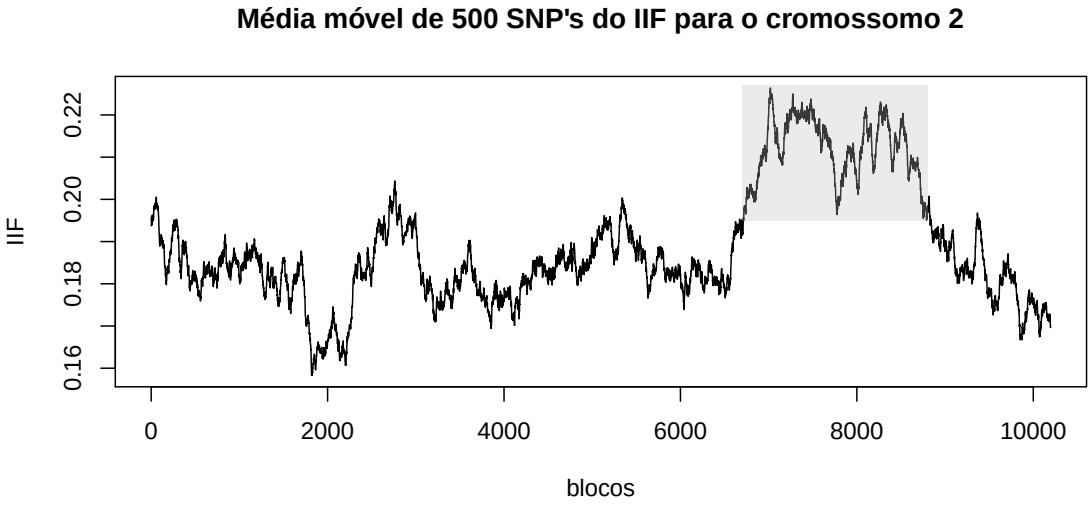



**Figura 4.11:** *Gráfico de média móvel de 500 SNP's do Índice de Influência Familiar - IIF - para os dados processados do cromossomo 2.*

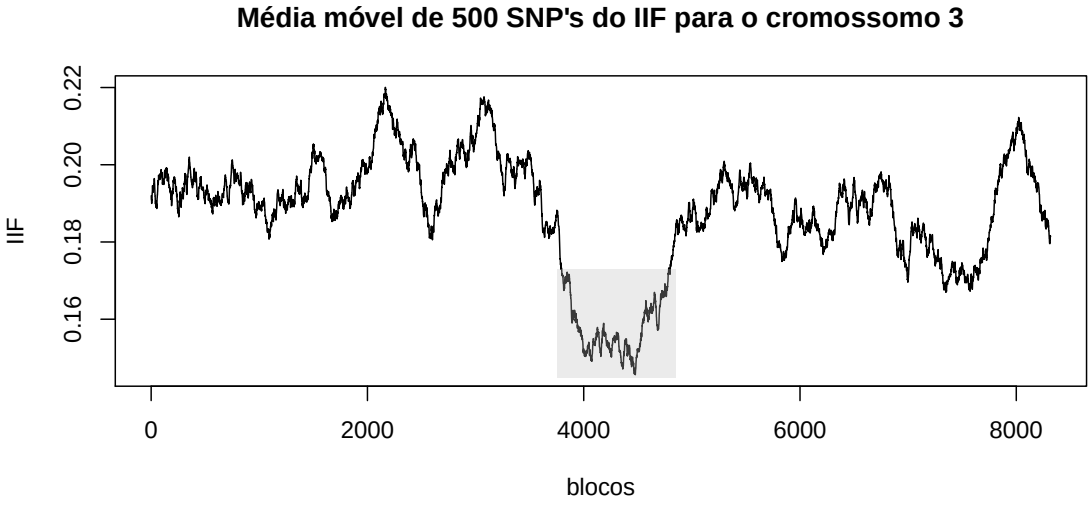



**Figura 4.12:** *Gráfico de média móvel de 500 SNP's do Índice de Influência Familiar - IIF - para os dados processados do cromossomo 3.*

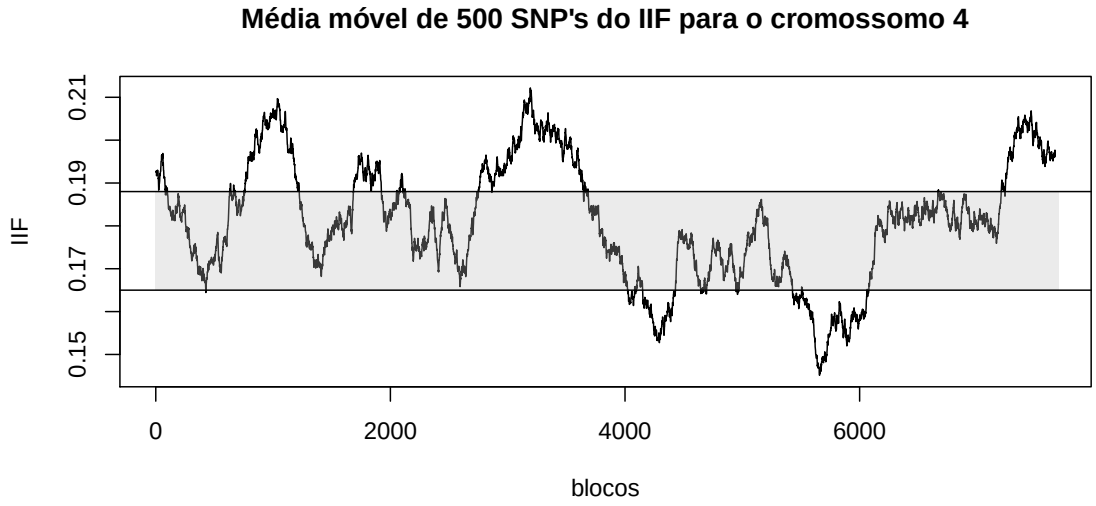



**Figura 4.13:** *Gráfico de média móvel de 500 SNP's do Índice de Influência Familiar - IIF - para os dados processados do cromossomo 4.*

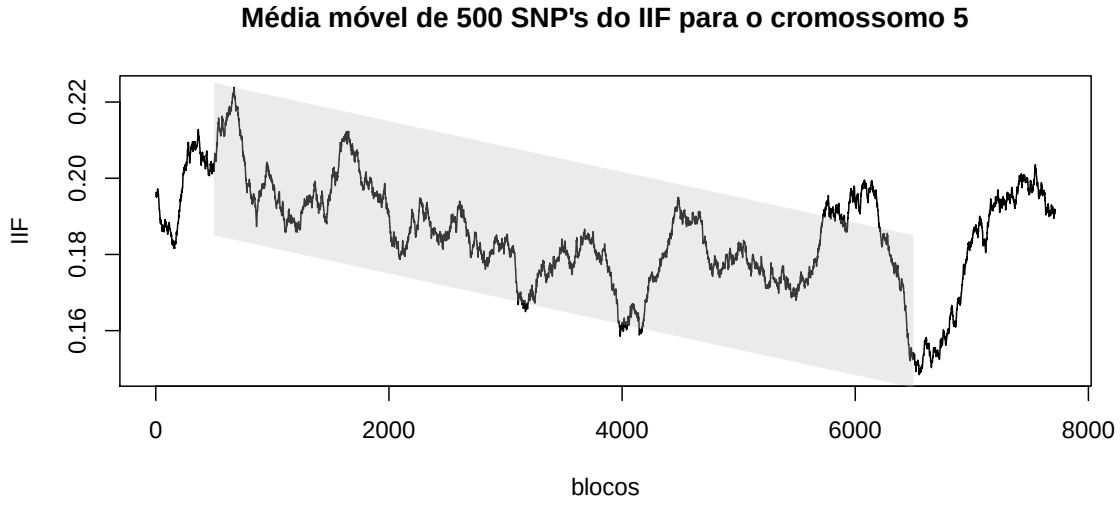



**Figura 4.14:** *Gráfico de média móvel de 500 SNP's do Índice de Influência Familiar - IIF - para os dados processados do cromossomo 5.*

Figura 4.14, vemos que o IIF varia do seu máximo até o seu mínimo dentro do cromossomo.

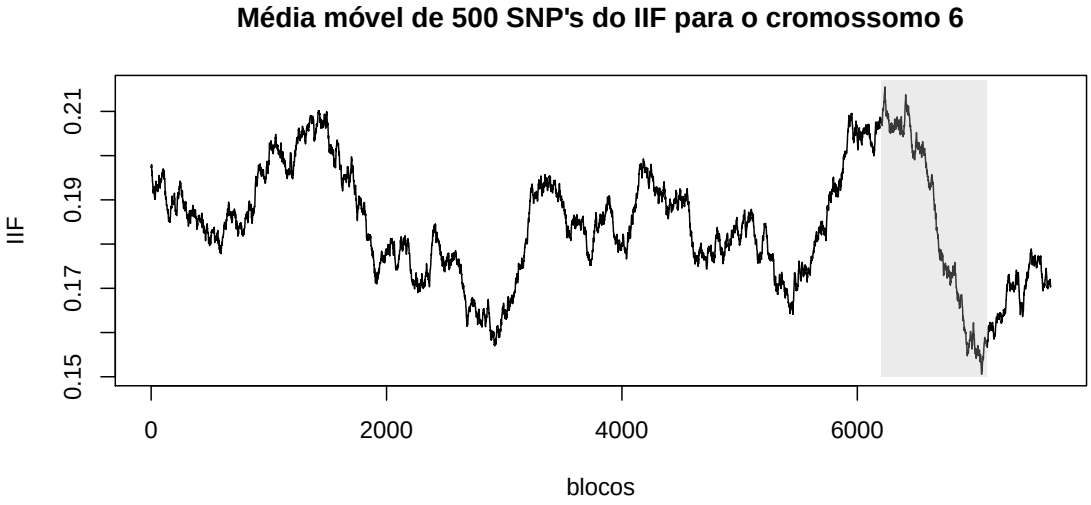



**Figura 4.15:** *Gráfico de média móvel de 500 SNP's do Índice de Influência Familiar - IIF - para os dados processados do cromossomo 6.*

Por fim, a Figura 4.15 evidencia que a variação do máximo para o mínimo do IIF no cromossomo 6 acontece um uma região bem menor, comparativamente ao cromossomo 5.

