UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Métodos eficientes para colapsagem e seleção de SNPs raros em dados familiares

Ana Fernanda Noli

Dissertação de Mestrado do Programa Interinstitucional de Pós-Graduação em Estatística (PIPGEs)

ICMC USP SÃO CARLOS

Ana Fernanda Noli

Métodos eficientes para colapsagem e seleção de SNPs raros em dados familiares

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP e ao Departamento de Estatística – DEs-UFSCar, como parte dos requisitos para obtenção do título de Mestra em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística. *VERSÃO REVISADA*

Área de Concentração: Estatística

Orientadora: Profa. Dra. Daiane Aparecida Zuanetti

USP – São Carlos
Agosto de 2024

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

N792m

Noli, Ana Fernanda
Métodos eficientes para colapsagem e seleção de SNPs raros em dados familiares / Ana Fernanda Noli; orientadora Daiane Aparecida Zuanetti. -- São Carlos, 2024.
84 p.

Dissertação (Mestrado - Programa Interinstitucional de Pós-graduação em Estatística) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2024.

1. GWAS. 2. Lasso. 3. Modelo linear misto. 4. NGS. I. Zuanetti, Daiane Aparecida, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Ana Fernanda Noli

Efficient methods for collapsing and selecting rare SNPs in family-based data

Dissertation submitted to the Institute of Mathematics and Computer Science – ICMC-USP and to the Department of Statistics – DEs-UFSCar – in accordance with the requirements of the Statistics Interagency Graduate Program, for the degree of Master in Statistics. *FINAL VERSION*

Concentration Area: Statistics

Advisor: Profa. Dra. Daiane Aparecida Zuanetti

USP – São Carlos
August 2024

Dedico este trabalho às pessoas mais importantes da minha vida, cujo amor e apoio foram fundamentais em minha jornada. Ao meu marido, Nícolas, por seu carinho e paciência inabaláveis. À minha mãe, Maria Neide, cuja força e dedicação são minha eterna inspiração. Aos meus irmãos, Bruno, Paula e Lila, por sempre acreditarem em mim. E aos meus sobrinhos, Giovanni, Luiza e Violeta, que trazem alegria e leveza aos meus dias. Agradeço também aos meus amigos, que me acompanharam e me ofereceram seu apoio e amizade em cada passo deste caminho. Por fim, dedico este esforço também a mim mesma, por perseverar diante dos desafios e completar esta etapa da minha vida.

A todos vocês, meu profundo agradecimento e amor.

AGRADECIMENTOS

Agradeço a Deus por me dar força e inspiração todos os dias, guiando meu caminho com luz e sabedoria durante esta jornada.

Gostaria de expressar minha profunda gratidão à Professora Daiane, minha orientadora, por sua inestimável orientação, paciência e amizade ao longo deste processo. Seu apoio foi fundamental para o desenvolvimento e conclusão desta pesquisa.

Agradeço imensamente a Francisco e Vitor, profissionais que foram essenciais não apenas para o sucesso deste trabalho, mas também por me ajudarem a ver o melhor em mim mesma durante os momentos mais desafiadores.

Meus agradecimentos também se estendem aos amigos do mestrado, especialmente Flavia e Rodrigo, e a todos os demais colegas que compartilharam comigo esta jornada acadêmica. Suas companhias, apoio e discussões enriqueceram enormemente minha experiência.

Sou extremamente grata aos meus amigos antigos Amanda, Gabriela e Nemer, por sempre estarem ao meu lado, me encorajando e celebrando cada conquista. Aos novos amigos, especialmente Vanessa, que trouxe novos ares e companheirismo durante os últimos meses, meu sincero agradecimento.

Sou grata aos professores e funcionários do departamento, cujo conhecimento e assistência foram vitais para minha formação acadêmica e pessoal. A dedicação de vocês ao ensino e ao bem-estar dos estudantes não passou despercebida e é sinceramente apreciada.

Por fim, agradeço a todas as pessoas que, direta ou indiretamente, contribuíram para a realização deste trabalho. Este sucesso é também de vocês.

RESUMO

NOLI, A. F. **Métodos eficientes para colapsagem e seleção de SNPs raros em dados familiares**. 2024. 84 p. Dissertação (Mestrado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2024.

O presente trabalho apresenta métodos para a identificação de variantes raras, que consistem em variações genéticas que ocorrem com baixa frequência na população, para a compreensão da saúde humana e como o sequenciamento de próxima geração (NGS, do inglês next-generation sequencing) permitiu uma avaliação mais completa da variação genética dos indivíduos. A detecção da variação genômica é feita por meio de marcadores SNP (Polimorfismos de Nucleotídeo Único). A abordagem do modelo linear misto (MLM) destaca-se nesses estudos por abranger efeitos fixos e aleatórios e, o método LASSO (do inglês Least Absolute Shrinkage and Selection Operator), para a seleção de variáveis. Este trabalho descreve, ainda, as abordagens mais recentes para estudos do tipo GWAS (do inglês *genome-wide association studies*) e NGS que abrangem dados de família e como variantes raras desempenham um papel importante na arquitetura genética de distúrbios complexos. Por fim, discutem-se as principais metodologias utilizadas nos estudos genômicos com base em colapsagem e a necessidade de trazer novas técnicas ou combiná-las para melhor compreender a influência de variantes raras em doenças complexas.

Palavras-chave: GWAS, Lasso, Modelo linear misto, NGS.

ABSTRACT

NOLI, A. F. **Efficient methods for collapsing and selecting rare SNPs in family-based data**. 2024. 84 p. Dissertação (Mestrado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2024.

The present work presents methods for identifying rare variants, which consist of genetic variations that occur with low frequency in the population, for understanding human health and how next-generation sequencing (NGS) has enabled a more comprehensive assessment of individuals' genetic variation. The detection of genomic variation is performed using SNP markers (Single Nucleotide Polymorphism). The linear mixed model (LMM) approach stands out in these studies for encompassing both fixed and random effects, and the LASSO (Least Absolute Shrinkage and Selection Operator) method is used for variable selection. This work also describes the most recent approaches for GWAS (genome-wide association studies) and NGS studies that include family data and how rare variants play an important role in the genetic architecture of complex disorders. Finally, the main methodologies used in genomic studies based on collapsing are discussed, as well as the need to bring new techniques or combine them to better understand the influence of rare variants on complex diseases.

Keywords: GWAS, Lasso, Linear Mixed Model, NGS.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

CAPÍTULO

1

INTRODUÇÃO

Ao longo dos últimos anos, o interesse na identificação de variantes raras que influenciam fenótipos de seres vivos levou ao desenvolvimento de muitos métodos estatísticos para testar a associação entre conjuntos de SNPs raros e características binárias ou quantitativas. Compreender o impacto dessas variantes e SNPs é essencial para entender a saúde e o aperfeiçoamento de técnicas estatísticas aliado à sua eficiência computacional vêm promovendo enorme avanço na análise de fenótipos complexos.

Antes do sequenciamento de próxima geração (do inglês *next-generation sequencing*, NGS), a análise de características complexas focava nas contribuições de marcadores (sejam eles microssatélites, SNPs, entre outros) comuns usando métodos de estudo de associação genômica ampla (do inglês *genome-wide association study*, GWAS). Esses estudos identificaram com sucesso muitas regiões genômicas significativas relacionadas a vários fenótipos complexos. No entanto, à medida que os tamanhos das amostras de estudo cresceram e o sequenciamento genético ficou mais barato e preciso, variantes recém-identificadas tiveram efeitos menores sobre os fenótipos. Os avanços na tecnologia NGS transformaram, portanto, a genética e biologia molecular na última década, permitindo uma avaliação mais completa da variação genética de um indivíduo, incluindo a análise de variantes raras (POVYSIL *et al.*, 2019).

Mais recentemente, a detecção da variação genômica baseia-se em marcadores SNP (*single nucleotide polymorphism* ou variante genética comum ou polimorfismo de nucleotídeo único) que consiste em uma variação na sequência de DNA, que afeta somente uma base nitrogenada na sequência do genoma entre indivíduos de uma espécie ou entre pares de cromossomos de um indivíduo. É uma variação em uma posição específica de DNA, que também está presente em uma proporção da população. Estas variações podem ou não causar alteração do fenótipo, mas podem também ser causa de resistência ou suscetibilidade para determinadas doenças, bem como de resposta a medicamentos.

Os estudos GWAS geralmente tem como alvo SNPs comuns. A frequência do alelo menor

(do inglês *minor allele frequency*, MAF) é a frequência relativa (ou proporção) na qual o segundo alelo mais comum ocorre em uma determinada população e é amplamente utilizado em estudos de genética populacional, porque fornece informações para diferenciar entre SNPs comuns e raros. Os raros são geralmente definidos como tendo uma frequência de alelo menor (MAF) inferior a 5%, 1% ou 0.5%. A esperança em estudos de SNPs raros é que estes forneçam uma visão complementar de fatores de risco de doenças ou outras características físicas e um novo conjunto de alvos para pesquisas translacionais[1] (NICOLAE, 2016). Em modelos tradicionais de GWAS, o efeito de SNPs raros, geralmente não identificado por técnicas usuais de modelagem e seleção, é considerado conjuntamente como um único efeito (geralmente aleatório), conhecido como efeito poligênico. No entanto, se torna cada vez mais necessário identificar individualmente ou em grupos menores de SNPs, o efeito delas.

Variações raras se devem, geralmente, a dois motivos. Um deles é o fato da variante ser jovem ou recente, ou seja, que surgiu em um indivíduo pela primeira vez e não é herdada de um dos pais e por isso estão pouco presentes na população. Elas podem ocorrer por acaso, por falhas no processo de replicação do DNA durante a divisão celular e formação dos gametas, por falhas no sistema de reparo de erros no processamento do DNA ou devido a eventos mutagênicos, como exposição à radiação e produtos químicos específicos. Outras variantes, apesar de já existirem há bastante tempo, podem estar segregando em poucas famílias, e por estarem restritas a indivíduos de poucas famílias são raras na população.

Para a análise de dados familiares, em específico, quando existe grau de parentesco entre os indivíduos da amostra, é indispensável a utilização de técnicas estatísticas que contemplem indivíduos (e observações) não independentes.

A abordagem do modelo linear misto (MLM) destaca-se nesses estudos por abranger efeitos fixos e aleatórios (JIANG; NGUYEN, 2007). O controle das associações de origem de parentesco se dá através de uma matriz de relacionamento genético atrelada a um vetor de efeitos aleatórios (cujo objetivo é capturar os efeitos do parentesco entre os indivíduos).

Uma das abordagens mais completas para estudos do tipo GWAS que abrange dados de família é proposto no *fastGWA* (JIANG *et al.*, 2019). Trata-se de modelo linear misto com o controle da estratificação da população via componentes principais de ancestralidade. Essa metodologia seleciona os SNPs comuns mais relevantes usando uma métrica derivada da função de verossimilhança, via teste de significância. Outras derivações de modelos lineares mistos que também têm sido utilizadas para seleção de SNPs comuns e raros em dados familiares são o *famSKAT-RC*, do inglês *family sequence kernel association test for rare and common variants* (SAAD; WIJSMAN, 2014), e o *Grid-LMM* (RUNCIE; CRAWFORD, 2019), sendo a principal

[1] A pesquisa translacional envolve um processo conhecido por translação do conhecimento, no qual o foco da investigação científica na área da saúde adquire um fluxo de informação bidirecional entre as pesquisas básica e clínica, resultando em uma aplicabilidade real do conhecimento e de novas tecnologias, melhorando a aplicação clínica de novos conceitos terapêuticos, e, finalmente, proporcionando benefícios diretos aos principais interessados neste processo: os pacientes.

vantagem desse último método em relação aos demais a possibilidade de definir vários efeitos aleatórios que não apenas o intercepto.

Um tema comum em estudos NGS recentes é que variações raras e, em particular, aquelas consideradas “ultra-raras” na população - ou seja, não observadas em coortes de referência disponíveis e provavelmente de origem muito jovem - também desempenham um papel importante na arquitetura genética de distúrbios complexos. Uma das principais motivações para estudos de variantes raras é baseada no argumento de que elas geralmente têm efeitos maiores do que os SNPs comuns ou de baixa frequência associados ao fenótipo (NICOLAE, 2016).