A possibilidade de representar a influência familiar ao longo de cada cromossomo e, em geral, ao longo de qualquer região mais específica do genoma, nos parece ser uma das principais contribuições de nosso trabalho. De modo geral, o comportamento do IIF em cada região pode, a exemplo das áreas que destacamos aqui, indicar as porções genômicas nas quais a influência familiar mais varia ou onde se mantém constante. A partir desses dados, é possível identificar quais SNP's são pertencentes a essas regiões e então, verificar a possível correspondência biológica que justifique cada padrão.

Alternativamente, é possível estudar o que acontece com o IIF em uma região de interesse, por exemplo, áreas gênicas ou que sabidamente estão envolvidas em processos metabólicos, ajudando a compreensão de quanto a influência familiar pode responder por determinados fenótipos.

Para ilustrar essas duas possibilidades, tomamos como exemplo duas análises adicionais, a primeira delas, avaliando a influência familiar em dois genes conhecidos. A segunda, avaliando como o IIF se comporta em uma região extremamente importante sob o ponto de vista imunológico: o HLA (do inglês *Human Leukocyte Antigens*).

## 4.3 Análise de genes MODY

Analisar como a influência familiar varia ao longo de um ou mais cromossomos, em quais regiões genômicas ela se manifesta mais ou em quais delas ela é menos importante,

pode ser indicativo para estudos que relacionam o nosso genoma com nossas características (fenótipos), nosso metabolismo e, claro, com as doenças que têm relação familiar.

Para o escopo deste estudo, entretanto, sem o devido conhecimento e interpretação biológica, pouco acrescentaria. Para analisar os resultados da solução proposta neste trabalho de uma forma mais objetiva, procuramos avaliar como se comporta a influência familiar em algumas regiões específicas do genoma, já documentadas.

Diversos estudos disponíveis conseguem identificar os genes envolvidos em algumas doenças, entre as quais, aquelas que sabidamente possuem um componente hereditário. Nessa classe encontramos a diabete, um distúrbio no qual a concentração de glicose no sangue encontra-se anormalmente elevada, pois o organismo não libera ou não utiliza a glicose de modo adequado (Brutti *et al.*, 2019).

Apesar de ser considerada uma doença complexa ou multifatorial[2], um dos tipos de diabete, denominada MODY (do inglês *Mature Onset Diabetes of the Young*[3]), é na verdade monogênica. Também chamada de doença Mendeliana, significa que apenas um gene é responsável pela disfunção.

Existem alguns subtipos na diabete tipo MODY, entre os quais seis deles se destacam (Urakami, 2019), relacionados na Tabela 4.9.

**Tabela 4.9:** *Principais subtipos da diabete tipo MODY, indicando o gene envolvido e o cromossomo no qual se localiza.*

| Subtipo | Gene | Cromossomo |
|---|---|---|
| MODY 1 | HNF-4-alpha | 20 |
| MODY 2 | GCK | 7 |
| MODY 3 | HNF-1-alpha | 12 |
| MODY 4 | PDX1 | 13 |
| MODY 5 | HNF-1-beta | 17 |
| MODY 6 | NEUROD1 | 2 |

Analisamos a região dos genes responsáveis pelos dois primeiros subtipos da diabete tipo MODY indicados na Tabela 4.9, para entender em que pontos a influência familiar poderia indicar alguma diferenciação.

Utilizando a ferramenta *Genome Browser* (Kent *et al.*, 2002) para consulta da localização dos genes e o banco de dados de SNP's *dbSNP* (Sherry *et al.*, 2001) (versão 153, genoma humano de referência hg38), foi possível identificar quais marcadores estão anotados em cada gene. Através de um cruzamento com nossa base de dados, encontramos os SNP's de cada gene presentes em nossa amostra, sendo 31 no gene HNF4-alpha e 11 no gene GCK.

A Tabela 4.10 mostra esses marcadores e sua posição em bp[4] (do inglês *base pair*) dentro

[2] Doenças complexas ou multifatoriais são aquelas provocadas por múltiplos genes em combinação com fatores de estilo de vida e ambientais.

[3] Em tradução livre, algo como "diabete diagnosticada na idade madura do jovem".

[4] Considerando que a molécula de cada cromossomo é formada por uma sequência de pares de bases, bp, do inglês *base pair*, indica em qual par de bases se localiza um determinado marcador.

do cromossomo, para os genes HNF4-alpha (a) e GCK (b).

**Tabela 4.10:** *SNP's anotados nos genes HNF4-alpha e GCK, encontrados na amostra de dados de Baependi.*

| SNP | Posição (bp) | SNP | Posição (bp) |
|---|---|---|---|
| rs6031546 | 42418769 | rs3181206 | 42465269 |
| rs16988991 | 42423191 | rs6130609 | 42466509 |
| rs11696298 | 42423383 | rs3212183 | 42468552 |
| rs6103716 | 42433044 | rs3212184 | 42468574 |
| rs6031559 | 42433253 | rs8114057 | 42469866 |
| rs6073418 | 42434004 | rs8122476 | 42470105 |
| rs6065725 | 42438429 | rs11574730 | 42470427 |
| rs6031579 | 42448522 | rs11574733 | 42470836 |
| rs6031580 | 42448695 | rs13041396 | 42475544 |
| rs4812831 | 42451674 | rs1800961 | 42475778 |
| rs717248 | 42459019 | rs11574738 | 42475994 |
| rs717247 | 42459198 | rs3212198 | 42477776 |
| rs2425639 | 42460924 | rs11086925 | 42480762 |
| rs2071199 | 42464280 | rs1028583 | 42484175 |
| rs2071200 | 42464370 | rs6031593 | 42484627 |
| rs6031586 | 42465058 | | |

**(a)** *Gene HNF4-alpha.*

| SNP | Posição (bp) |
|---|---|
| rs2971675 | 44161363 |
| rs17832252 | 44166524 |
| rs758989 | 44169531 |
| rs7793213 | 44171519 |
| rs12540369 | 44172039 |
| rs2300586 | 44185381 |
| rs1303722 | 44185599 |
| rs2300584 | 44185863 |
| rs741038 | 44190004 |
| rs2908289 | 44190467 |
| rs6952751 | 44193238 |
| | |
| | |
| | |
| | |
| | |

**(b)** *Gene GCK.*

Utilizando a solução proposta neste trabalho, foram levantados os blocos da estrutura de dependência de cada um desses genes, utilizando as duas modelagens (GLM e GMM). Os resultados obtidos estão mostrados nas Figuras 4.16 e 4.17, para os genes HNF4-alpha e GCK, respectivamente.

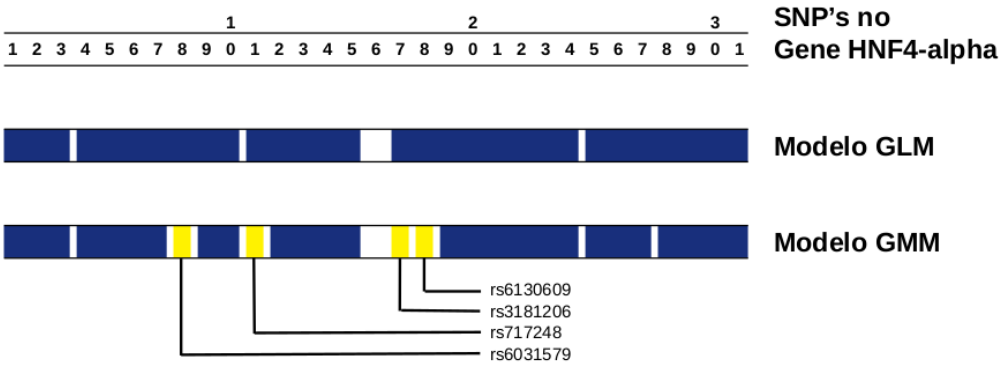



**Figura 4.16:** *Blocos de dependência obtidos para a região do gene HNF4-alpha.*

Analisando a estrutura de dependência inferida a partir das duas modelagens, vemos que alguns marcadores que estão inseridos em um bloco GLM, aparecem como independentes quando o modelo é o GMM. Assim, esses SNP's em destaque nas Figuras 4.16 e 4.17, são o exemplo da dependência que pode ser explicada pela relação familiar. Poderiam, portanto, ser possíveis alvos em estudos de associação para compreender a contribuição da co-segregação familiar no padrão hereditário desse tipo de doença.