Estudos recentes de SNPs raros não apenas levaram à identificação de genes que mostram associação significativa definitiva do genoma com doenças e outras características físicas, mas também forneceram informações sobre questões-chave que, de outra forma, seriam difíceis de abordar, incluindo a identificação de variantes específicas que contribuem para o risco de doenças, e avaliar aspectos da arquitetura genética (ASIMIT; ZEGGINI, 2010; LETTRE, 2014; MÄGI *et al.*, 2012; BOMBA; WALTER; SORANZO, 2017; PETROVSKI *et al.*, 2017; ALLEN *et al.*, 2017; CIRULLI *et al.*, 2020; CIRULLI *et al.*, 2019).

Nos concentramos nesse trabalho nas principais metodologias utilizadas nos estudos genômicos GWAS e NGS que visam compreender o efeito de SNPs comuns e raros em determinados fenótipos ou outras características. Veremos que os esforços empregados em trazer novas técnicas são indispensáveis, pois as metodologias de GWAS não performam bem o suficiente na presença de SNPs raros. Os autores Ionita-Laza *et al.* (2014) asseveram que em qualquer estudo envolvido com SNPs raros, a maioria deles é observada apenas um pequeno número de vezes. Essa natureza esparsa dos dados apresenta dificuldades estatísticas não triviais, e os métodos estatísticos tradicionais empregados para testes de associação não são poderosos nesse contexto (SUN *et al.*, 2011).

Além disso, traremos no decorrer desse estudo algumas abordagens de colapsagem baseada em genes, nas quais SNPs que satisfazem critérios específicos (também chamadas de SNPs de qualificação) são agrupados como equivalentes para identificar contribuições de SNPs raros para doenças ou outras características em estudo (POVYSIL *et al.*, 2019).

Este estudo está organizado da seguinte maneira. No Capítulo 2, discutimos os fundamentos genômicos e apresentamos as principais metodologias para estudar associações de variações genéticas. No Capítulo 3, detalhamos as metodologias para colapsagem, tais como SKAT e Burden, procedimento de eliminação backward, modelo hierárquico para estimar as razões de chance de SNPs raros individuais, e o uso conjunto do procedimento de eliminação backward e modelo hierárquico. No Capítulo 4, descrevemos o banco de dados GAW17 utilizado em nossas análises, o tratamento dos dados, a métrica de desempenho para seleção de SNPs e a criação da variável de colapsagem. No Capítulo 5, apresentamos os resultados das diferentes abordagens previamente discutidas. Finalmente, no Capítulo 6, discutimos os achados e apresentamos as conclusões do estudo.

CAPÍTULO

2

METODOLOGIAS PARA SELEÇÃO DE SNPS

Neste capítulo, apresentamos alguns conceitos genéticos e as principais metodologias existentes para estudos de associação genômica de SNPs com características físicas de interesse.

2.1 Definição de SNP

Um SNP (do inglês *Single Nucleotide Polymorphism*) é definido como a variação encontrada em um único nucleotídeo ou outra sequência compartilhada em um indivíduo da mesma espécie. Em outras palavras, cada SNP representa uma variação em uma única base nitrogenada no DNA. Em uma via de DNA, um nucleotídeo é composto por um grupo açúcar, grupo fosfato e uma base nitrogenada. Os quatro tipos de base nitrogenada geralmente se referem à adenina (A), timina (T), citosina (C) e guanina (G). Um SNP é identificado, por exemplo, quando o nucleotídeo citosina (C) é trocado pelo nucleotídeo timina (T). Aqui, os SNPs são as variações genéticas ocorridas e caracterizam alterações de um único par de bases na sequência de DNA.

Esses SNPs têm grande potencial como marcadores moleculares na análise de mapeamento (GRIFFITHS *et al.*, 2009) e são essenciais em estudos relacionados a genes, pois estão fortemente associados a certas doenças tais como diabetes e doenças cardíacas, além de outras características físicas, tais como peso, pressão arterial, entre outras. Para os estudos genéticos, os SNPs são geralmente usados para marcar uma região genômica, com a grande maioria deles tendo um impacto nos sistemas biológicos. SNPs abundantes divergem de SNPs raros no que tange a características raras complexas, que são geralmente causadas por variantes genéticas extremamente raras. Esse tipo de ocorrência eventualmente traz mudanças adversas na função da proteína, que então atua como o caminho para o estado de uma doença ou característica complexa.

Estima-se que um SNP ocorra em uma frequência de 1 em 1000 pares de bases na sequência de DNA. SNPs podem cair em sequências codificantes de genes, regiões não codificantes de

genes ou nas regiões intergênicas (regiões entre genes). No entanto, uma vez que apenas 2.5% de todo código genético é uma região codificante (que carrega a informação para a síntese de proteínas), a maioria dos SNPs ocorrem em regiões não codificantes. Logo, grande parte dessas mutações podem não causar alterações significativas no organismo. Importante também ressaltar que SNPs dentro de uma sequência de codificação não alteram necessariamente a sequência de aminoácidos da proteína que é produzida, como veremos adiante (POVYSIL *et al.*, 2019).

Um SNP pode ser classificado conforme a troca de base como:

- **Transição:** quando ocorre a troca de uma purina por outra purina (Adenina e Guanina) ou uma pirimidina por outra pirimidina (Citocina e Timina); ou

- **Transversão:** quando ocorre a substituição de uma purina por uma pirimidina e vice-versa, por exemplo A por C.

SNPs na região de codificação são de dois tipos: SNPs sinônimos (*synonymous*) e não-sinônimos (*non-synonymous*). SNPs sinônimos, também conhecidos como SNPs silenciosos, não afetam a sequência da proteína, enquanto SNPs não-sinônimos, alteram a sequência de aminoácidos da proteína e podem fazer com que ela perca ou diminua sua função. Dentre os SNPs não-sinônimos, encontramos dois tipos:

- ***nonsense*:** quando a alteração de um nucleotídeo codifica para o encerramento da tradução, finalizando prematuramente a síntese de uma proteína (códon de parada prematuro), também conhecido como *stop-gain*; ou

- ***missense*:** quando a alteração de um nucleotídeo modifica a codificação, sintetizando um aminoácido diferente e podendo afetar a função da proteína. Alguns SNPs dessa categoria podem ser benignos, enquanto outros podem ter efeitos significativos na estabilidade, atividade ou interações de proteínas com outras moléculas.

A frequência do alelo menor em uma específica região do DNA, do inglês *minor allele frequency* (MAF) é a frequência relativa ou proporção na qual o segundo alelo mais comum ocorre em uma determinada população nessa específica região. Essa medida é amplamente utilizada em estudos genômicos porque fornece informações para diferenciar variantes (alelos) comuns de variantes raras na população. A maioria dos autores considera como raros os SNPs que apresentam MAF≤ 0.01. Alguns consideram raros as SNPs que apresentam MAF≤ 0.05. Outros autores consideram ainda a divisão em SNPs comuns (MAF ≥ 0.05), SNPs de baixa frequência ($0.01 <$MAF< 0.05) e SNPs raros (MAF≤ 0.01). Nesse estudo, consideramos como raros os SNPs com MAF≤ 0.05.

2.2 Metodologias estatísticas

Avanços recentes aprimoram constantemente a capacidade de técnicas baseadas em classificação e regressão para analisar SNPs comuns ou raros aliados à interação gene-ambiente (fatores ambientais) e identificar os que estão associados à presença de uma específica doença ou característica física (JIANG *et al.*, 2019).

Esta análise, geralmente complexa, pode ser realizada utilizando métodos de associação genômica ampla (GWAS, do inglês *Genome Wide Association Studies*), metodologias de colapsagem (na presença de muitos SNPs raros), classificadores e técnicas de ensemble[1], todos cada vez mais utilizados para analisar o polimorfismo de nucleotídeo único (SNP).

Nesta seção, vamos apresentar o modelo linear misto e métodos derivados dele utilizados tradicionalmente em estudos genômicos com dados familiares, como o *FamSKAT-RC* (SAAD; WIJSMAN, 2014), *Grid-LMM* (RUNCIE; CRAWFORD, 2019) e *fastGWA* (JIANG *et al.*, 2019).

2.2.1 Modelo linear misto

O modelo linear misto (MLM) assemelha-se ao modelo de regressão linear, porém com um termo adicional de efeitos aleatórios (JIANG; NGUYEN, 2007), ou seja, o modelo misto é um modelo de regressão que contempla efeitos fixos e aleatórios, sendo que os últimos possuem uma distribuição de probabilidade associada. Ao contemplar esses efeitos, os parâmetros de efeito fixo são comuns a todas as unidades amostrais, ao passo que os efeitos aleatórios, por terem uma distribuição associada, podem assumir diferentes valores entre as unidades da amostra.

Essa característica inerente a esse modelo faz com que ele seja útil para descrever dados que apresentem correlação entre medidas de uma mesma unidade amostral ou com outro tipo de dependência entre as unidades, como medidas obtidas de indivíduos relacionados (familiares, por exemplo), caso muito frequente nos dados de sequenciamento genético. Através dos efeitos aleatórios também podemos modelar e descrever comportamentos específicos das unidades amostrais, geralmente não capturados pelos efeitos fixos.

Os efeitos fixos são tratados como parâmetros desconhecidos e comuns a todas as observações. Logo, para as variáveis que consideramos não afetar a variável resposta em cada unidade de maneira diferente, a melhor forma de tratar seus efeitos é considerá-los como fixos. Já o efeito aleatório pode diferir para cada unidade amostral e, devido a isso, estes geralmente são os efeitos das variáveis que possuem impacto diferente em cada unidade, como os efeitos decorrentes do parentesco.

Para análise de dados genômicos, em especial, a notação do modelo linear misto é geralmente adaptada e escrita em função de termos genéticos. Almasy e Blangero (1998) traz

[1] Técnicas que visam melhorar a precisão dos resultados combinando vários modelos. Exemplos de técnicas ensemble tradicionais são: *bagging*, *boosting*, *Bayesian model averaging, Bayesian model combination*, entre outros.

uma das notações mais tradicionalmente utilizadas. Assumindo apenas o intercepto como efeito aleatório, sendo esta a situação mais comum em análise de dados genéticos, e seja Y_i o valor de um fenótipo quantitativo para o i-ésimo indivíduo ou unidade amostral, o modelo linear misto pode ser escrito como

$$Y_i = \mathbf{x}_{0i}\boldsymbol{\alpha}_0 + \mathbf{x}_i\boldsymbol{\beta} + u_i + \varepsilon_i,\ i = 1,\ldots,n, \tag{2.1}$$

em que $\boldsymbol{\alpha}_0$ é um vetor de parâmetros (fixos) que medem o efeito de variáveis ambientais, de comportamento ou ancestralidade presentes em $\mathbf{x}_{0i} = (x_{0i1},\ldots,x_{0ib})$ no valor esperado de Y_i, $\boldsymbol{\beta}$ é outro vetor de parâmetros (fixos) que medem o efeito das variáveis genéticas presentes em $\mathbf{x}_i = (x_{i1},\ldots,x_{ip})$ para a i-ésima unidade amostral, u_i é o intercepto aleatório que reflete o comportamento individual do fenótipo do i-ésimo indivíduo e o relaciona com o comportamento do fenótipo de seus familiares e ε_i é o erro aleatório associado à i-ésima unidade amostral. Em geral, assumimos que $u_i \sim Normal(0,\sigma_g^2)$ e $\varepsilon_i \sim Normal(0,\sigma_\varepsilon^2)$ e, u_i independente de ε_i.

Se considerarmos que não existem variáveis ambientais, de comportamento ou ancestralidade disponíveis para a análise, o Modelo (2.1) pode ser definido matricialmente como

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} + \boldsymbol{\varepsilon} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}, \tag{2.2}$$

em que $\mathbf{Z}$ é uma matriz identidade de dimensão $(n \times n)$, $\boldsymbol{u} \sim Normal_n(\mathbf{0}, \mathbf{G} = \sigma_g^2\Pi = \sigma_g^2 2\Phi)$, $\boldsymbol{\varepsilon} \sim Normal_n(\mathbf{0}, \mathbf{R} = \sigma_\varepsilon^2\mathbf{I})$ e $\boldsymbol{u}$ e $\boldsymbol{\varepsilon}$ são independentes. Nesse caso,

- $\mathbf{0}$ é um vetor de zeros de dimensão $(n \times 1)$;
- $\mathbf{G}$ é a matriz de variância e covariância dos efeitos aleatórios. Como há apenas os interceptos como efeitos aleatórios, é uma matriz com dimensão $(n \times n)$ e cada elemento dela representa o quanto o intercepto e, consequentemente, a variável resposta de cada par de unidades amostrais estão correlacionados;
- $\Pi = 2\Phi$ é a matriz com os coeficientes de relacionamento entre os pares de unidades amostrais que será melhor discutida posteriormente, e σ_g^2 é a variância genética associada ao fenótipo em estudo; e
- $\mathbf{R}$ é a matriz de variância e covariância do erro aleatório com dimensão $(n \times n)$, assumida como $\sigma_\varepsilon^2\ \mathbf{I}$, sendo σ_ε^2 a variância constante dos erros e $\mathbf{I}$, a matriz identidade de dimensão $(n \times n)$.

Uma outra maneira de representar o Modelo (2.2) na genética é considerar os interceptos aleatórios como independentes entre si e a matriz $\mathbf{Z} = \Pi^{1/2} = (2\Phi)^{1/2}$ ou $\mathbf{Z}$ como a matriz triangular inferior da decomposição de Cholesky de Π. Observe que em ambas as representações, a distribuição marginal do vetor $\mathbf{Y}$ é a mesma e dada por $\mathbf{Y} \sim Normal_n(\mathbf{X}\boldsymbol{\beta}, \sigma_g^2 2\Phi + \sigma_\varepsilon^2\mathbf{I})$.

A matriz de relacionamento Π é igual a duas vezes a matriz de parentesco Φ que, por sua vez, é definida como

$$\Phi = \begin{pmatrix} \frac{1}{2} & \frac{1}{2}^{r_{12}} & \cdots & \frac{1}{2}^{r_{1n}} \\ \frac{1}{2}^{r_{21}} & \frac{1}{2} & \cdots & \frac{1}{2}^{r_{2n}} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{1}{2}^{r_{n1}} & \frac{1}{2}^{r_{n2}} & \cdots & \frac{1}{2} \end{pmatrix},$$

em que r_{ij} é o grau de parentesco entre os indivíduos i e j. Por exemplo, $r_{ij} = 1$ se irmãos gêmeos idênticos, $r_{ij} = 2$ se pai e filho ou irmãos completos, $r_{ij} = 3$ se avô e neto, e assim sucessivamente. O coeficiente de parentesco entre dois indivíduos, i e j, é representado por ϕ_{ij}. O coeficiente de parentesco entre um indivíduo não endogâmico e ele próprio, ϕ_{ii}, é igual a $\frac{1}{2}$. Isso se deve ao fato de que os humanos são diploides, o que significa que a única maneira de os alelos escolhidos aleatoriamente serem idênticos por descendência é se o mesmo alelo for escolhido duas vezes (probabilidade $\frac{1}{2}$).

Sob essa formulação, as covariâncias entre as variáveis respostas (fenótipos) dos indivíduos i e j são dadas por:

- $Cov(Y_i, Y_i) = Var(Y_i) = \sigma_g^2 + \sigma_\varepsilon^2$;
- $Cov(Y_i, Y_j) = \sigma_g^2 2\phi_{ij}$ se $i \neq j$ e parentes em algum grau; e
- $Cov(Y_i, Y_j) = 0$ se $i \neq j$ e não parentes.

Se a estrutura de parentesco entre as unidades amostrais não estiver disponível, a matriz de relacionamento pode ser estimada através dos marcadores genéticos (geralmente considerando SNPs comuns) disponíveis na amostra (a GRM-SNP-*derived genetic relationship matrix*). Vários métodos para calcular a matriz de parentesco com base em marcadores genéticos estão disponíveis (AMADEU *et al.*, 2023), sendo um dos mais comuns o IBS (do inglês, *Identical-by-state*) que calcula uma medida de associação entre os genótipos para cada par de indivíduos usando o número de alelos compartilhados entre eles em cada SNP.

No entanto, o modelo linear misto pode acomodar vários outros efeitos aleatórios que não apenas o intercepto. Nesse caso, novamente considerando que não existem variáveis ambientais, de comportamento ou ancestralidade disponíveis para a análise, o modelo pode ser escrito compactamente como:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{u} + \boldsymbol{\varepsilon}, \tag{2.3}$$

sendo que $\boldsymbol{u} \sim Normal_{nq}(\mathbf{0}, \mathbf{G})$ e $\boldsymbol{\varepsilon} \sim Normal_n(\mathbf{0}, \mathbf{R})$. Ao assumirmos $\boldsymbol{u}$ e $\boldsymbol{\varepsilon}$ independentes, teremos $\mathbb{E}(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$ e $\mathbb{V}(\mathbf{Y}) = \mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}^T + \mathbf{R}$ em que

- $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ é um vetor de tamanho $(n \times 1)$ de fenótipos, em que n é o tamanho amostral;

- $\boldsymbol{\beta}$ é o vetor de coeficientes fixos das variáveis genotípicas com dimensão $(p \times 1)$, em que p é o número de variáveis genotípicas consideradas nos efeitos fixos;
- $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$ é a matriz de planejamento de $\boldsymbol{\beta}$ codificada conforme o genótipo dos SNPs para cada unidade amostral com dimensão $(n \times p)$;
- $\boldsymbol{u} = (\boldsymbol{u}_1^T, \ldots, \boldsymbol{u}_n^T)^T$ é o vetor de coeficientes aleatórios com dimensão $(nq \times 1)$, em que q é o número de variáveis consideradas nos efeitos aleatórios;
- $\mathbf{Z}$ é a matriz de planejamento bloco diagonal de $\boldsymbol{u}$ com dimensão $(n \times nq)$. Caso apenas o intercepto seja considerado como efeito aleatório, como visto anteriormente, essa matriz pode ser a identidade;
- $\boldsymbol{\varepsilon}$ é o vetor dos erros aleatórios com dimensão $(n \times 1)$;
- $\mathbf{G}$ é a matriz de variância e covariância dos efeitos aleatórios, definida com base no grau de parentesco entre os indivíduos no caso de dados genéticos e assumindo variâncias constantes para cada efeito. Caso haja apenas o intercepto nos efeitos aleatórios, como visto anteriormente, ela será uma matriz de variância e covariância com dimensão $(nq \times nq)$, em muitos casos a $\sigma_g^2 \Pi$;
- $\mathbf{R}$ é a matriz de variância e covariância do erro aleatório com dimensão $(n \times n)$, geralmente assumida como $\sigma_\varepsilon^2\ \mathbf{I}$, em que $\mathbf{I}$ é a matriz identidade $(n \times n)$ e σ_ε^2 a variância constante dos erros; e
- $\mathbf{0}$ é um vetor de zeros.

A distribuição marginal de $\mathbf{Y}$ é $Normal_n(\mathbf{X}\boldsymbol{\beta}, \mathbf{ZGZ}^T + \mathbf{R})$ e sua distribuição condicional é, assumindo $\mathbf{u}$ conhecido, $Normal_n(\mathbf{X}\boldsymbol{\beta} + \mathbf{Zu}, \mathbf{R})$.

Na perspectiva frequentista de estimação, os parâmetros são fixos, porém, desconhecidos. Uma forma de estimação recai sobre o método de mínimos quadrados generalizados (GLS) que contempla qualquer estrutura não singular da matriz $\mathbf{V} = \mathbb{V}(\mathbf{Y})$ e consiste em encontrar o valor de $\boldsymbol{\beta}$ que minimize a expressão $(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T \mathbf{V}^{-1} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$ (DEMIDENKO, 2013). Assumindo como conhecidos $\mathbf{G}$ e $\mathbf{R}$, a estimação simultânea dos efeitos fixos e predição dos efeitos aleatórios pode ser obtida pelas equações dadas por:

$$\begin{bmatrix} \mathbf{X}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}^T\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^T\mathbf{R}^{-1}\mathbf{Y} \\ \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Y} \end{bmatrix}.$$

Essas equações são conhecidas como equações de Henderson. Resolvendo esse sistema chegamos nos seguintes estimadores para os efeitos fixos e aleatórios, respectivamente,

$$\hat{\boldsymbol{\beta}} = [\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}]^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{Y}$$

$$\hat{\mathbf{u}} = \mathbf{G}\mathbf{Z}^T\mathbf{V}^{-1}[\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}] = \mathbf{G}\mathbf{Z}^T\mathbf{Q}\mathbf{Y},$$

em que $\mathbf{Q} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}[\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}]^{-1}\mathbf{X}^T\mathbf{V}^{-1}$, $\hat{\boldsymbol{\beta}}$ é estimador de mínimos quadrados generalizados (GLS) ou melhor estimador linear não viciado (BLUE) de $\boldsymbol{\beta}$ e $\hat{\mathbf{u}}$ é denominado de melhor preditor linear não viciado (BLUP) de $\mathbf{u}$ (SINGER; ANDRADE, 1986).

Propriedades de $\hat{\boldsymbol{\beta}}$ e $\hat{\mathbf{u}}$ podem ser verificadas em Henderson (1975). O autor mostra, ainda, que

$$\mathbb{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta},\ \mathbb{V}(\hat{\boldsymbol{\beta}}) = [\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}]^{-1},$$

$$\mathbb{E}(\hat{\mathbf{u}}) = \mathbf{0},\ \mathbb{V}(\hat{\mathbf{u}}) = \mathbf{G}\mathbf{Z}^T\mathbf{Q}\mathbf{Z}\mathbf{G}.$$

Caso as matrizes $\mathbf{G}$ e $\mathbf{R}$ não sejam conhecidas, deve-se inserir estimativas de $\mathbf{V}$ no problema de minimização por GLS. Para tal, deve-se encontrar estimativas razoáveis dos componentes de variância relacionados a $\mathbf{G}$ e $\mathbf{R}$ por meio de algum método de estimação. Os métodos mais utilizados nesses casos são os procedimentos ANOVA baseado no método dos momentos, método de máxima verossimilhança (MV) e método de máxima verossimilhança restrita (MVR ou do inglês REML).

Os métodos que apresentaremos a seguir se tratam de adaptações do modelo linear misto para análise de dados genéticos e seleção de SNPs associados ao fenótipo.

2.2.2 FastGWA

Esta seção será baseada em Jiang *et al.* (2019). Os autores propõem o seguinte modelo denominado *fastGWA*

$$\mathbf{Y} = \mathbf{X}_0\boldsymbol{\alpha}_0 + \mathbf{X}_{snp}\beta_{snp} + \mathbf{u} + \boldsymbol{\varepsilon}, \tag{2.4}$$

em que

- $\mathbf{Y}$ é um vetor de tamanho $(n \times 1)$ de fenótipos;
- $\mathbf{X}_0$ é a matriz de variáveis ambientais, de comportamento e de estratificação populacional ou ancestralidade;
- $\boldsymbol{\alpha}_0$ é vetor de coeficientes da matriz $\mathbf{X}_0$;
- $\mathbf{X}_{snp}$ é um vetor coluna com a informação genotípica de um único marcador de interesse (SNP);
- $\boldsymbol{\beta}_{snp}$ é o efeito fixo do SNP em estudo;
- $\mathbf{u}$ é o vetor dos efeitos genéticos totais capturados por parentesco de linhagem (um intercepto para cada unidade amostral), $\mathbf{u} \sim Normal_n(\mathbf{0}, \mathbf{G})$, com $\mathbf{G} = \sigma_g^2\Pi$;

- $\Pi_{n\times n}$ é a matriz de relacionamento familiar (FAM - *family relatedness matrix*) construída com base na estrutura do parentesco, como já descrito anteriormente;

- $\boldsymbol{\varepsilon}$ é um vetor de erros, $\boldsymbol{\varepsilon} \sim Normal_n(\mathbf{0}, \mathbf{R})$, com $\mathbf{R} = \mathbf{I}\sigma^2_\varepsilon$.

Fazendo a analogia com o modelo descrito na Equação (2.3), temos que $\mathbf{Y} \sim Normal_n(\mathbf{X}_{snp}\beta_{snp} + \mathbf{X}_0\boldsymbol{\alpha}_0, \mathbf{V})$. Ainda, $\mathbf{Z}$ será dado por uma matriz identidade de tamanho $(n \times n)$. Por isso, a matriz de variâncias e covariâncias de $\mathbf{Y}$ é dada por $\mathbf{V} = \mathbf{G} + \mathbf{R} = \Pi\sigma^2_g + \mathbf{I}\sigma^2_\varepsilon$.