Considerando a influência familiar de cada região descrita, os valores calculados para o IIF foram de 0,0682 (ou 13,64% do máximo 0,5) e 0,2095 (ou 41,90%), para os genes

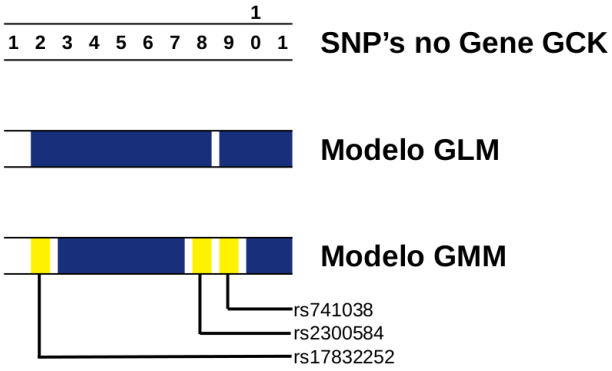



**Figura 4.17:** *Blocos de dependência obtidos para a região do gene CGK.*

HNF4-alpha e GCK, respectivamente (Tabela 4.11). Por essa métrica e baseado nos dados disponíveis em nossa amostra, parece que a estrutura de dependência na região do gene GCK, responsável pelo subtipo 2 da MODY, pode ser mais explicada pela influência familiar do que a estrutura do gene HNF4-alpha, do subtipo 1.

Cumpre ressaltar que esse resultado foi obtido a partir da amostra de Baependi, a qual, talvez, não tenha mapeado um conjunto relevante de marcadores nessas regiões específicas. Uma pesquisa mais aprofundada poderia, por exemplo, utilizar plataformas de SNP's com uma densidade maior nessas regiões de interesse para buscar resultados mais acurados. Essa aplicação contribui, contudo, para ilustrar o potencial da solução proposta.

**Tabela 4.11:** *Valores do Índice de Influência Familiar (IIF) calculados para a região dos genes dos subtipos 1 e 2 da diabete MODY.*

| Subtipo | Gene | IIF | IIF % |
|---|---|---|---|
| MODY 1 | HNF4-alpha | 0,0682 | 13,64% |
| MODY 2 | GCK | 0,2095 | 41,90% |

## 4.4 HLA - *Human Leukocyte Antigen*

A chamada região HLA é uma porção genômica localizada no cromossomo 6 (Figura 4.20), responsável pela resposta imunológica do ser humano. Apesar de relativamente pequena, é uma região rica em polimorfismos (Balding *et al.*, 2008) e muito estudada dada sua relação com doenças imunológicas, em particular, doenças auto-imunes e alguns tipos de câncer e é de fundamental importância para determinação de compatibilidade em caso de transplantes. A conclusão diagnóstica para muitas doenças, atualmente, envolve levantamento genotípico de marcadores específicos da região HLA.

Apenas como base de comparação de densidade de polimorfismos, novamente utilizando o banco de dados de SNP's *dbSNP* (Sherry *et al.*, 2001) (versão 153, genoma humano de referência hg38), a densidade de SNP's anotados no cromossomo 6, fora da região HLA, é de 1 SNP a cada 4.9 pares de bases, enquanto que na região HLA temos 1 SNP a cada 4.4 pares de bases.

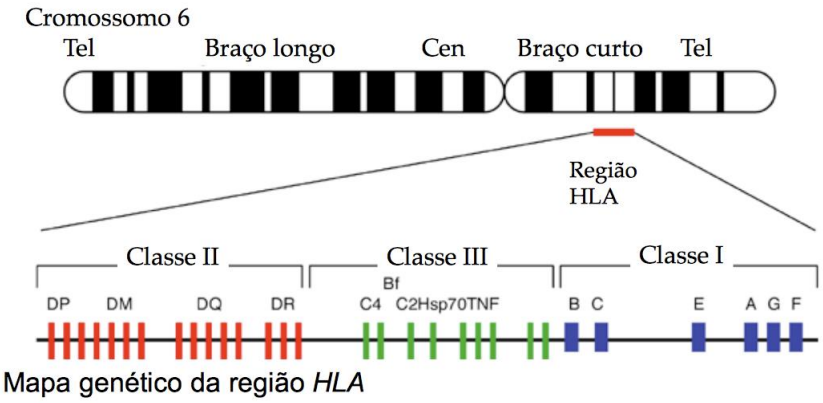



**Figura 4.18:** *Representação esquemática do cromossomo 6 com a região HLA em destaque (Bitarello, 2011) - Expert Reviews in Molecular Medicine©2003 Cambridge University Press.*

Estima-se que, em grande parte, a presença do ser humano na Terra e sua sobrevivência em face à exposição a inúmeros agentes externos, esteja relacionada ao mecanismo que a evolução natural impôs a esta região genômica (Fernandes, 2016).

Sendo assim, apesar de bastante polimórfica, é razoável esperar que seja uma região conservada do genoma, garantindo nossa defesa imunológica, independentemente da ancestralidade. De fato, no trabalho de Fernandes, 2016, foi feita uma avaliação das estruturas de dependência dessa região genômica, nas 12 populações mundiais em estudo (do projeto HapMap) além de uma amostra de brasileiros. Foram encontrados blocos de dependência com tamanho médio superior àqueles encontrados no restante do cromossomo (Figura 4.19), evidenciando o efeito de conservação.

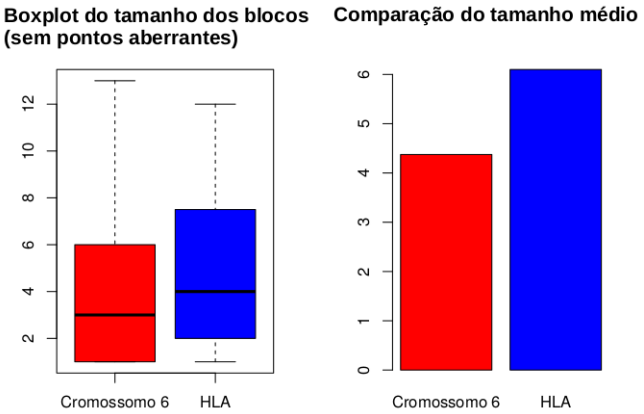



**Figura 4.19:** *Comparação do tamanho dos blocos de dependência na região HLA e no cromossomo 6 (Fernandes, 2016).*

No mesmo estudo, foi proposto um índice de similaridade para comparar as estruturas de dependência encontradas nas diversas populações. Os valores calculados desse índice foram superiores à mediana do cromossomo em 82% dos casos. Assim, com base nos resultados para indivíduos independentes, a região HLA se mostrou uma região conservada e homogênea em termos de estrutura de dependência.

Sendo essa a região alvo para determinar a compatibilidade entre dois indivíduos no caso de transplantes e sabendo que essa compatibilidade é mais provável entre pessoas da mesma família; lembrando também que muitas doenças de origem imunológica têm um forte componente hereditário e, finalmente, considerando os resultados de homogeneidade e conservação obtidos para indivíduos independentes, cabe questionar o que esperar da estrutura de dependência quando analisamos pessoas com parentesco e, sobretudo, quanto da dependência nessa região pode ser explicada pela influência familiar?

Dos SNP's anotados no HLA, foram encontrados 5.052 marcadores em nossa amostra de dados. Analisando os resultados obtidos para o Índice de Influência Familiar (IIF) para o cromossomo 6, com foco na região HLA, o que se nota é uma queda consistente da média móvel do índice ao longo dessa porção genômica, como pode ser visto na Figura 4.20, que destaca a região.

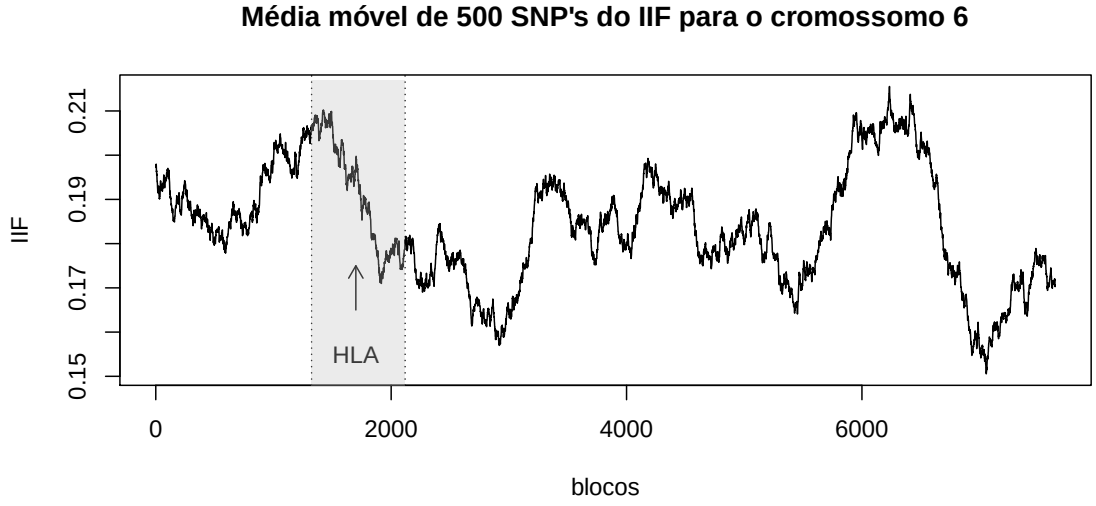



**Figura 4.20:** *Média móvel do IIF ao longo do cromossomo 6, com destaque para a região HLA.*

Esse resultado nos parece bastante coerente com a realidade biológica. Apesar do fato dessa região ser muito polimórfica e com grandes blocos de dependência em sua estrutura, esta dependência não deveria ser devida ao efeito de família. Caso o fosse, representaria dizer que seria possível encontrar famílias com mais chances de sobreviver do que outras, o que não parece ser o que vemos na verdade.