Na prática, se a informação de parentesco estiver faltando ou estiver amplamente incompleta, Π pode ser substituída por uma matriz de relacionamento genético derivado de SNPs, como já comentado anteriormente. Jiang *et al.* (2019) apresentam duas versões intimamente relacionadas do método, o *fastGWA* (baseado em GRM esparsa calculada a partir de dados dos SNPs) e o *fastGWA-Ped* (baseado em matriz FAM construída a partir de informações de parentesco). Ambos seguem o mesmo modelo sendo diferenciados apenas pela origem da matriz Π.

O modelo *fastGWA* impõe controle sobre parentesco por informações de linhagem ou por GRM esparsa calculada, com o efeito de estratificação da população capturado pelos componentes principais derivados de SNPs. Os componentes de variância σ^2_ε e σ^2_g são desconhecidos ($\mathbf{G}$ e $\mathbf{R}$ são, portanto, desconhecidas), mas podem ser estimados pelo algoritmo de máxima verossimilhança restrita (MVR ou REML).

Para tal, os autores implementaram algoritmo REML de busca baseado em uma grade eficiente de possíveis valores para essas variâncias (*grid-search-based REML algorithm* denominado *fastGWA-REML*) para estimar σ^2_ε e σ^2_g sem a necessidade de calcular $\mathbf{V}^{-1}$.

Na presença de efeitos ambientais comuns moderados a fortes compartilhados entre parentes, a variância genética estimada $\hat{\sigma}^2_g$ de indivíduos intimamente relacionados (por exemplo, pares de indivíduos com coeficientes de parentesco > 0.05) pode ser uma quantidade mais útil do que a estimada com base no parentesco genético entre todos os indivíduos de pares na amostra, como é o caso na maioria dos métodos existentes baseados em MLM. Isso ocorre porque $\hat{\sigma}^2_g$ estimado de parentes que vivem próximos captura as variações devido aos efeitos genéticos e ambientais comuns (JIANG *et al.*, 2019).

Uma vez que as estimativas dos componentes de variância são obtidas, a matriz de variância-covariância $\mathbf{V}$ e sua inversa podem ser calculadas eficientemente usando os algoritmos de matriz esparsa implementados na biblioteca Eigen C++[2]. Portanto, $\hat{\beta}_{snp}$ pode ser estimado usando a abordagem dos mínimos quadrados generalizados como

$$\hat{\beta}_{snp} = \frac{\mathbf{X}^T_{snp}\mathbf{V}^{-1}\mathbf{Y}}{\mathbf{X}^T_{snp}\mathbf{V}^{-1}\mathbf{X}_{snp}} \text{ com } \mathbb{V}(\hat{\beta}_{snp}) = \frac{1}{\mathbf{X}^T_{snp}\mathbf{V}^{-1}\mathbf{X}_{snp}}, \tag{2.5}$$

[2] http://eigen.tuxfamily.org.

em que os parâmetros usados para calcular $\mathbf{V}$ são desconhecidos, mas podem ser replicados pelas estimativas de REML (máxima verossimilhança restrita) esparsa sob a hipótese nula $\hat{\beta}_{snp} = 0$, como é o caso na maioria dos métodos existentes. A estimação pelo método proposto baseia-se em duas etapas:

1. Etapa da estimação: estimar σ_g^2, $\boldsymbol{\alpha}_0$ e os outros parâmetros sob o modelo nulo ou basal (ou seja, $\mathbf{Y} = \mathbf{X}_0\boldsymbol{\alpha}_0 + \mathbf{u} + \boldsymbol{\varepsilon}$);

2. Etapa do teste de associação: realizar o teste Score para cada marcador SNP disponível.

Na etapa da estimação, os autores utilizam o método denominado *fastGWA-GLMM-REML* por eles desenvolvido, para estimar os componentes de variância no modelo de uma maneira robusta. Na etapa do teste de associação, com base nas estimativas obtidas na etapa anterior, a estatística do teste Score para cada SNP pode ser calculada pela seguinte equação:

$$T_{score} = \mathbf{X}_{snp}^T(\mathbf{Y} - \hat{\mathbf{Y}}) \text{ com } \mathbb{V}(T_{score}) = \mathbf{X}_{snp}^T\mathbf{P}\mathbf{X}_{snp}, \tag{2.6}$$

sendo $\frac{T_{score}^2}{\mathbb{V}(T_{score})} \sim \chi_{gl=1}^2$, em que $\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}_c(\mathbf{X}_c^T\mathbf{V}^{-1}\mathbf{X}_c)\mathbf{X}_c^T\mathbf{V}^{-1}$ e $\hat{\mathbf{Y}}$ é o vetor com valores preditos da variável resposta para as unidades amostrais através do modelo completo em (2.4) considerando o SNP em teste.

2.2.3 FamSKAT-RC

Características complexas tornaram-se um dos grandes focos dos estudos genéticos nas últimas décadas. Uma das discussões mais pertinentes nesta área de estudo compreende a hipótese de que SNPs raros e comuns estão associados a traços complexos. Estudos de associação de dados familiares relativamente grandes são adequados para identificação de SNPs raros e comuns.

Diversos testes de associação baseados em kernel existem na literatura. Para maiores detalhes sobre tal metodologia vide Svishcheva, Belonogova e Axenovich (2014). Saad e Wijsman (2014) estenderam o *Sequence Kernel Association Test* para SNPs raros[3] e comuns (*SKAT-RC*), originalmente proposto para dados de indivíduos não relacionados, para dados familiares e deram o nome a essa extensão de *FamSKAT-RC*. A relação familiar entre os indivíduos é explicada através da matriz de parentesco, de modo que a mesma determina a variância e covariância entre os efeitos aleatórios, que são os coeficientes de regressão associados aos SNPs, sejam eles comuns ou raros.

Considerando que n seja o tamanho da amostra, a matriz de parentesco terá dimensão $(n \times n)$ e cada posição carrega um valor entre 0 e 1 que representa a associação genética entre cada par de indivíduos do conjunto de dados, como já explicado anteriormente.

[3] Os autores consideram raros as SNPs que apresentam MAF≤ 0.01.

Segundo Saad e Wijsman (2014), o modelo *FamSKAT-RC* pode ser escrito como:

$$\mathbf{Y} = \mathbf{X}_0\boldsymbol{\alpha}_0 + \mathbf{R}\boldsymbol{\beta}_r + \mathbf{C}\boldsymbol{\beta}_c + \mathbf{u} + \boldsymbol{\varepsilon}, \tag{2.7}$$

em que

- $\mathbf{Y}$ é um vetor de tamanho $(n \times 1)$ de fenótipos;
- $\mathbf{X}_0$ é a matriz de b variáveis não genéticas (ambientais, de comportamento ou ancestralidade) com dimensão $(n \times b)$;
- $\boldsymbol{\alpha}_0$ é o vetor de tamanho $(b \times 1)$ de coeficientes de regressão fixos da matriz $\mathbf{X}_0$;
- $\mathbf{R} = (v_{R1}R_1, v_{R2}R_2, \ldots, v_{Rr}R_r)$ é a matriz genotípica ponderada de dimensão $(n \times r)$ de r SNPs raros, sendo $p = c + r$ o total de SNPs considerados no modelo com c SNPs comuns;
- $\boldsymbol{\beta}_r$ é o vetor de tamanho $(r \times 1)$ de efeitos aleatórios dos SNPs raros;
- $\mathbf{C} = (v_{C1}C_1, v_{C2}C_2, \ldots, v_{Cc}C_c)$ é a matriz genotípica ponderada de dimensão $(n \times c)$ de c SNPs comuns, sendo $p = c + r$ o total de SNPs considerados no modelo com r SNPs raros;
- $\boldsymbol{\beta}_c$ é o vetor de tamanho $(c \times 1)$ de efeitos aleatórios dos SNPs comuns;
- $\mathbf{u} = (u_1, \ldots, u_n)^T \sim Normal(\mathbf{0}, \sigma_g^2\Pi)$ é outro vetor de efeitos aleatórios, de modo que σ_g^2 corresponde à variância genética;
- $\Pi = 2\Phi$ é a matriz de relacionamento, de dimensão $(n \times n)$; e
- $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$ é um vetor de erros, $\boldsymbol{\varepsilon} \sim Normal(\mathbf{0}, \sigma_\varepsilon^2\mathbf{I})$, tal que σ_ε^2 representa a variância residual e $\mathbf{I}$ uma matriz identidade de dimensão $(n \times n)$.

As variáveis $v_{C.}$ são os pesos para os SNPs comuns, de modo que $\sqrt{v_{C.}}$ é valor da função densidade da distribuição $Beta(0.5, 0.5)$ calculada nos MAFs desses SNPs, e, analogamente, as variáveis $v_{R.}$ são os pesos para os SNPs raros, com $\sqrt{v_{R.}}$ o valor da função densidade da $Beta(1, 25)$ calculada nos MAFs dos SNPs raros. Para mais informações, ver Ionita-Laza *et al.* (2013b) e Wu *et al.* (2011). Assim, SNPs raros possuem, em média, peso maior do que SNPs comuns.

Ainda, de acordo com Saad e Wijsman (2014), algumas restrições são impostas, quais sejam: $\mathbb{E}(\boldsymbol{\beta}_r) = \mathbb{E}(\boldsymbol{\beta}_c) = \mathbf{0}$, $\mathbb{V}(\boldsymbol{\beta}_r) = \phi\tau\mathbf{I}$ e $\mathbb{V}(\boldsymbol{\beta}_c) = (1-\phi)\tau\mathbf{I}$, em que τ é um componente de variância e $0 \leq \phi \leq 1$ é a parte da variância explicada pelos SNPs raros.

O logaritmo da função de verossimilhança é dado por

$$l = \mathbf{K} - \frac{1}{2}\log|\mathbf{V}| - \frac{1}{2}(\mathbf{Y} - \mathbf{X}_0\boldsymbol{\alpha}_0)\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}_0\boldsymbol{\alpha}_0),$$

no qual $\mathbf{V} = \tau\phi\mathbf{R}\mathbf{R}^T + \tau(1-\phi)\mathbf{K}\mathbf{K}^T + \sigma_g^2\Pi + \sigma_\varepsilon^2\mathbf{I} = \mathbb{V}(\mathbf{Y})$.

A significância dos SNPs é verificada via teste de hipóteses baseado no modelo estimado. Já que se assume o efeito dos SNPs como aleatório, a hipótese nula é dada por $\tau = 0$ (não há efeito dos SNPs no fenótipo de interesse), e a hipótese alternativa $\tau > 0$ (há efeito dos SNPs no fenótipo de interesse). Isso porque testar $\boldsymbol{\beta}_r = \boldsymbol{\beta}_c = \mathbf{0}$ equivale a testar $\tau = 0$.

Sob a hipótese nula, o estimador de variância do vetor $\mathbf{Y}$, representado por $\hat{\mathbf{V}}$, é dado por $\hat{\sigma}_g^2\Pi + \hat{\sigma}_\varepsilon^2\mathbf{I}$ e a estatística do teste é descrita como sendo:

$$\mathbf{Q} = \phi\mathbf{Q}_r + (1-\phi)\mathbf{Q}_c,$$

em que $\mathbf{Q}_r = (\mathbf{Y} - \mathbf{X}_0\hat{\boldsymbol{\alpha}}_0)^T\hat{\mathbf{V}}^{-1}\mathbf{R}\mathbf{R}^T\hat{\mathbf{V}}^{-1}(\mathbf{Y} - \mathbf{X}_0\hat{\boldsymbol{\alpha}}_0)$ e $\mathbf{Q}_c = (\mathbf{Y} - \mathbf{X}_0\hat{\boldsymbol{\alpha}}_0)^T\hat{\mathbf{V}}^{-1}\mathbf{C}\mathbf{C}^T\hat{\mathbf{V}}^{-1}(\mathbf{Y} - \mathbf{X}_0\hat{\boldsymbol{\alpha}}_0)$. Ainda, sob a hipótese nula, a estatística $\mathbf{Q}$ segue a soma de p distribuições qui-quadrado com 1 grau de liberdade, ou seja,

$$\mathbf{Q} \sim \sum_{k=1}^{p} \lambda_k \chi_1^2,$$

de modo que λ_ks são os autovalores da matriz $\mathbf{P}^{\frac{1}{2}}(\phi\mathbf{R}\mathbf{R}^T + (1-\phi)\mathbf{C}\mathbf{C}^T)\mathbf{P}^{\frac{1}{2}}$ tal que $\mathbf{P} = \hat{\mathbf{V}}^{-1} - \hat{\mathbf{V}}^{-1}\mathbf{X}_0(\mathbf{X}_0^T\hat{\mathbf{V}}^{-1}\mathbf{X}_0)^{-1}\mathbf{X}_0^T\hat{\mathbf{V}}^{-1}$. Caso seja utilizado $\phi = 1$ estaríamos utilizando a metodologia apenas para SNPs raros, ou seja, considerando no modelo apenas os SNPs com MAF igual ou inferior à 0.05. Por outro lado, se definirmos $\phi = 0$, estaremos empregando o teste de associação de kernel *FamSKAT-RC* apenas aos SNPs comuns.