Neste ponto cumpre ressaltar a diferença entre o que pode ser entendido de hereditariedade e o que aqui chamamos de influência familiar. Hereditariedade ou ainda, o conceito estatístico de herdabilidade, é uma medida do quanto da variabilidade total de um fenótipo (quantitativo) pode ser devida a fatores genéticos (o coeficiente de herdabilidade, dado por uma proporção entre variabilidade do efeito genético e variabilidade total, é também um coeficiente de correlação intra-indivíduos com grau de parentesco). A influência familiar, por outro lado, procura indicar o quanto da estrutura de dependência do genoma pode ser explicada pelo parentesco entre os indivíduos.

A região HLA consegue mostrar com mais clareza essa diferença. Apesar de ser uma região com muitos marcadores associados a doenças hereditárias, seu índice de influência

familiar não é alto. Isto significa que a estrutura de dependência dessa porção genômica, apesar de conter grandes blocos, não pode ser explicada pela família. Enquanto humanos, é esperado que nosso sucesso na luta contra patógenos, deve ser o mesmo, dada a característica conservada e homogênea do HLA, não importando a família à qual pertencemos, o que explica o decréscimo da média do índice de influência familiar, ao longo dessa porção genômica.

Em outras palavras, há sim relações de dependência entre os marcadores dessa região, talvez até favorecendo a ocorrência da hereditariedade de algumas doenças, mas essa dependência se deve a outros fatores. Possivelmente, entre esses fatores esteja justamente a característica dessa região em ser conservada, esta sim garantidora de nossa resposta imunológica e de sobrevivência.

Assim, apesar de ser uma região genômica com um componente hereditário sabidamente alto, a dependência encontrada ao longo dessa porção não é devida à componente familiar, mas ao contrário, segregada na população em geral. Mais uma vez, isso nos parece bastante coerente, uma vez que os blocos dessa região seriam conservados por conta de garantir nossa imunidade, independentemente da origem familiar ou população. Os blocos genômicos que são passados de pai para filho nessa região, provavelmente garantem a compatibilidade, especialmente no caso de transplantes entres parentes, por manterem os mesmos nucleotídeos. Indivíduos de outras famílias, mesmo tendo diferentes nucleotídeos nos blocos respectivos, têm os mesmos blocos de dependência.

Quando combinamos os resultados obtidos em Fernandes, 2016, que considerou dados independentes e encontrou blocos grandes e homogêneos, com a solução deste trabalho, que considerou dados de família e encontrou baixa influência familiar, acabamos por caracterizar o HLA, em termos de estrutura de dependência, como uma região conservada, com blocos de dependência homogêneos comparativamente às populações e com baixa influência familiar, tendo em vista que a dependência encontrada não pode ser explicada pelo parentesco.

# Capítulo 5

# Considerações Finais

## 5.1 Conclusões, contribuições e extensões

Muitas questões em aberto relativas à evolução do ser humano, variabilidade, doenças, habilidades e até mesmo como aumentar a nossa expectativa de vida, talvez tenham suas respostas escritas em algum ponto do nosso genoma. Compreender como os blocos de material genético se combinam para formar indivíduos da próxima geração, quais porções desse material se mantêm e quais se misturam, parece ser de fundamental importância na busca pelo melhor entendimento de nossa herança, pela medicina de precisão, pela medicina personalizada.

As chamadas doenças genéticas que, em poucas palavras, seriam aquelas cuja prevalência em determinadas famílias é maior do que na população geral, são objeto de vários estudos, muitas vezes desafiadores em termos de compreensão. Ainda, maior desafio reside em entender o padrão de (co)segregação de doenças multifatoriais, cuja regulação envolve tanto componentes específicos de famílias, bem como aqueles presentes na população geral, além de interações com variáveis ambientais. Como consequência, a análise de dados de famílias é de fundamental importância e técnicas estatísticas que conseguem lidar com dados correlacionados e identificar a natureza dos blocos de dependência do genoma passam a ser uma exigência.

É neste contexto que este trabalho encontra a sua importância. Procuramos uma alternativa analítica para responder à questão de como os blocos genômicos estão organizados e qual sua natureza, utilizando, para isso, informações de indivíduos com parentesco, combinando assim, as duas demandas, compreensão do nosso genoma e caracterização de porções com maior e menor influência familiar.

Este trabalho é uma extensão natural dos desenvolvimentos alcançados em Fernandes, 2016, sob os quais foi proposta uma solução baseada em campos Markovianos para inferir a estrutura de dependência do genoma humano de diversas populações, com ancestralidades distintas, usando amostras de indivíduos independentes. Quando as amostras envolvem dados de indivíduos e seus familiares, considerando que parentes compartilham grandes porções de material genético, diferentemente do que é esperado entre indivíduos não parentes, há um desafio adicional no sentido de encontrar as porções genômicas que

segregam na população geral. Adicionalmente, explicar o quanto de cada bloco é devido à relação familiar (são compartilhados por indivíduos parentes) e quanto é oriundo de outras influências, entre as quais, citamos até mesmo, garantir nossa sobrevivência.

Nossa proposta apresenta uma alternativa inédita, até onde é do conhecimento dos autores, para inferir a estrutura de dependência do genoma, usando amostras de indivíduos com algum grau de parentesco, e identificar os blocos que são compartilhados por indivíduos da população geral, retirando assim, a parte da dependência devido ao parentesco. Indicamos também, baseada na solução proposta, uma opção para mostrar o quanto da dependência de cada porção genômica foi retirada devido à influência familiar.

A exemplo do trabalho de Fernandes, 2016, utilizamos marcadores do tipo SNP (do inglês *Single Nucleotide Polymorphism*) para amostrar o genoma. Comparativamente, a solução modifica a função de pseudo-verossimilhança parametrizada pelo campo Markoviano, adequando à distribuição (Binomial) formulada no nível de indivíduo, e levando em conta o grau de parentesco, em uma formulação de modelos lineares generalizados mistos.

A solução combinada proposta reúne vantagens de duas ferramentas estatísticas, os modelos lineares generalizados mistos e os processos estocásticos representados por campos Markovianos. A primeira permite o tratamento de dados correlacionados, em particular, os chamados dados de família, enquanto que a segunda explora a relação de dependência de uma sequência de variáveis aleatórias, no nosso caso, marcadores tipo SNP. O resultado final é um processo que, considerando a estrutura de correlação dos dados da amostra, infere a estrutura de dependência entre as variáveis resposta (marcadores SNP) que não está sob influência familiar.

Comparando como o campo Markoviano processa as estimativas fornecidas pela modelagem via o modelo linear generalizado misto e o modelo linear generalizado (sem efeito aleatório), conseguimos caracterizar uma nova estrutura de dependência presente nos marcadores, desta vez, aquela que está sob a influência familiar. Finalmente, a comparação entre os dois resultados permitiu identificar o quanto da dependência em uma região genômica, poderia ser explicada pela relação de parentesco. Um índice descritivo foi proposto como métrica para a influência familiar na estrutura de dependência genômica.

Dados foram simulados sob diferentes cenários de dependência entre marcadores, bem como de estruturas familiares, por meio dos quais foi possível caracterizar propriedades da solução proposta. Em particular, foi possível, a partir do entendimento de como simular dados para atender à finalidade do trabalho, recuperar características dos dados simulados, como a identificação da influência familiar em uma determinada porção genômica. Cabe aqui fazer uma ressalva sobre o racional que norteou a escolha da estratégia usada na simulação dos dados. Optamos por simular dados sob princípios coerentemente biológicos, em particular usando recursos do pacote SimPed (Leal *et al.*, 2005). Neste caso, a estrutura de dependência no genoma é decorrente de haplótipos, com probabilidades específicas de ocorrência, em que um progenitor pode passar para um filho, não somente porções do genoma que segregam na população geral, mas também haplótipos co-segregando em poucas famílias. Esta escolha nos permite verificar, o quanto da dependência biológica gerada, a solução combinada é capaz de recuperar. Neste sentido, os resultados encontrados neste trabalho mostraram consistência com nossa proposta.

Contudo, uma alternativa que poderia ter sido adotada, seria simular dados a partir de árvores de contextos incluindo efeitos aleatórios para gerar correlação familiar entre os indivíduos, possivelmente usando o algoritmo proposto em Zanin Zambom *et al.*, 2022. Acreditamos que esta última alternativa poderia mais diretamente mostrar a consistência de nossa proposta, contudo, possivelmente, não garantiria a coerência biológica que existe em dados reais e que este trabalho objetiva inferir. Caracterizar propriedades destas diferentes formas de gerar os dados, é tema que propomos para trabalhos futuros. Além disso, simular dados sob uma coerência biológica, como por meio do Simped, permite mais eficientemente comparar os resultados de nossa solução combinada com a de outros trabalhos que visam encontrar regiões de dependência genômica (Druet e Georges, 2010; Geraci, 2010; Delaneau *et al.*, 2019).

Aplicada a uma amostra da população brasileira, contendo mais de dois mil indivíduos distribuídos em quase uma centena de famílias, avaliada em cerca de 800 mil marcadores SNP da plataforma Affymetrics 6.0, a solução proposta mostrou coerência com o que é esperado do ponto de vista teórico e confirmado nas simulações.

Duas regiões genômicas específicas, referentes a dois genes responsáveis por subtipos da doença diabete, foram analisadas com os dados amostrais disponíveis, tendo sido possível indicar alguns marcadores moleculares localizados em regiões candidatas a explicar a influência familiar (sabidamente conhecida) nesses subtipos.