Os autores compararam o desempenho do *FamSKAT-RC* e vários outros testes de ponderação e associação de kernel existentes. Em dados de família simulados, os resultados mostraram um aumento da precisão de seleção com o uso da combinação de abordagens com ponderação. Além disso, os resultados mostraram melhor desempenho do *FamSKAT-RC* em comparação com os outros testes considerados, na maioria dos cenários investigados por eles.

Ressaltamos que este método está disponível no software `R` através do pacote `famSKATRC` (KUNJI; SAAD, 2017).

2.2.4 Grid-LMM

Ainda tendo por foco dados familiares, Runcie e Crawford (2019) apresentam o *Grid-LMM* [4], um algoritmo extensível para ajustar repetidamente modelos lineares mistos complexos que respondem por múltiplas fontes de heterogeneidade, como variância genética aditiva e não aditiva, heterogeneidade espacial e interações genótipo-ambiente.

O *Grid-LMM* pode calcular estatísticas de teste frequentistas aproximadas (mas precisas) ou Bayesianas em uma fração do tempo em comparação com os métodos de uso geral existentes. No entanto, evidenciaremos a abordagem frequentista e sua extensão, combinada com LASSO.

Ressaltam os autores que a maioria das aplicações genômicas usa apenas a forma mais simples dos modelos lineares mistos (LMM, sigla comumente utilizada para o termo em inglês

[4] https://github.com/deruncie/GridLMM.

linear mixed model) em decorrência das grandes demandas computacionais para o ajuste do modelo. Desenvolveram, então, uma abordagem flexível para ajustar modelos lineares mistos a dados de escala genômica que reduz sua carga computacional e fornece flexibilidade para os usuários escolherem o melhor modelo estatístico (em termos de número de parâmetros do efeito aleatório) para sua análise de dados.

Como foi mencionado, os modelos de efeitos mistos são ferramentas dominantes na maioria das pesquisas genéticas por permitirem a análise de dados familiares e evitarem problemas relacionados ao desvio da suposição fundamental subjacente aos modelos lineares de que as observações são independentes. Termos de efeito aleatório são usados em MLM para explicar correlações específicas entre observações. Ajustar um MLM requer estimar a importância de cada efeito aleatório, chamado de componente de variância.

As ferramentas de uso geral para tal estimação são lentas para serem usadas em conjuntos de dados em escala genômica com milhares de observações e milhões de marcadores genéticos. Conforme descrevem os autores, essa falta de escalabilidade é causada principalmente por dois fatores, a saber: (i) soluções de forma fechada da função de máxima verossimilhança ou máxima verossimilhança restrita (MV ou MVR) ou estimativas das posterioris dos componentes de variância não estão disponíveis e as rotinas de otimização numérica exigem a avaliação repetida da função de verossimilhança inúmeras vezes, e (ii) cada avaliação da verossimilhança requer a inversão da matriz de covariância dos efeitos aleatórios, uma operação que escala de forma cúbica o número de observações. Repetir todo esse processo milhões de vezes rapidamente se torna inviável.

Recentemente, algoritmos de aprendizado para aproximação foram desenvolvidos para múltiplos efeitos aleatórios, mas poucos deles fornecem garantias em termos de precisão de estimativa. O *Grid-LMM* adota uma abordagem diferente para ajustar MLM: em vez de otimizar diretamente os componentes de variância separadamente para cada teste, define-se uma grade abrangendo todos os valores válidos dos componentes de variância e modelos lineares simples são ajustados em cada local da grade.

Cada avaliação envolve uma única decomposição de Cholesky da matriz de covariância de efeito aleatório, que é então reutilizada para calcular soluções MV de forma fechada para todos os testes separadamente, resultando em grande economia de tempo em configurações peculiares de GWAS. Após repetir esses cálculos em toda a grade, a pontuação MV (ou MVR) mais alta para cada marcador ou grupo de marcadores é selecionada para calcular a razão de verossimilhança aproximada e os testes de significância de Wald.

A abordagem do *Grid-LMM* consiste em uma reparametrização da estrutura MLM típica de componente de variância individual σ_g^2 para proporção de componentes de variância $h_g^2 = \frac{\sigma_g^2}{\sigma^2}$, em que σ^2 representa a soma de todos os componentes de variância (incluindo o residual) do fenótipo em análise. Tal proporção é também conhecida como herdabilidade da característica fenotípica em estudo.

Como os componentes de variância devem ser não negativos, suas proporções ficam restritas ao intervalo unitário $[0,1]$ e somam 1, formando um Simplex[5]. Portanto, uma grade finita pode abranger todos os valores válidos desses parâmetros, embora o tamanho da grade aumente rapidamente com o número de efeitos aleatórios considerados.

Os autores afirmam que essa estratégia de condicionar os componentes de variância em uma grade e, em seguida, combinar soluções pode ser aplicada a muitas outras ferramentas em genética quantitativa, incluindo testes de conjuntos para SNPs raros, modelos de regressão de todo o genoma, como LASSO, e mapeamento de QTL em cruzamentos controlados.

Segundo Runcie e Crawford (2019), para L termos de efeito aleatório, o modelo *Grid-LMM* considera a seguinte parametrização do MLM:

$$\mathbf{Y} = \mathbf{X}_0\boldsymbol{\alpha}_0 + \mathbf{X}\boldsymbol{\beta} + \sum_{l=1}^{L} \mathbf{Z}_l\mathbf{u}_l + \boldsymbol{\varepsilon}, \tag{2.8}$$

em que

- $\mathbf{Y}$ é um vetor de tamanho $(n \times 1)$ de fenótipos;
- $\mathbf{X}_0$ é a matriz de b covariáveis não genéticas (ambientais, de comportamento ou ancestralidade) com dimensão $(n \times b)$;
- $\boldsymbol{\alpha}_0$ é o vetor de tamanho $(b \times 1)$ de coeficientes de regressão fixos da matriz $\mathbf{X}_0$;
- $\mathbf{X}$ é a matriz genotípica (SNPs) de dimensão $(n \times p)$;
- $\boldsymbol{\beta}$ é o vetor de tamanho $(p \times 1)$ de coeficientes fixos dos SNPs;
- $\mathbf{Z}_l$ representa a l-ésima matriz de planejamento de tamanho $(n \times r_l)$ correspondente aos efeitos aleatórios;
- $\mathbf{u}_l \sim Normal_{rl}(\mathbf{0}, \sigma^2 h_l^2 \mathbf{K}_l)$ é o vetor de efeitos aleatórios referente ao l-ésimo termo;
- $\mathbf{K}_l$ é uma matriz positiva semi-definida conhecida de dimensão $(r_l \times r_l)$ para o l-ésimo termo;
- $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T \sim Normal_n(\mathbf{0}, \sigma^2 h_\varepsilon^2 \mathbf{I})$ é um vetor de erros aleatórios, tal que $\sigma_\varepsilon^2 = \sigma^2 h_\varepsilon^2$ representa a variância residual, σ^2 soma dos componentes de variância do modelo (erros e efeitos) aleatórios, $\mathbf{h}^2 = (h_1^2, \ldots, h_L^2, h_\varepsilon^2)$ a proporção de variância atribuída a cada termo do efeito aleatório ou erro residual e $\mathbf{I}$ uma matriz identidade de dimensão $(n \times n)$. Todos elementos de $\mathbf{h}^2$ são não-negativos e somam 1, formando um Simplex L-dimensional.

[5] O método Simplex é utilizado na otimização matemática sequencial para avaliar a melhor solução possível de um problema complexo, dadas determinadas condições operacionais e a quantidade de recursos.

Ao considerarmos apenas o intercepto como efeito aleatório, ou seja, fixando $L = 1$ (que é o modelo considerado nesse estudo), podemos reescrever a Equação (2.8) na seguinte forma

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{I}\mathbf{u} + \boldsymbol{\varepsilon},$$

em que $\mathbf{I}$ consiste na matriz identidade $n \times n$ e $\mathbf{u} = (u_1, \ldots, u_n)^T$ representa o vetor de interceptos aleatórios de tal modo que esperamos que indivíduos relacionados assumam valores próximos nesse termo e indivíduos não relacionados apresentem valores distantes. Neste caso, $r_1 = n$ e $\mathbf{K}_1 = \Pi = 2\Phi$.

Consoante os autores, em aplicações GWAS assumimos que o interesse principal é a inferência de $\boldsymbol{\beta}$, uma vez que a matriz genotípica $\mathbf{X}$ varia a cada cromossomo. Já em aplicações de estimativa de herdabilidade, tem-se por foco inferir o vetor $\mathbf{h}^2$. Em ambos os casos, os vetores $\mathbf{u}_l$ e $\boldsymbol{\varepsilon}$ são parâmetros perturbadores e podemos integrá-los.

Com a integração dos vetores aleatórios $\mathbf{u}_l$ e $\boldsymbol{\varepsilon}$ e considerando apenas o intercepto como efeito aleatório, obtemos o seguinte modelo equivalente:

$$\mathbf{Y} \sim Normal_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{U}), \tag{2.9}$$

em que $\mathbf{U} = h_1^2\mathbf{K}_1 + h_\varepsilon^2\mathbf{I}$ e, se essa matriz for de posto completo (o que é garantido se $h_\varepsilon^2 > 0$), podemos utilizar a decomposição de Cholesky $\mathbf{U} = \mathbf{L}\mathbf{L}^T$ de modo a transformar a Equação (2.9) para:

$$\mathbf{Y}^* \sim Normal(\mathbf{X}^*\boldsymbol{\beta}, \sigma^2\mathbf{I}), \tag{2.10}$$

em que $\mathbf{Y}^* = \mathbf{L}^{-1}\mathbf{Y}$ e $\mathbf{X}^* = \mathbf{L}^{-1}\mathbf{X}$. A Equação (2.10) trata-se de modelo linear simples para $\mathbf{Y}^*$ cuja função de log-verossimilhança é dada por:

$$l_F(\mathbf{Y}; \beta, \sigma^2|\mathbf{h}^2) = \frac{1}{2}\left[-n\log(2\pi\sigma^2) - \frac{1}{\sigma^2}(\mathbf{Y}^* - \mathbf{X}^*\boldsymbol{\beta})^T(\mathbf{Y}^* - \mathbf{X}^*\boldsymbol{\beta})\right],$$

através da qual métodos eficientes para inferência de $\boldsymbol{\beta}$ e σ^2 conhecidos podem ser aplicados. Estimativas desses parâmetros são encontradas via método de máxima verossimilhança restrita, definida com base em dados transformados e não originais.

Note que os parâmetros $\mathbf{h}^2$ estão fixos na função de verossimilhança. Isso porque todos os seus possíveis valores são considerados numa grade de tamanho 0.01 e, então, escolhe-se o melhor valor de modo que ele siga as restrições de que todos $h_l^2 \geq 0$ e $\sum_{l=1}^{L} h_l^2 < 1$. Assim sendo, calcula-se a verossimilhança em cada vértice da grade para cada marcador (SNP) ou grupo de marcadores.

Segundo Runcie e Crawford (2019), essa grade pode ser atrelada à origem (ou seja, $h_l^2 = 0, \forall l$) ou à estimativa da função de máxima verossimilhança restrita do modelo nulo (sem efeito significativo dos SNPs no fenótipo). A busca pelos valores de h_l^2 gera um vetor de g escores de verossimilhança restrita para cada marcador ou grupo de marcadores, um para cada

ponto da grade. Para cada SNP ou grupo de SNPs seleciona-se o valor que corresponde ao maior escore e o empregamos no cálculo das estimativas do teste de Wald através da abordagem Grid, utilizadas para calcular a verossimilhança para os modelos sob as hipóteses nulas e alternativas.