Adicionalmente, foi possível estudar a região HLA (do inglês *Human Leukocyte Antigens*), uma pequena porção no cromossomo 6 responsável por nossa resposta imunológica. Os resultados obtidos revelaram que nessa região, a influência familiar decai, o que nos parece biologicamente razoável, por ser uma região considerada conservada do genoma, logo, a estrutura de dependência nessa região não deveria ser diferente de família para família.

Sob o ponto de vista biológico, a solução proposta neste trabalho, além de inferir a estrutura de dependência do genoma humano usando dados de família, pode contribuir indicando a influência familiar em cada porção genômica. Por conta disso, uma possível contribuição é lançar luz sobre regiões genômicas com alta influência familiar que possam indicar alvos para estudos de doenças com componente hereditário.

Em termos da modelagem estatística, nossa proposta é flexível, uma vez que permite a utilização de diferentes distribuições da variável resposta (número de alelos de referência de acordo com o genótipo do indivíduo no SNP), bem como de estruturas de correlação entre as observações, sem, contudo, impor uma estrutura específica de dependência para as variáveis (como a uniforme, autocorrelação, etc). Neste trabalho, consideramos a distribuição Trinomial, bem como a Binomial, mas, para casos específicos, como organismos poliplóides, estes modelos probabilísticos podem ser estendidos para outras classes. Também, adotamos a matriz de parentesco sob a definição estrita , isto é, obtida pela autodeclaração de parentesco, contudo, diretamente, nossa solução pode ser estendida para incorporar a matriz de relacionamento genômico entre indivíduos (Wang e Thompson, 2019).

No caso da aplicação utilizada neste trabalho, os dados foram extraídos de uma única população (a brasileira) e não foram consideradas questões de ancestralidade. Estudos

genéticos, em geral, devem levar em conta os impactos da miscigenação e estratificação populacional. Uma das abordagens para isso é através do uso de marcadores específicos de ancestralidade[1]. Outra abordagem é obter componentes principais de ancestralidade (Price *et al.*, 2006; De Andrade *et al.*, 2015). Em ambos os casos, uma das alternativas para encontrar as regiões de dependência genômica ajustada por efeitos de miscigenação e estratificação populacional, é usar essas variáveis como covariáveis na modelagem. A solução aqui desenvolvida, permite o uso dessa abordagem de forma direta pelo ajuste dos modelos GMM ou GLM com a inclusão de covariáveis.

Mais do que isso, a modelagem proposta permite a inclusão de outras covariáveis ou variáveis de ajuste nos modelos tais quais, gênero, idade, população de origem, entre outras, o que pode refinar ainda mais a estimativa do parâmetro da distribuição de probabilidades e, consequentemente, ajustar o contexto necessário para o campo Markoviano. Assim, a exemplo do que foi proposto aqui, comparando modelos com e sem a informação das covariáveis, deve ser possível verificar quanto da estrutura de dependência pode ser explicada por elas.

Também, o presente trabalho, desde o modelo inicial para indivíduos independentes, à modelagem de dados de famílias, foi desenvolvido utilizando dados genotípicos, ou seja, a informação genética dos alelos em *loci* de cromosssomos homólogos, o que conduziu à avaliação dos contextos em uma amostra de $n$ indivíduos à partir do alfabeto $\{0, 1, 2\}$. Uma alternativa de extensão seria desenvolver a formulação da solução proposta para dados haplotípicos, que consideram cada cromossomo homólogo separadamente, e, portanto, induziriam a um alfabeto $\{0, 1\}$. Conforme já discutido no Capítulo 1, essa abordagem tem a desvantagem de exigir passos adicionais de estimação para determinar a fase alélica, contudo, além de dobrar o tamanho amostral (de $n$ indivíduos para $2n$ cromossomos), permite a comparação com outros trabalhos que se utilizam dessa codificação.

Apesar do trabalho ter sido, majoritariamente, baseado em resultados de simulações e interpretação dos resultados dos modelos GLM e GMM e a sua combinação com o campo Markoviano, uma boa parte da pesquisa foi feita para o entendimento da diferença entre as duas abordagens. Isso incluiu a diferença entre as matrizes de correlação para dados de família (de relacionamento) e para dados independentes (identidade). Alguns esforços foram feitos no sentido de comparar as duas matrizes e tirar, da diferença entre elas, a explicação analítica para sustentar os resultados de consistência esperados e obtidos por simulação. Além disso, os critérios propostos nesta tese, para encontrar as regiões genômicas com influência familiar, foram obtidos sob um contexto descritivo, sendo ainda necessário construir um teste de significância estatística para declarar a influência familiar em um bloco genômico. Para essa finalidade, o teste pode ser construído, por exemplo, a partir da significância do componente de variância genético incorporado no GMM. Contudo, nessa abordagem, seria necessário combinar todos os testes exigidos para se estabelecer a vizinhança de um SNP, bem como na concatenação das vizinhanças de SNPs adjacentes. Faz parte, portanto, dos nossos objetivos futuros, incluir no desenvolvimento analítico da solução proposta, testes de significância para a declaração de influência familiar.

Por fim, um dos grandes desafios enfrentados na realização deste trabalho, foi a exigên-

[1] Marcadores de ancestralidade são marcadores que possuem grandes diferenças entre subtipos populacionais, chamados marcadores AIM, do inglês *Ancestry Informative Marker* (Laird e Lange, 2010).

cia computacional da solução, a qual foi implementada utilizando a linguagem **R** (R Core Team, 2020). Conforme pode ser visto no Capítulo 4, o número de modelos calculados facilmente se enquadra na ordem de centenas de milhares, considerando cada cromossomo. O tempo total de processamento acaba sendo um limitante. Vários esforços foram feitos no sentido de contornar esse problema, entre os quais vale destacar a paralelização do processamento e o uso de supercomputadores disponíveis na Universidade de São Paulo (USP). O programa foi reescrito para permitir um processamento incremental, ou seja, processar SNP's separadamente para depois consolidar o resultado e usar largura de janela variável. É possível, por exemplo, processar os SNP's considerando uma vizinhança de 2 SNP's para cada lado e depois acrescentar o processamento apenas das vizinhanças do terceiro SNP de ambos os lados.

Todos esses esforços foram implementados e acabaram por permitir um processamento distribuído, o que viabilizou a execução da solução, mas ainda com alto custo de tempo computacional, o que diretamente restringiu a avaliação de um maior número de marcadores e de réplicas em nossos estudos de simulação. Utilizando uma ferramenta de análise de tempo gasto no processamento, depois de todas essas melhorias na estrutura do programa, identificamos que o maior custo de tempo passou a ser o ajuste do modelo linear generalizado misto dentro do pacote lme4qtl. Assim, uma possibilidade para melhorar a performance da solução poderia ser usar a formulação do modelo linear generalizado misto de famílias via a decomposição espectral da matriz de parentesco, como é proposto em Blangero *et al.*, 2013. É um de nossos objetivos futuros tornar mais eficiente o código de implementação computacional da solução combinada proposta.

Este trabalho atingiu o objetivo principal perseguido, qual seja, inferir a estrutura de dependência do genoma humano, refletindo as regiões genômicas que segregam na população geral, utilizando dados de indivíduos com algum grau de parentesco. Adicionalmente, foram propostas alternativas, de base qualitativa bem como quantitativa, visando descrever o que seria a influência familiar, aqui descrita como as porções genômicas que segregam nas famílias. Neste segundo ponto, cabe uma reflexão.

A construção de conhecimento é, principalmente, feita de observações de fatos, em alguns, talvez muitos, casos. O próprio processo científico tem por base a formulação de hipóteses, seguidas do planejamento e da coleta de dados, nos quais as hipóteses são testadas e os resultados são observados, discutidos e, sob condições controladas, interpretados.

Na biologia, em especial, grande parte do desenvolvimento científico se dá através de estudos, experimentais ou observacionais. Na área exata, por outro lado, procura-se provar por desenvolvimento analítico e raciocínio lógico, hipóteses ou conjecturas, que não necessariamente são factuais.

Neste trabalho, levantamos hipóteses de base factual, propomos uma solução estatística para investigar essa hipótese e, acreditamos, que nossos esforços contribuem para o avanço na análise de dados genômicos. Enfatizamos, no entanto, que para alcançarmos o objetivo último da ciência, que é a reprodutibilidade dos resultados, no contexto do problema que estamos tratando, ainda há pontos de pesquisa em aberto.

Finalmente, a pesquisa científica é contínua na busca de expandir as fronteiras do conhecimento. Nosso trabalho deve seguir esse mesmo caminho de continuidade, nesse

sentido, é preciso aprofundar os estudos para elaborar a base analítica que sustente melhor o uso da diferença entre as modelagens GLM e GMM como preditora da influência familiar nos contextos do campo Markoviano. Por outro lado, os resultados obtidos indicam que a abordagem proposta é promissora e nos encorajam a aplicá-la em novos conjuntos de dados, seja para aprimorá-la ou efetivamente responder a perguntas cujas respostas ainda se encontram escondidas dentro de nosso genoma.

# Apêndice A

# Pseudocódigo

Neste apêndice colocamos a lógica, em formato de pseudo-código, dos programas computacionais utilizados para:

A.1 Calcular o valor da função de pseudo-verossimilhança empírica para cada SNP e vizinhança;

A.2 Determinar a vizinhança de cada SNP, por máxima pseudo-verossimilhança empírica penalizada;

A.3 Montagem dos blocos de dependência a partir da concatenação das vizinhanças dos SNP's contíguos.