A significância dos efeitos dos SNPs, fixado um cromossomo, pode ser avaliada via teste de Wald, cuja hipótese nula associada repousa sob $\mathbf{M}\boldsymbol{\beta} = \mathbf{0}$ (não há efeito dos SNPs no fenótipo de interesse), contra a hipótese alternativa $\mathbf{M}\boldsymbol{\beta} \neq \mathbf{0}$ (há efeito dos SNPs no fenótipo de interesse), em que $\mathbf{M}$ é uma matriz arbitrária de dimensão $m \times p$ que especifica os testes de interesse em relação aos coeficientes de regressão. Caso haja interesse de testar um SNP individualmente, a matriz $\mathbf{M}$ tratar-se-á de vetor linha composto pelo valor 1 na posição correspondente ao SNP de interesse, e valor 0 nos demais elementos. Portanto, m representa o número de SNPs testados conjuntamente ou m combinações lineares dos seus efeitos.

Para apreciar tais hipóteses, recorremos à estatística teste $F_{Wald} = \frac{\hat{\boldsymbol{\beta}}^T \mathbf{M}^T (\mathbf{M}(\mathbf{X}^T \mathbf{U}^{-1} \mathbf{X})^{-1} \mathbf{M}^T)^{-1} \mathbf{M}\hat{\boldsymbol{\beta}}}{m}$, em que $\hat{\boldsymbol{\beta}}$ é a estimativa de $\boldsymbol{\beta}$ calculada através da função de máxima verossimilhança restrita para $\mathbf{U}$. Sob a hipótese nula, a estatística teste F_{Wald} tem distribuição F de Snedecor com m e $n-p$ graus de liberdade. SNPs com efeitos significativos são considerados aqueles que possuem p-valor menor que um nível de significância pré-fixado. Se um grande número de SNPs for testado individualmente, é usual fazermos a correção de Bonferroni no nível de significância de comparação para uma conclusão conjunta sobre a significância deles.

No artigo mencionado, Runcie e Crawford (2019) demonstraram precisão aprimorada para testes de associação genética, maior poder para descobrir variantes genéticas causais e a capacidade de fornecer resumos precisos da incerteza do modelo usando exemplos de dados simulados e reais.

Outra forma de seleção dos SNPs significativos se dá através da extensão do modelo *Grid-LMM* combinado com LASSO, que veremos a seguir.

2.2.5 Grid-LMM combinado com LASSO

O LASSO (do inglês *least absolute shrinkage and selector operator*) foi proposto por Tibshirani (1996) e tem por principal objetivo encontrar um estimador dos parâmetros na regressão linear múltipla que apresente risco de predição menor que o de mínimos quadrados. Este método apresenta duas principais finalidades: a estimação de parâmetros e a seleção de modelos, aliadas a um procedimento computacional eficiente. Isso se deve ao fato de que durante o procedimento, as estimativas de vários coeficientes ficam iguais a zero, levando à construção de modelos esparsos. Na prática, descartam-se as variáveis relacionadas a tais coeficientes, e selecionam-se as demais (IZBICKI; SANTOS, 2020).

Na abordagem do modelo em estudo, o LASSO consiste num método que seleciona os marcadores (SNPs) mais associados ao fenótipo através da adição de uma restrição na fórmula do *Grid-LMM*. Essa restrição é da forma $\sum_{k=1}^{p} |\beta_k| \leq c$, em que $c = c(\lambda)$.

Essa penalização, conforme discutido em Tibshirani (1996), faz com que muitos parâmetros (β_ks) sejam estimados com o valor 0, descartando do modelo as variáveis a eles associadas. No modelo aqui considerado, os SNPs acompanhados de um coeficiente com estimativa 0 são descartados, ou seja, não são considerados como associados ao fenótipo em estudo.

Assim sendo, a metodologia do *Grid-LMM* aliada ao LASSO busca (IZBICKI; SANTOS, 2020):

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\max}\frac{1}{2}\left[-n\log(2\pi\sigma^2) - \frac{1}{\sigma^2}(\mathbf{Y}^* - \mathbf{X}^*\boldsymbol{\beta})^T(\mathbf{Y}^* - \mathbf{X}^*\boldsymbol{\beta})\right] - \lambda\sum_{k=1}^{p}|\beta_k|,$$

em que $\lambda \geq 0$ é um parâmetro de regularização. Ademais, quanto maior o valor de λ, mais restrita é a limitação e menor a quantidade de SNPs associados. Por consequência, menos complexo é o modelo.

Desse modo, a fim de obtermos o melhor modelo, faz-se necessário escolher um valor para λ que não assuma valor nulo ou valores muito grandes, uma vez que acarretaria, no primeiro caso, a seleção de todas as variáveis, e, no segundo caso, a eliminação de todas elas com a permanência apenas do intercepto (β_0) e do vetor de efeitos aleatórios. A escolha do valor deve ocorrer para não apenas minimizar o super-ajuste (um ajuste perfeito na amostra de estimação), mas também para controlar a seleção de variáveis.

Uma maneira de escolher o valor para λ, para cada cromossomo, é aquele que possui o menor erro médio de predição dentre todos os valores para ele testados.

O risco observado dado pelo erro quadrático médio de predição é um estimador muito otimista do real risco preditivo pois ele leva ao super-ajuste. Para solucionar este problema divide-se o conjunto de dados em duas partes, treinamento (estimação) e validação. Neste contexto, utilizaremos uma extensão desse artifício, o método de validação cruzada k-fold (IZBICKI; SANTOS, 2020).

Esse método consiste na divisão aleatória do banco de dados em k partes (subamostras) mutuamente exclusivas e aproximadamente de mesmo tamanho. O modelo é estimado com $k-1$ partes e testado na única parte remanescente. Para cada valor de λ, são propostos k modelos e, para cada um, é calculado o erro médio de predição (EP) dado por

$$EP = \frac{\sum_i (y_i - \hat{y}_i)^2}{n_k}, \tag{2.11}$$

em que y_i e $\hat{y}_i$ consistem em, respectivamente, valor observado e valor predito pelo modelo *Grid-LMM* aliado ao LASSO referentes ao fenótipo dos indivíduos que fazem parte da subamostra não utilizada na estimação e n_k é o tamanho dessa subamostra.

Portanto, para cada valor de λ, teremos k erros médios de predição e podemos calcular a média deles. O valor escolhido para λ consiste naquele que apresentar o menor risco médio observado, ou seja, a menor média de EP. Valores muito utilizados para k são 5 ou 10, sendo o segundo o valor automático de muitos pacotes estatísticos disponíveis.

2.2.6 MLM combinado com LASSO

Outra alternativa aqui proposta para selecionar os SNPs mais associados ao fenótipo em estudo e, ao mesmo tempo, predizer efeitos aleatórios relevantes através dos modelos mistos combinados com o LASSO, mas sem utilizar a metodologia de estimação do *Grid-LMM* que pode ser complicada em alguns cenários, é usar o modelo descrito na Eq. (2.2) em que $\mathbf{Z} = \Pi^{1/2} = (2\Phi)^{1/2}$ ou $\mathbf{Z}$ é a matriz triangular inferior da decomposição de Cholesky de Π (nesse estudo assumiremos a segunda alternativa).

Nessa situação, o modelo pode ser resumido matricialmente como

$$\mathbf{Y} = [\mathbf{X}, \mathbf{Z}] \begin{bmatrix} \boldsymbol{\beta} \\ \mathbf{u} \end{bmatrix} + \boldsymbol{\varepsilon} = \mathbf{B}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}. \tag{2.12}$$

Esse modelo possui as mesmas propriedades e características do Modelo (2.2), com $p+n$ efeitos (fixos e aleatórios) em $\boldsymbol{\gamma}$. No entanto, como a nova matriz de planejamento $\mathbf{B}$ tem tamanho $n \times (p+n)$ e não possui posto completo, não seria possível uma estimação única de $\boldsymbol{\gamma}$ através de metodologias tradicionais.

Entretanto, pela metodologia LASSO, esses efeitos podem ser estimados ou preditos como

$$\hat{\boldsymbol{\gamma}} = \underset{\boldsymbol{\gamma} \in \mathbb{R}^{n+p}}{\arg\min} (\mathbf{Y} - \mathbf{B}\boldsymbol{\gamma})^T (\mathbf{Y} - \mathbf{B}\boldsymbol{\gamma}) + \lambda \sum_{k=1}^{p+n} |\gamma_k|, \tag{2.13}$$

em que $\lambda \geq 0$ é o parâmetro de regularização e controla a quantidade de variáveis selecionadas. O melhor valor de λ pode ser obtido através de métodos de validação cruzada como definido na seção anterior. Quanto maior o valor de λ, maior será a quantidade de coeficientes iguais a zero e os SNPs e efeitos aleatórios com estimativas zero são considerados como não relevantes ou não associados ao fenótipo em estudo.

Vários algoritmos eficientes estão disponíveis na literatura para encontrar a solução do problema em (2.13) para diferentes valores de λ. Se for necessária a estimação dos componentes de variância que não é automática por esse processo de estimação, σ^2_ε pode ser estimado com a variância observada nos resíduos $y_i - \mathbf{b}_i\hat{\boldsymbol{\gamma}}$ e a variância genética σ^2_g como a variância observada em $\hat{\mathbf{u}}$.

Modelos com mais efeitos aleatórios que não apenas o intercepto podem ser considerados nessa metodologia desde que a matriz $\mathbf{Z}$ seja completamente conhecida e não dependa de valores a serem estimados.

CAPÍTULO

3

METODOLOGIAS PARA COLAPSAGEM

As metodologias apresentadas no capítulo anterior são muito utilizadas para estudos de associação entre SNPs e características físicas de interesse em dados familiares. No entanto, exceto pela metodologia *FamSKAT-RC* que considera o efeito de SNPs raros e comuns, geralmente elas apresentam bons resultados de seleção quando lidamos com SNPs comuns, mas resultados deficientes na presença de SNPs raros.

Várias abordagens alternativas têm sido propostas para superar o problema da raridade (baixíssima frequência) combinando múltiplos SNPs raros em um gene, comumente denominadas métodos de colapsagem (SUN *et al.*, 2011). Neste capítulo, mostramos as principais técnicas existentes para colapsagem baseadas em uma unidade de análise, geralmente um gene.

3.1 Metodologias estatísticas

Sabemos que estudos de associação com variantes raras apresentam alguns desafios. A perda de poder dos testes, por exemplo, pode ser mitigada pela análise conjunta de vários SNPs raros agrupados em uma unidade de análise.

Esses agrupamentos podem aumentar o poder reduzindo a carga de testes múltiplos de milhões para dezenas de milhares de testes. Segundo Nicolae (2016), a agregação de SNPs em unidades de análise é feita usando informações de muitas fontes, incluindo (a) localização física em relação aos genes (por exemplo, uma unidade de análise pode consistir em todos os SNPs exônicos[1] de um determinado gene), (b) um subconjunto de SNPs exônicos considerados deletérios[2] com base em seu impacto funcional, (c) um subconjunto de SNPs exônicos com altos escores funcionais de ferramentas computacionais como GERP++ (*Genomic Evolutionary Rate Profiling*++) (DAVYDOV *et al.*, 2010) e (d) unidades obtidas de escores que agregam

[1] Os éxons são sequências de bases transcritas e traduzidas, enquanto os íntrons são sequências transcritas e não traduzidas.

[2] Cuja mutação provoca alterações no fenótipo.

informações de múltiplas fontes, como CADD (*Combined Annotation Dependent Depletion*) (KIRCHER *et al.*, 2014). O autor ainda ressalta que cada um desses agrupamentos corresponde a uma hipótese específica que é explorada, por exemplo, analisar apenas SNPs deletérios previstos implica que apenas esses SNPs são relevantes para o fenótipo sob investigação.

A investigação dos fatores de confusão, como ancestralidade, prevê o ajuste para estratificação populacional que apresenta desafios diferentes em estudos de SNPs raros em comparação aos GWAS de SNPs comuns. Ferramentas clássicas, como estimar um fator de inflação global (uma constante de todo o genoma que reflete a inflação na estatística de teste devido à ancestralidade) não são apropriadas (NICOLAE, 2016). O modelo *fastGWA*, por exemplo, além de recorrer ao MLM, agrega o efeito de estratificação da população através dos componentes principais derivados de SNPs (que já se caracteriza como técnica de colapsagem). A maioria dos autores afirma que esse controle é mais relevante ainda em estudos de SNPs raros do que em estudos de SNPs comuns.

Conceitualmente, as abordagens de SNPs raros baseados em genes funcionam de maneira ideal quando existe uma expectativa de heterogeneidade alélica[3] entre um ou muitos genes associados a doenças ou outras características. Nessas situações, espera-se que cada alelo causal individual explique apenas uma fração muito pequena dos casos em estudo, mas diferentes SNPs no mesmo gene podem ter uma contribuição cumulativa maior (POVYSIL *et al.*, 2019).

Os testes de componente de variância foram desenvolvidos para permitir uma mistura de efeitos em um conjunto de SNPs raros e efeitos de magnitudes diferentes. Os testes projetados para considerar efeitos variados incluem C-alfa (NEALE *et al.*, 2011), o modelo misto para teste de associação agrupada (EMMPAT, do inglês *evolutionary mixed model for pooled association testing*) (KING; RATHOUZ; NICOLAE, 2010), o teste *sum of square* (SSU) (PAN, 2009) e o teste de associação kernel de sequência (SKAT, do inglês *sequence kernel association test*) (WU *et al.*, 2011). No entanto, um dos métodos mais tradicionais para estudo de associação genética com SNPs raros é o teste de Burden (LI; LEAL, 2008; MORGENTHALER; THILLY, 2007), que visa resumir as informações contidas em alguns SNPs em uma única pontuação de carga genética colapsada que pode ser usada diretamente para a análise de associação.

Primeiramente, revisamos os fundamentos dos testes de associação baseados em colapsagem e, em seguida, descrevemos outros métodos complementares propostos por Ionita-Laza *et al.* (2014) para estudos de SNPs raros.

3.1.1 Metodologias SKAT e Burden

Assumimos que n indivíduos foram sequenciados em uma região de interesse (por exemplo, um gene), que contém p SNPs raros. Seja $\mathbf{W}$ a matriz de genótipos de tamanho $(n \times p)$.

[3] Quando diferentes mutações em um mesmo locus podem produzir fenótipos semelhantes.

Consideramos o modelo de regressão para o i-ésimo indivíduo definido como

$$g[\mathbb{E}(Y_i)|\boldsymbol{\beta}] = \beta_0 + \mathbf{x}_{0i}\boldsymbol{\alpha}_0 + \mathbf{w}_i\boldsymbol{\beta}, \tag{3.1}$$

em que $g(.)$ é uma função de ligação e pode ser definida como a função identidade quando o fenótipo em estudo é contínuo (como consideramos em todo Capítulo 2) ou a função logito quando a característica física em análise é dicotômica; $\boldsymbol{\alpha}_0 = (\alpha_{01}, \ldots, \alpha_{0b})^T$ são coeficientes de regressão das variáveis ambientais, de comportamento ou ancestralidade, $\mathbf{x}_{0i} = (x_{0i1}, \ldots, x_{0ib})$, que queremos considerar (por exemplo, sexo, idade, componentes principais relacionados à ascendência, entre outras); $\mathbf{w}_i = (w_{i1}, \ldots, w_{ip})$ é o vetor de genótipos para o i-ésimo indivíduo, e Y_i seu fenótipo correspondente (característica fenotípica em estudo); $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ denota os coeficientes de regressão para os p SNPs no conjunto testado.

Estamos interessados em testar a hipótese nula de nenhum efeito genético significativo, $H_0 : \boldsymbol{\beta} = \mathbf{0}$. Testar individualmente cada $\beta_k = 0$, para $k = 1, \ldots, p$, ou usar múltiplos testes diminui o poder devido à esparsidade dos dados e aos numerosos SNPs em um gene. Portanto, precisamos impor certas suposições para β_k a fim de tornar o teste mais poderoso (reduzir os graus de liberdade do modelo).

Um dos testes mais amplamente usados, o teste de Burden, assume que todos os β_ks têm essencialmente o mesmo valor, digamos β, e o modelo de regressão em (3.1) equivale a $g[\mathbb{E}(Y_i)|\boldsymbol{\beta}] = \beta_0 + \mathbf{x}_{0i}\boldsymbol{\alpha}_0 + \beta \sum_{k=1}^{p} w_{ik}$. Dessa maneira, para cada indivíduo, somamos o número de alelos raros de todos os marcadores SNPs na região genética para formar uma pontuação de variação genética agregada. Outra maneira de colapsar os SNPs para o teste de Burden, que não será utilizada neste estudo, é a dicotomização, onde a pontuação colapsada indica se o indivíduo correspondente carrega o(s) alelo(s) raro(s) ou não (1 ou 0).

O kernel linear ponderado, por sua vez, é construído sob a suposição de que os β_ks são independentes. Lee, Wu e Lin (2012) propõem uma nova família de kernels que incorpora explicitamente a correlação entre os efeitos dos SNPs. De maneira mais geral, assume-se que $\boldsymbol{\beta}$ é um vetor aleatório com $\mathbb{E}(\beta_k) = 0$, $\mathbb{V}(\beta_k) = v_k^2\tau$ e $corr(\beta_k, \beta_j) = \rho$ para k e j diferentes. Note que o efeito dos SNPs é então considerado como efeito aleatório e para testar a hipótese nula de nenhum efeito genético significativo $H_0 : \boldsymbol{\beta} = \mathbf{0}$, a estatística teste proposta do componente de variância $\tau = 0$ é

$$Q_\rho = (\mathbf{Y} - \hat{\boldsymbol{\mu}}_0)^T \mathbf{K}_\rho (\mathbf{Y} - \hat{\boldsymbol{\mu}}_0), \tag{3.2}$$

em que $\mathbf{K}_\rho = \mathbf{W}\mathbf{V}\mathbf{R}_\rho\mathbf{V}\mathbf{W}^T$, e $\mathbf{R}_\rho = (1-\rho)\mathbf{I} + \rho\mathbf{1}\mathbf{1}^T$, especifica uma matriz de correlação intercambiável (onde os elementos diagonais, autocorrelações, são iguais a 1, e todos os elementos fora da diagonal principal, correlações entre diferentes observações, são iguais a uma constante ρ), e $\mathbf{V} = diag(v_1, \ldots, v_p)$ é uma matriz diagonal de pesos, na qual cada peso pode ser relacionado, por exemplo, ao efeito funcional previsto de um SNP (por exemplo, escore PolyPhen-2 (ADZHUBEI *et al.*, 2010) ou SIFT (SIM *et al.*, 2012)). Para um fenótipo dicotômico, $\hat{\boldsymbol{\mu}}_0$ é um

vetor de probabilidades estimadas de $\mathbf{Y}$ sob o modelo nulo ou basal. Para um fenótipo contínuo, $\hat{\boldsymbol{\mu}}_0$ se dá pelo vetor de valores preditos de $\mathbf{Y}$ sob o modelo nulo ou basal, ou pela média amostral de $\mathbf{Y}$ no caso de não haver variáveis ambientais, de comportamento ou ancestralidade disponíveis no estudo. Embora esta classe de testes seja mais geral, os dois testes comumente usados são o teste de Burden ($\rho = 1$) e o teste SKAT ($\rho = 0$). Essas estatísticas testes podem ser escritas simplesmente como:

$$\text{SKAT: } Q_{\rho=0} = \sum_{k=1}^{p} v_k^2 \left[\sum_{i=1}^{n} (Y_i - \hat{\mu}_{0i}) w_{ik} \right]^2, \text{ e} \tag{3.3}$$

$$\text{Burden: } Q_{\rho=1} = \left[\sum_{k=1}^{p} v_k \sum_{i=1}^{n} (Y_i - \hat{\mu}_{0i}) w_{ik} \right]^2. \tag{3.4}$$

Podemos assumir peso $v_k = 1$ para todos os SNPs, mas também é comum na literatura assumirmos $v_k = \frac{1}{[\text{MAF}_k(1-\text{MAF}_k)]^{1/2}}$. Outra abordagem bastante comum e implementada nos algoritmos disponíveis determina que os pesos v_k assumem o valor da densidade da distribuição $Beta(1,25)$ calculada no MAF do SNP em análise.

A distribuição de Q_ρ sob H_0 é aproximada por uma mistura de distribuições χ^2_1. O método de Davies (DAVIES, 1977) ou ajuste de momentos pode ser empregado para calcular o p-valor. Conjuntos de SNPs raros que apresentam efeito significativo para o fenótipo são os que possuem valor-p menor que o nível de significância especificado ou esse nível corrigido por Bonferroni (ou outro método parecido) em caso de teste para inúmeros grupos de SNPs raros.

O desempenho relativo dos dois testes dependerá do verdadeiro modelo subjacente. O teste de Burden tende a ser mais poderoso quando os SNPs associados ao fenótipo são todos do mesmo tipo (de risco ou protetores) e com efeitos de magnitude semelhante. O teste SKAT tende a ser mais poderoso quando há uma mistura de SNPs de risco e proteção, e também quando apenas uma pequena porcentagem de SNPs em uma região é causal.

No entanto, essa versão original do teste de Burden e SKAT supõe independência entre os indivíduos da amostra, que não é verdade para indivíduos com parentesco. Uma estrutura paralela para estudos baseados em dados de família com aplicação da colapsagem através do teste SKAT é aquela proposta na metodologia *FamSKAT-RC* na Seção 2.2.3. Para mais informações sobre a adequação dos testes de colapsagem para dados familiares vide Ionita-Laza *et al.* (2013a).

Para o teste de Burden, desconhecemos uma versão já proposta que se aplique a dados em família. No entanto, podemos utilizar a ideia central do método que é agregar a informação dos SNPs raros dentro da unidade de análise através da contagem dos alelos raros para cada unidade amostral e utilizar essa nova pontuação de variação genética como variável comum nos modelos MLM, *fastGWA* ou *Grid-LMM*. Essa abordagem que se aproxima à colapsagem

proposta no teste de Burden, cria uma nova variável, D_i, que consiste em

$$D_i = \sum_{k=1}^{p} w_{ik}, \tag{3.5}$$

em que w_{ik} corresponde ao genótipo codificado. Mais adiante, na Seção 4.1.1, descreveremos a codificação utilizada no contexto dessa pesquisa.

Nesse estudo, usaremos essa ideia de colapsagem aliada ao modelo *Grid-LMM* e os testes de significância associados para selecionar os grupos de SNPs raros mais relevantes, assim como o MLM combinado com o LASSO.

3.1.2 Procedimento de eliminação backward

Os testes de associação por colapsagem descritos acima testam a associação ao nível de unidade de análise (gene, em geral), mas não são capazes de identificar SNPs causais individualmente. Contudo, uma vez que um gene tenha demonstrado conter SNPs associados ao fenótipo (por exemplo, usando os testes Burden ou SKAT), identificar os SNPs causais individuais entre os muitos marcadores de um gene é de interesse considerável, pois pode levar a uma melhor compreensão dos mecanismos moleculares subjacentes a um traço complexo, e é essencial para o trabalho de validação experimental posterior (IONITA-LAZA *et al.*, 2014).

Iniciando a análise com um teste de associação por colapsagem, uma maneira natural de identificar SNPs que são individualmente de efeito fraco é avaliar sua contribuição para um determinado conjunto de SNPs removendo o SNP do conjunto e avaliando o efeito resultante, por exemplo, o p-valor para o conjunto reduzido. O algoritmo iterativo descrito a seguir foi projetado para esse propósito:

Passo 1. Inicie com um conjunto de p SNPs raros $W = (w_1, \ldots, w_p)$. O conjunto atual é dado por $W_c = W$. Calcule a estatística teste Q_ρ a partir das Equações (3.3) e (3.4) ($\rho = 0$ ou $\rho = 1$, respectivamente) para o conjunto atual W_c, e calcule o p-valor p_{W_c};

Passo 2. Remova cada um dos p SNPs, um por vez, de W_c, ou seja, considere os conjuntos $W_{-k} = (w_1, \ldots, w_{k-1}, w_{k+1}, \ldots, w_p)$ com $k = 1, \ldots, p$ e, em seguida, calcule a estatística teste correspondente e o p-valor para cada um desses conjuntos reduzidos $p_{W_{-k}}$;

Passo 3. Se $\min(p_{W_{-1}}, \ldots, p_{W_{-p}}) \leq p_{W_c}$ então remova o SNP k que leva ao menor p-valor, sendo

$$k = \arg\min(p_{W_{-1}}, \ldots, p_{W_{-p}}). \tag{3.6}$$

O conjunto atual torna-se $W_c = W_{-k}$ e as etapas 2 e 3 são repetidas. Se o p-valor atual não puder ser melhorado, vá para o Passo 4;

Passo 4. Retorne o conjunto atual de SNPs.