**Programa A.1** Determinação da pseudo-verossimilhança empírica para SNP/vizinhança.

```
1   carrega(pacotes)  ▷ carga dos pacotes necessários
2   BD ← carrega(database)  ▷ carga dos dados de SNP's
3   KS ← carrega(Kinship)  ▷ carga da matriz de relacionamento
4
5   SNP_total ← ncols(database)  ▷ número de SNP's na base de dados
6
7   FUNCAO Ajusta(SNP, dados, KA, modelo)  ▷ ajuste dos modelos
8       . ▷ SNP é o SNP de interesse
9       . ▷ dados é o SNP de interesse, os SNP's da vizinhança considerada e o ID de cada
            indivíduo
10      . ▷ KA é a matriz Kinship ajustada
11      . ▷ modelo é o tipo de modelo que será ajustado
12
13      se (var(SNP)==0)  ▷ Ajusta probabilidade para SNP degenerado
14        se (SNP==0) P0 ← 1; P1 ← 0; P2 ← 0
15        se (SNP==1) P1 ← 0; P1 ← 1; P2 ← 0
16        se (SNP==2) P2 ← 0; P1 ← 0; P2 ← 1
```

*cont* ⟶

⟶ *cont*

```
17      senao
18        se (modelo==GLM0)  ▷ ajusta modelo GLM sem vizinhança
19          ajustado ← glm(SNP ~ 1, dados, binomial)
20        se (modelo==GLM1)  ▷ ajusta modelo GLM com vizinhança
21          ajustado ← glm(SNP ~ 1 + vz, dados, binomial)
22        se (modelo==GMM0)  ▷ ajusta modelo GMM sem vizinhança
23          ajustado ← gmm(SNP ~ (1|ID), dados, KA, binomial)
24        se (modelo==GMM1)  ▷ ajusta modelo GMM com vizinhança
25          ajustado ← gmm(SNP ~ vz + (1|ID), dados, KA, binomial)
26        p_estimado ← fitted(ajustado)
27        P0 ← (1 − p_estimado)^2
28        P1 ← 2*(1 − p_estimado)*p_estimado
29        P2 ← (p_estimado)^2
30      fim
31
32      L0 ← soma(log(P0,base=3)) quando SNP==0
33      L1 ← soma(log(P1,base=3)) quando SNP==1
34      L2 ← soma(log(P2,base=3)) quando SNP==2
35
36      Devolva (L0+L1+L2)
37  fim
38
39
40
41
42  FUNCAO Principal (SNP_total, SNP_inicial, SNP_final, Largura_inicial, Largura_final)
43
44      enquanto (j entre SNP_inicial e SNP_final)
45          l_min ← max(1, (j − Largura_final))
46          l_max ← max(1, (j − 1))
47          r_min ← min(SNP_total, (j + 1))
48          r_max ← min(SNP_total, (j + Largura_final))
49
50          enquanto (l entre l_min e j)
51              enquanto (r entre r_max e j)
52                  se ((l > j−Largura_inicial) & (r < j+Largura_inicial))
53                      incrementa(r)
54                      proxima iteracao
55                  fim
56                  se (l==j) janela_l ← concatena()
57                  senao janela_l ← concatena(l até l_max)
58                  se (r==j) janela_r ← concatena()
59                  senao janela_r ← concatena(r_min até r)
60
61                  janela ← concatena(janela_l,janela_r)
62
```

*cont* ⟶

⟶ *cont*

```
63                  se (janela == ())  ▷ estimação sem vizinhança
64                      SNP ← casos_completos(BD[j])  ▷ valores não faltantes do SNP_j
65                      dados ← concatena(SNP,ID)  ▷ valores não faltantes do SNP e ID
66                      n ← nrows(dados)  ▷ número de registros (indivíduos) nos dados
67                      KA ← KS[ID,ID]  ▷ sem vizinhança, KA é igual a KS
68                      SomaL_GLM0 ← Ajusta(SNP, dados, KA, GLM0)
69                      SomaL_GMM0 ← Ajusta(SNP, dados, KA, GMM0)
70                      Modelos_GLM ← incluir(concatena(SNP, l, r, n, SomaL_GLM0))
71                      Modelos_GMM ← incluir(concatena(SNP, l, r, n, SomaL_GMM0))
72                  senao  ▷ estimação com vizinhança
73                      dados ← casos_completos(BD[janela,j])  ▷ não faltantes SNP_j e
                            vizinhança
74                      SNP ← dados[SNP]
75                      dados ← concatena(dados, ID)
76                      n ← nrows(dados)  ▷ número de indivíduos nos dados
77                      KA ← 0  ▷ montagem da matriz Kinship ajustada
78                      estados ← distintos(dados[janela])  ▷ diferentes estados de
                            vizinhança
79                      enquanto (estado entre 1 e nrows(estados))
80                          estrato ← estado(ID)  ▷ indivíduos de cada estado de vizinhança
81                          KA ← KS[estrato,estrato]
82                      fim
83                      se (nrows(estados) > 1)  ▷ tem mais de um estado de vizinhança
84                        SomaL_GLM1 ← Ajusta(SNP, dados, KA, GLM1)
85                        SomaL_GMM1 ← Ajusta(SNP, dados, KA, GMM1)
86                        Modelos_GLM ← incluir(concatena(SNP, l, r, n, SomaL_GLM1))
87                        Modelos_GMM ← incluir(concatena(SNP, l, r,n,SomaL_GMM1))
88                      senao
89                        SomaL_GLM0 ← Ajusta(SNP, dados, KA, GLM0)
90                        SomaL_GMM0 ← Ajusta(SNP, dados, KA, GMM0)
91                        Modelos_GLM ← incluir(concatena(SNP, l, r, n, SomaL_GLM0))
92                        Modelos_GMM ← incluir(concatena(SNP, l, r,n,SomaL_GMM0))
93                      fim
94                  fim
95                  incrementa(r)
96              fim
97              incrementa(l)
98          fim
99          incrementa(j)
100     fim
101     Devolva Modelos_GLM
102     Devolva Modelos_GMM
103 fim
```

**Programa A.2** Determinação da vizinhança de cada SNP.

```
1   ▷ carga da pseudo-verossimilhança calculada para cada SNP e vizinhança
2   Modelos ← Modelos_GLM
3   ▷ ou
4   Modelos ← Modelos_GMM
5
6   PL ← concatena(Modelos, wl← SNP−l, wr=r−SNP)  ▷ acrescenta largura das janelas
7   PL ← concatena(PL, W← wl+wr)  ▷ acrescenta largura da janela total
8
9   PL ← concatena(PL, penal← (3^W · log(n, base = 3)))  ▷ penalização
10  PL ← concatena(PL, LLpenal← (LL − penal))  ▷ pseudo-verossimilhança penalizada
11
12  PL ← agrupa(PL por SNP)
13  Max_PL ← max(PL[LLpenal])
14
15  Vz_SNP ← concatena(Max_PL[SNP], Max_PL[l], Max_PL[r])  ▷ vizinhança de cada SNP
```

**Programa A.3** Montagem dos Blocos de dependência.

```
1   vizinhancas ← seleciona(Vz_SNP onde W ≥ 1)  ▷ SNP's não independentes
2   comeco ← min(vizinhancas[l])
3   enquanto (comeco ≤ max(vizinhancas[r])
4       finais ← seleciona(vizinhancas onde l==comeco)
5       final ← max(finais[r])
6       enquanto (max(seleciona(vizinhancas, onde l ≥ comeco & l ≤ final)[r]) > final)
7           finais ← seleciona(vizinhancas onde l ≥ comeco & l ≤ final)
8           final ← max(finais[r])
9       fim
10      bloco ← concatena(comeco, final)
11      Blocos ← incluir(bloco)
12      comecos ← seleciona(vizinhancas onde l>final)
13      comeco ← min(comecos[l])
14  fim
15
16  ▷ inclusão dos SNP's independentes
17  indice ← 1
18  SNPs ← Vz_SNP[SNP]  ▷ separa todos os SNP's
19  enquanto (indice < nrows{Blocos})
20      bloco ← seleciona_registro(Blocos, indice)
21      sequencia ← seleciona(blocos onde SNPs ≥ comeco & SNPs ≤ final)
22      dependentes ← incluir(sequencia)
23  fim
24  independentes ← retira(SNPs, dependentes)
25  Blocos ← incluir(independentes)
26  Blocos ← ordena(Blocos[comeco])
```

# Apêndice B

# Cálculo do Número de Vizinhanças

O número de vizinhanças e, consequentemente, o número de modelos que precisam ser ajustados, para um determinado SNP de interesse, depende de quantos marcadores vamos considerar para cada lado, dentro do campo Markoviano.

Assim, se para um determinado SNP, adotarmos, por exemplo, 2 marcadores para cada lado, teremos que considerar 9 diferentes vizinhanças, conforme mostra a Tabela B.1.

**Tabela B.1:** *Indicação do número de vizinhanças considerando um intervalo de 2 marcadores para cada lado do SNP de interesse.*

| Limite à esquerda ($l_j$) | Limite à direita ($r_j$) | Tamanho vizinhança |
|---|---|---|
| $j-2$ | $j+2$ | 4 marcadores |
| $j-2$ | $j+1$ | 3 marcadores |
| $j-2$ | $j$ | 2 marcadores |
| $j-1$ | $j+2$ | 3 marcadores |
| $j-1$ | $j+1$ | 2 marcadores |
| $j-1$ | $j$ | 1 marcador |
| $j$ | $j+2$ | 2 marcadores |
| $j$ | $j+1$ | 1 marcador |
| $j$ | $j$ | nenhum marcador (sem vizinhança) |

Vamos chamar o número de marcadores para cada lado do SNP de interesse como largura da vizinhança e denotar pela letra $W$. Assim, para um SNP, o número de diferentes tamanhos de vizinhança que precisam ser avaliados será $(W+1)^2$. No caso da Tabela B.1, $W=2$ e, portanto, o número de vizinhanças será 9.

Quando avaliamos uma sequência de SNP's, o número de vizinhanças a considerar será $(\#SPN's)\cdot(W+1)^2$, porém, se nesse intervalo de SNP's estiverem os marcadores iniciais e/ou finais da amostra, o número total precisa ser reduzido.

De fato, nas extremidades, teremos vizinhanças apenas em um dos lados. A Tabela B.2

mostra esse cálculo para os SNP's iniciais e finais, considerando uma amostra de 10 SNP's como exemplo.

**Tabela B.2:** *Número de vizinhanças a considerar para cada SNP, supondo uma largura de vizinhança igual a 2, em uma amostra de 10 SNP's.*

| # SNP | # vizinhanças |
|---|---|
| 1 | 3 |
| 2 | 6 |
| 3 | 9 |
| 4 | 9 |
| 5 | 9 |
| 6 | 9 |
| 7 | 9 |
| 8 | 9 |
| 9 | 6 |
| 10 | 3 |
| Total | 72 |

Assim, para cada extremidade, precisamos descontar uma quantidade de vizinhanças dada pela somatória:

$$\sum_{i=1}^{W}(W+1)\cdot i$$

Chamando então de $N$ o número de SNP's extremos dentro de uma sequência considerada, chegamos a:

$$\#\text{vizinhanças} = (\#SNP's)\cdot(W+1)^2 - N\cdot\sum_{i=1}^{W}(W+1)\cdot i. \tag{B.1}$$

Importante notar que é preciso haver pelo menos $W$ SNP's além do SNP de interesse, para considerar que ele não está no extremo da amostra. A Figura B.1 ilustra o valor de $N$ para os diferentes casos considerados, para uma sequência de 10 marcadores, inseridos em uma amostra.


**Figura B.1:** *Atribuição do valor de N para cada configuração de sequência de SNP's. No primeiro caso, os SNP's estão no meio da amostra e, portanto, $N = 0$. No segundo caso, os SNP's são os iniciais, logo $N = 1$. Finalmente, se os SNP's forem todos os marcadores da amostra, $N = 2$, uma vez que teremos que tratar as duas extremidades.*

Voltando então à expressão (B.1), temos que:

$$\sum_{i=1}^{W}(W+1)\cdot i = (W+1)\cdot\sum_{i=1}^{W} i,$$

e $\sum_{i=1}^{W} i$ é a soma dos termos de uma progressão aritmética de razão unitária, ou seja:

$$\sum_{i=1}^{W} i = \left[(1+W)\cdot\frac{W}{2}\right].$$

Substituindo na expressão (B.1) chegamos a:

$$\#\text{vizinhanças} = (W+1)^2\cdot(\#SNP's) - N\cdot(W+1)\cdot\left[(1+W)\cdot\frac{W}{2}\right],$$

e, finalmente:

$$\#\text{vizinhanças} = (W+1)^2\left[\#SNP's - N\cdot\frac{W}{2}\right], N\in\{0,1,2\}, \tag{B.2}$$

sendo que:

$N = 0 \Rightarrow$ a sequência de SNP's de interesse está longe pelo menos $W$ SNP's do início e $W$ SNP's do final da amostra;

$N = 1 \Rightarrow$ uma das extremidades da sequência de interesse coincide com o início ou com o final da amostra;

$N = 2 \Rightarrow$ a sequência de SNP's de interesse corresponde à toda amostra.

# Referências

[Andrade e Pinheiro 2002] Mariza Andrade e Hildete Prisco Pinheiro. *Métodos estatísticos aplicados em genética humana*. ABE, 2002 (citado nas pgs. 2, 5, 22, 24, 32).

[Avery e Henderson 1999] Peter J Avery e Daniel A Henderson. "Fitting markov chain models to discrete state series such as dna sequences". Em: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 48.1 (1999), pgs. 53–61 (citado na pg. 13).

[Balding *et al.* 2008] David J Balding, Martin Bishop e Chris Cannings. *Handbook of statistical genetics*. John Wiley & Sons, 2008 (citado na pg. 76).

[Bansal *et al.* 2008] Vikas Bansal, Aaron L Halpern, Nelson Axelrod e Vineet Bafna. "An mcmc algorithm for haplotype assembly from whole-genome sequence data". Em: *Genome research* 18.8 (2008), pgs. 1336–1346 (citado na pg. 5).

[Bates *et al.* 2015] Douglas Bates, Martin Mächler, Ben Bolker e Steve Walker. "Fitting linear mixed-effects models using lme4". Em: *Journal of Statistical Software* 67.1 (2015), pgs. 1–48. doi: 10.18637/jss.v067.i01 (citado na pg. 43).

[Besag 1975] Julian Besag. "Statistical analysis of non-lattice data". Em: *The statistician* (1975), pgs. 179–195 (citado na pg. 18).

[Bianchi 2009] André Jucovsky Bianchi. *Vizinhanças e Janelas de Influência em Polimorfismos de UmUnico Nucleotıdeo*. 2009 (citado nas pgs. 18, 19).

[Bitarello 2011] Bárbara Domingues Bitarello. "Seleção natural em genes HLA: uma investigação da localização molecular e temporal dos eventos de seleção". Tese de dout. Universidade de São Paulo, 2011 (citado nas pgs. xii, 77).

[Blangero *et al.* 2013] John Blangero, Vincent P Diego, Thomas D Dyer, Marcio Almeida, Juan Peralta, Jack W Kent Jr, Jeff T Williams, Laura Almasy e Harald HH Göring. "A kernel of truth: statistical advances in polygenic variance component models for complex human pedigrees". Em: *Advances in genetics*. Vol. 81. Elsevier, 2013, pgs. 1–31 (citado nas pgs. 8, 22, 31, 85).

[Bonat e Jørgensen 2016] Wagner Hugo Bonat e Bent Jørgensen. "Multivariate covariance generalized linear models". Em: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 65.5 (2016), pgs. 649–675 (citado nas pgs. 22, 23, 30).

[Bourgain *et al.* 2002] Catherine Bourgain, Emmanuelle Genin e Francoise Clerget-Darpoux. "Comparison of family based haplotype methods using intragenic snps in candidate genes". Em: *European Journal of Human Genetics* 10.5 (2002), pgs. 313–319 (citado na pg. 5).

[Breslow e Clayton 1993] Norman E Breslow e David G Clayton. "Approximate inference in generalized linear mixed models". Em: *Journal of the American statistical Association* 88.421 (1993), pgs. 9–25 (citado na pg. 31).

[Browning e Browning 2011] Sharon R Browning e Brian L Browning. "Haplotype phasing: existing methods and new developments". Em: *Nature Reviews Genetics* 12.10 (2011), pgs. 703–714 (citado na pg. 7).

[Brutti *et al.* 2019] Bruna Brutti, Jéssica Flores, Juliana Hermes, Giovana Martelli, Deise da Silva Porto e Elenir Terezinha Rizzetti Anversa. "Diabete mellitus: definição, diagnóstico, tratamento e mortalidade no brasil, rio grande do sul e santa maria, no período de 2010 a 2014". Em: *Brazilian Journal of Health Review* 2.4 (2019), pgs. 3174–3182 (citado na pg. 74).

[Bühlmann e Wyner 1999] Peter Bühlmann e Abraham J Wyner. "Variable length markov chains". Em: *The Annals of Statistics* 27.2 (1999), pgs. 480–513 (citado na pg. 13).

[Chen e Conomos 2016] H Chen e M Conomos. *Generalized Linear Mixed Model Association Tests*. 2016. url: https://content.sph.harvard.edu/xlin/dat/GMMAT_user_manual_v0.7.pdf (citado na pg. 23).

[Consortium 2015] 1000 Genomes Project Consortium. "A global reference for human genetic variation". Em: *Nature* 526.7571 (2015), pg. 68 (citado na pg. 7).

[Csiszár e Talata 2006] Imre Csiszár e Zsolt Talata. "Context tree estimation for not necessarily finite memory processes, via bic and mdl". Em: *IEEE Transactions on Information theory* 52.3 (2006), pgs. 1007–1016 (citado nas pgs. 19, 28).

[De Andrade *et al.* 2002] Mariza De Andrade, René Guéguen, Sophie Visvikis, Catherine Sass, Gérard Siest e Christopher I Amos. "Extension of variance components approach to incorporate temporal trends and longitudinal pedigree data analysis". Em: *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society* 22.3 (2002), pgs. 221–232 (citado na pg. 22).

[De Andrade *et al.* 2015] Mariza De Andrade, Debashree Ray, Alexandre C Pereira e Júlia P Soler. "Global individual ancestry using principal components for family data". Em: *Human heredity* 80.1 (2015), pgs. 1–11 (citado na pg. 84).

[Delaneau *et al.* 2019] Olivier Delaneau, Jean-François Zagury, Matthew R Robinson, Jonathan L Marchini e Emmanouil T Dermitzakis. "Accurate, scalable and integrative haplotype estimation". Em: *Nature communications* 10.1 (2019), pg. 5436 (citado na pg. 83).

[Ding *et al.* 2003] Keyue Ding, Kaixin Zhou, Fuchu He e Yan Shen. "Lda—a java-based linkage disequilibrium analyzer". Em: *Bioinformatics* 19.16 (2003), pgs. 2147–2148 (citado na pg. 4).

[Druet e Georges 2010] Tom Druet e Michel Georges. "A hidden markov model combining linkage and linkage disequilibrium information for haplotype reconstruction and quantitative trait locus fine mapping". Em: *Genetics* 184.3 (2010), pgs. 789–798 (citado nas pgs. 5, 83).

[Epstein *et al.* 2009] Michael P Epstein, Jessica E Hunter, Emily G Allen, Stephanie L Sherman, Xihong Lin e Michael Boehnke. "A variance-component framework for pedigree analysis of continuous and categorical outcomes". Em: *Statistics in biosciences* 1.2 (2009), pgs. 181–198 (citado na pg. 23).

[Fernandes 2016] Francisco José de Almeida Fernandes. "Caracterização da estrutura de dependência do genoma humano usando campos markovianos: estudo de populações mundiais e dados de SNPs". Diss. de mestr. Universidade de São Paulo, 2016 (citado nas pgs. xii, 5, 20, 21, 28, 77, 79, 81, 82).

[Galdino 2015] Maicon Vinícius Galdino. "Modelos lineares generalizados mistos e equações de estimação generalizadas para dados binário aplicados em anestesiologia veterinária". Em: (2015) (citado nas pgs. 23, 31).

[Geraci 2010] Filippo Geraci. "A comparison of several algorithms for the single individual snp haplotyping reconstruction problem". Em: *Bioinformatics* 26.18 (2010), pgs. 2217–2225 (citado nas pgs. 3, 83).

[Greenspan e Geiger 2006] G Greenspan e Dan Geiger. "Modeling haplotype block variation using markov chains". Em: *Genetics* 172.4 (2006), pgs. 2583–2599 (citado na pg. 5).

[Huggins 1993] RM Huggins. "On the robust analysis of variance components models for pedigree data". Em: *Australian Journal of Statistics* 35.1 (1993), pgs. 43–57 (citado na pg. 23).

[Jorde 1995] Lynn B Jorde. "Linkage disequilibrium as a gene-mapping tool." Em: *American journal of human genetics* 56.1 (1995), pg. 11 (citado na pg. 5).

[Kent *et al.* 2002] W James Kent, Charles W Sugnet, Terrence S Furey, Krishna M Roskin, Tom H Pringle, Alan M Zahler e David Haussler. "The human genome browser at ucsc". Em: *Genome research* 12.6 (2002), pgs. 996–1006 (citado na pg. 74).

[Kim *et al.* 2008] Yunjung Kim, Sheng Feng e Zhao-Bang Zeng. "Measuring and partitioning the high-order linkage disequilibrium by multiple order markov chains". Em: *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society* 32.4 (2008), pgs. 301–312 (citado nas pgs. 4, 5).

[Klein e Sato 2000] JAN Klein e Akie Sato. "The hla system". Em: *New England Journal of Medicine* 343.10 (2000), pgs. 702–709 (citado na pg. vii).

[Konishi e Rao 1992] Sadanori Konishi e C Radhakrishna Rao. "Principal component analysis for multivariate familial data". Em: *Biometrika* 79.3 (1992), pgs. 631–641 (citado na pg. 22).

[Laird e Lange 2010] Nan M Laird e Christoph Lange. *The fundamentals of modern statistical genetics*. Springer Science & Business Media, 2010 (citado nas pgs. 5, 22, 24, 84).

[Lange 2002] Kenneth Lange. *Mathematical and statistical methods for genetic analysis*. Springer Science & Business Media, 2002 (citado na pg. 4).

[Leal *et al.* 2005] Suzanne M Leal, Kai Yan e Bertram Müller-Myhsok. "Simped: a simulation program to generate haplotype and genotype data for pedigree structures". Em: *Human heredity* 60.2 (2005), pgs. 119–122 (citado nas pgs. 43, 82).

[Leonardi 2007] Florencia Graciela Leonardi. "Cadeias estocásticas parcimoniosas com aplicaçoesa classificaçao e filogenia das seqüências de proteınas". Tese de dout. Universidade de São Paulo, 2007 (citado na pg. 19).

[Lesk 2005] Arthur M Lesk. *Introdução à bioinformática*. Artmed, 2005 (citado nas pgs. 3–6).

[Liang e Zeger 1986] Kung-Yee Liang e Scott L Zeger. "Longitudinal data analysis using generalized linear models". Em: *Biometrika* 73.1 (1986), pgs. 13–22 (citado nas pgs. 23, 34).

[Löcherbach e Orlandi 2011] Eva Löcherbach e Enza Orlandi. "Neighborhood radius estimation for variable-neighborhood random fields". Em: *Stochastic Processes and their Applications* 121.9 (2011), pgs. 2151–2185 (citado nas pgs. 14, 29).

[McCullagh 1983] Peter McCullagh. "Quasi-likelihood functions". Em: *The Annals of Statistics* 11.1 (1983), pgs. 59–67 (citado na pg. 23).

[McCullagh 2019] Peter McCullagh. *Generalized linear models*. Routledge, 2019 (citado na pg. 36).

[Monteiro *et al.* 2004] Antônio Miguel Vieira Monteiro, G Câmara, MS Carvalho e S Druck. "Análise espacial de dados geográficos". Em: *Brasília: Embrapa* (2004) (citado na pg. 13).

[Ott e Rabinowitz 1999] Jürg Ott e Daniel Rabinowitz. "A principal-components approach based on heritability for combining phenotype information". Em: *Human heredity* 49.2 (1999), pgs. 106–111 (citado na pg. 22).

[Oualkacha *et al.* 2012] Karim Oualkacha, Aurelie Labbe, Antonio Ciampi, Marc-Andre Roy e Michel Maziade. "Principal components of heritability for high dimension quantitative traits and general pedigrees". Em: *Statistical Applications in Genetics and Molecular Biology* 11.2 (2012) (citado na pg. 22).

[Price *et al.* 2006] Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick e David Reich. "Principal components analysis corrects for stratification in genome-wide association studies". Em: *Nature genetics* 38.8 (2006), pgs. 904–909 (citado na pg. 84).

[R Core Team 2020] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2020. URL: https://www.R-project.org/ (citado nas pgs. 43, 85).

[R Development Core Team 2009] R Development Core Team. *R: A Language and Environment for Statistical Computing*. ISBN 3-900051-07-0. R Foundation for Statistical Computing. Vienna, Austria, 2009. URL: http://www.R-project.org (citado na pg. 23).

[Rissanen *et al.* 1983] Jorma Rissanen *et al.* "A universal data compression system". Em: *IEEE Transactions on information theory* 29.5 (1983), pgs. 656–664 (citado na pg. 14).

[Ross *et al.* 1996] Sheldon M Ross, John J Kelly, Roger J Sullivan, William James Perry, Donald Mercer, Ruth M Davis, Thomas Dell Washburn, Earl V Sager, Joseph B Boyce e Vincent L Bristow. *Stochastic processes*. Vol. 2. Wiley New York, 1996 (citado na pg. 13).

[Schlesinger 2010] David Schlesinger. "Ancestralidade da população de São Paulo e correlação com alterações neuropatológicas no idoso". Tese de dout. Universidade de São Paulo, 2010 (citado na pg. 21).

[Sherry *et al.* 2001] Stephen T Sherry, M-H Ward, M Kholodov, J Baker, Lon Phan, Elizabeth M Smigielski e Karl Sirotkin. "Dbsnp: the ncbi database of genetic variation". Em: *Nucleic acids research* 29.1 (2001), pgs. 308–311 (citado nas pgs. 74, 76).

[uk 2005] International HapMap Consortium Altshuler David altshuler@ molbio. mgh. harvard. edu Donnelly Peter donnelly@ stats. ox. ac. uk. "A haplotype map of the human genome". Em: *Nature* 437.7063 (2005), pgs. 1299–1320 (citado na pg. 20).

[Urakami 2019] Tatsuhiko Urakami. "Maturity-onset diabetes of the young (mody): current perspectives on diagnosis and treatment". Em: *Diabetes, metabolic syndrome and obesity: targets and therapy* 12 (2019), pg. 1047 (citado na pg. 74).

[Wang e Thompson 2019] Bowen Wang e Elizabeth Thompson. "Realized genome sharing in heritability estimation using random effects models". Em: *G3: Genes, Genomes, Genetics* 9.5 (2019), pgs. 1385–1391 (citado nas pgs. 53, 83).

[Wang *et al.* 2015] Tao Wang, Peng He, Kwang Woo Ahn, Xujing Wang, Soumitra Ghosh e Purushottam Laud. "A re-formulation of generalized linear mixed models to fit family data in genetic association studies". Em: *Frontiers in genetics* 6 (2015), pg. 120 (citado na pg. 23).

[Wedderburn 1974] Robert WM Wedderburn. "Quasi-likelihood functions, generalized linear models, and the gauss—newton method". Em: *Biometrika* 61.3 (1974), pgs. 439–447 (citado na pg. 23).

[Zanin Zambom *et al.* 2022] Adriano Zanin Zambom, Seonjin Kim e Nancy Lopes Garcia. "Variable length markov chain with exogenous covariates". Em: *Journal of Time Series Analysis* 43.2 (2022), pgs. 312–328 (citado na pg. 83).

[Ziyatdinov *et al.* 2018] Andrey Ziyatdinov, Miquel Vazquez-Santiago, Helena Brunel, Angel Martinez-Perez, Hugues Aschard e Jose Manuel Soria. "Lme4qtl: linear mixed models with flexible covariance structure for genetic studies of related individuals". Em: *BMC Bioinformatics* (2018), btw080. URL: http://dx.doi.org/10.1186/s12859-018-2057-x (citado nas pgs. 43, 51).