Observe que nesse algoritmo, o modelo inicial considera todos os SNPs e os menos significativos são eliminados um a um em um algoritmo *backward*, daí seu nome. Ele é aplicável

quando o número de SNPs com os quais começamos no Passo 1 não é muito grande (caso contrário, a contribuição do SNP fraco para um conjunto grande é de difícil avaliação). No entanto, sequenciar um gene em milhares de indivíduos pode levar à detecção de centenas de SNPs potenciais ou mais.

Portanto, os autores sugerem a utilização de um procedimento de reamostragem, através do qual um pequeno número de SNPs é escolhido por vez do grande número de marcadores sequenciados em um gene (digamos $p \sim 10$ a 20), de tal forma que o algoritmo acima é aplicado a conjuntos pequenos inúmeras vezes (nos exemplos, Ionita-Laza *et al.* (2014) usaram 2.000 reamostragens, embora esse número possa ser aumentado no caso de um grande número de SNPs no gene).

Finalmente, para cada SNP do gene, calculamos o número de ocorrências na Etapa 4 e a esse número damos o nome de "contagem de retorno" para um SNP. Um procedimento de reamostragem semelhante foi aplicado no contexto da interação gene a gene em Lo e Zheng (2004).

O objetivo é usar a amostra de contagens de retorno para dividir os SNPs em dois grupos: "interessante" (contagens de retorno mais altas) e "não interessante" (contagens de retorno mais baixas), com a expectativa de que a categoria "interessante" contenha SNPs associados ao fenótipo em estudo. Os autores utilizaram métodos do tipo EM (do inglês *expectation–maximization*) para identificar os dois subgrupos (BENAGLIA; CHAUVEAU; HUNTER, 2009).

3.1.2.1 *Integrando a anotação funcional no algoritmo backward*

É bem reconhecido que certas categorias funcionais são mais propensas a serem reconhecidas entre SNPs causais do que outras. Um exemplo óbvio são os SNPs raros não-sinônimos (*non-synonymous*), conhecidos por serem frequentes entre os SNPs causadores de doenças. Da mesma forma, os SNPs de perda de função[4] (LoF, do inglês *loss-of-function*), incluindo mutações *nonsense*, *splice-site*[5] e *frameshift*[6], são fortemente encontrados entre os SNPs causais.

Portanto, a estratificação de SNPs por diferentes categorias funcionais pode diminuir as taxas de falso positivo. Podemos incorporar informações sobre anotação funcional no algoritmo de eliminação *backward*. Isso pode ser feito simplesmente aplicando o algoritmo em diferentes classes, digamos não-sinônimos e sinônimos. Além disso, outros escores funcionais (como PolyPhen-2, SIFT e GERP++) podem ser explicitamente incorporados nas próprias estatísticas

[4] SNPs que perturbam seriamente a função dos genes codificadores de proteínas.

[5] Mutação genética que insere, exclui ou altera uma série de nucleotídeos no limite de um éxon e um íntron. Essa alteração pode interromper o *splicing* do RNA, resultando na perda de éxons ou na inclusão de íntrons e em uma sequência de codificação de proteína alterada.

[6] Uma inserção ou deleção envolvendo um número de pares de bases que não é um múltiplo de três, o que consequentemente interrompe o quadro de leitura de tripletos (aminoácidos) de uma sequência de DNA. Esses SNPs geralmente levam à criação de um códon de terminação prematura (parada) e resultam em um produto proteico truncado (mais curto que o normal).

teste de SKAT e Burden (como pesos associados a SNPs individuais nas Equações (3.3) e (3.4)).

3.1.3 Modelo hierárquico para estimar as razões de chance de SNPs raros individuais

Uma abordagem complementar ao procedimento de eliminação *backward* descrito anteriormente é a do modelo hierárquico. A modelagem hierárquica tem várias vantagens importantes na análise de dados de SNPs raros, porque pode integrar naturalmente vários escores de predição funcional para SNPs individuais. Esse conhecimento prévio será essencial para identificar os SNPs causais prováveis em um gene, especialmente para variantes causais que são raras o suficiente para aparecer apenas algumas vezes em um estudo.

Para tais SNPs, as frequências observadas em casos e controles claramente não são suficientes para distingui-los da grande maioria da variação aleatória (estudos mostraram que mais de 74% das variantes raras eram *singletons* ou *doubletons* (NELSON *et al.*, 2012), ou seja, ocorriam uma única vez ou duas vezes na amostra). Informações sobre o efeito funcional putativo de um SNP na proteína ou o grau de conservação evolutiva podem ser um indicador importante da probabilidade de um SNP ser causal.

Tais informações funcionais podem ser incorporadas por meio de um modelo hierárquico (IONITA-LAZA *et al.*, 2014). Na primeira etapa, o valor do fenótipo $\mathbf{Y}$ é relacionado aos genótipos por meio do seguinte modelo:

$$g[\mathbb{E}(Y_i)|\boldsymbol{\beta}] = \beta_0 + \mathbf{x}_{0i}\boldsymbol{\alpha}_0 + \mathbf{w}_i\boldsymbol{\beta}, \text{ para } i = 1,\ldots,n, \tag{3.7}$$

com notações semelhantes às do modelo descrito em (3.1).

Um modelo na segunda etapa relaciona os riscos de SNPs individuais a informações anteriores conhecidas (por exemplo, anotação funcional) sobre os SNPs dado por

$$\boldsymbol{\beta} = \mathbf{H}\boldsymbol{\eta} + \boldsymbol{\delta}, \tag{3.8}$$

em que $\mathbf{H}$ é uma matriz de dimensão $(p \times r)$ para as r variáveis dos SNPs (por exemplo, informações funcionais); $\boldsymbol{\eta}$ é um vetor de tamanho $(r \times 1)$ de parâmetros de regressão para as variáveis do segundo estágio e $\boldsymbol{\delta}$ é um vetor de tamanho $(p \times 1)$ de erros aleatórios normalmente distribuídos, assumidos (por conveniência) como independentes. Uma vantagem principal da estrutura de modelagem hierárquica é que ela pode facilmente incorporar várias anotações funcionais.

Combinando os dois modelos acima, obtém-se o seguinte modelo generalizado de efeitos mistos lineares:

$$g[\mathbb{E}(Y_i)|\boldsymbol{\beta}] = \beta_0 + \mathbf{x}_{0i}\boldsymbol{\alpha}_0 + \mathbf{w}_i\mathbf{H}\boldsymbol{\eta} + \mathbf{w}_i\boldsymbol{\delta}, \text{ para } i = 1,\ldots,n. \tag{3.9}$$

Os parâmetros do modelo acima podem ser estimados usando uma abordagem de pseudo-verossimilhança Bayesiana híbrida que realiza a estimativa Bayesiana do componente de variância do modelo e, em seguida, conduz a estimativa de pseudo-verossimilhança dos efeitos fixos e aleatórios usando esta variação estimada de efeitos aleatórios (CAPANU; BEGG, 2011).

Utilizam-se as estimativas resultantes para as razões de chance e seus erros padrão para então classificar (ranquear) SNPs em um gene. Naturalmente, os mais difíceis de identificar serão os SNPs causais que ocorrem apenas algumas vezes. As estimativas de *odds ratio* para tais SNPs dependerão fortemente das variáveis hierárquicas, como informações sobre o efeito funcional previsto para um SNP. Exemplificando, para um SNP que ocorre com pouca frequência em um conjunto de dados (por exemplo, 2 vezes nos casos e 0 vezes nos controles), saber que é um SNP LoF aumenta sua probabilidade de ser um SNP causal em comparação com um SNP sinônimo com a mesma frequência.

3.1.4 Procedimento de eliminação backward e modelo hierárquico

A ideia principal é de combinar os dois métodos anteriores, o procedimento de eliminação *backward* e a estrutura de modelagem hierárquica, como segue. Com essa combinação, identificamos a lista de SNPs "interessantes" através do algoritmo de eliminação *backward* e, para cada um desses SNPs, relacionamos a estimativa do tamanho do efeito (coeficiente) e o erro padrão associado obtido do modelo hierárquico. Os SNPs na lista "interessantes" podem ser classificados (ranqueados) naturalmente de acordo com essas estimativas de efeito.

Os autores Ionita-Laza *et al.* (2014) demonstram que restringir a atenção apenas à lista de SNPs "interessantes" pode melhorar a classificação de SNPs causais e que essa abordagem combinada funciona bem nos cenários que investigaram (síndrome de Cohen, autismo, e estudo do coração de Dallas).

É possível incorporar um preditor funcional para os SNPs diretamente no procedimento de eliminação *backward* (como um peso na estatística do teste de Burden). Entretanto, não fica claro escolher uma única anotação funcional entre várias anotações disponíveis. Portanto, o modelo hierárquico tem a vantagem importante de que vários preditores funcionais podem ser incluídos e, como mostraram os autores no artigo mencionado, a classificação dos SNPs causais melhora com a adição de vários preditores.

O principal objetivo dos métodos propostos é combinar dados de sequenciamento com previsões funcionais sobre a prejudicialidade de SNPs[7] para identificar um conjunto de SNPs promissores (candidatos), enriquecido em SNPs causais. Ademais, os SNPs selecionados podem ser classificados de acordo com suas contagens de retorno do procedimento de eliminação *backward* ou com os efeitos $\hat{\boldsymbol{\beta}}$ estimados do modelo hierárquico (a classificação baseada em

[7] Do inglês *deleteriousness of variants*. Essa expressão é usada na genética para se referir à capacidade de certos SNPs causarem danos ou efeitos negativos em um organismo. Em outras palavras, trata-se da medida em que um determinado SNP pode ser prejudicial para um organismo ou população.

escores de **H** proporcionou resultados semelhantes nos estudos dos autores). As medidas para avaliar o desempenho dos métodos utilizadas no estudo mencionado foram: (1) a classificação geral dos verdadeiros SNPs causais entre os SNPs do gene e (2) o viés e a precisão de cobertura na estimativa dos efeitos para os SNPs do modelo hierárquico.

CAPÍTULO

4

BANCO DE DADOS GAW17

Neste capítulo, discutiremos o banco de dados GAW17 utilizado em nossas análises. Descreveremos detalhadamente o processo de limpeza e codificação das variantes, etapas essenciais para a aplicação das metodologias em estudo. Além disso, abordaremos as métricas de desempenho utilizadas para avaliar os resultados e a criação da nova variável inspirada no teste de Burden, conforme discutido na Seção 3.1.1.

4.1 Banco de dados GAW17

O conjunto de dados utilizado neste estudo corresponde ao GAW17 (ALMASY *et al.*, 2011), composto por dados simulados desenvolvidos especificamente para o *Genetic Analysis Workshop 17*. Este foi elaborado para replicar um subconjunto de dados que poderia ser gerado em um rastreamento exômico completo para um distúrbio complexo e fatores de risco associados, facilitando assim a exploração de questões relacionadas ao planejamento de estudos e análises genéticas estatísticas por parte dos participantes do workshop.

Utilizando dados de sequência real do *Projeto 1000 Genomas*, esta simulação modelou uma característica de doença comum com prevalência de 30%, juntamente com três fenótipos de risco quantitativos, em uma amostra de 697 indivíduos de parentesco extenso. O modelo incorporou variantes comuns e raras, com frequências alélicas variando de 0,07% a 25,8%, e uma ampla gama de tamanhos de efeito.

O *Projeto 1000 Genomas*, uma iniciativa global para mapear a diversidade genética humana, incluiu indivíduos de diversas ascendências como europeia, asiática oriental, asiática meridional, africana ocidental e indígena americana. Com a conclusão de três projetos piloto em 2010, o projeto realizou sequenciamentos genômicos de baixa e alta cobertura. Os dados exônicos publicamente disponíveis foram empregados para simular a prevalência de doença e outros traços quantitativos no contexto do GAW17. As simulações utilizaram esses dados para

modelar a distribuição e frequência de polimorfismos de nucleotídeo único (SNPs) permitindo aos participantes do workshop investigar profundamente as implicações do planejamento de estudo e análise estatística genética. Um genoma humano masculino baseado na sequência de referência 36 do Centro Nacional de Informações sobre Biotecnologia (RefSeq36) foi utilizado como sequência de genoma de referência para os alinhamentos de indivíduos do sexo masculino e feminino.

Uma doença comum, com prevalência de 30%, foi simulada juntamente com três fatores de risco quantitativos relacionados, denominados Q1, Q2 e Q4. A condição de fumante (com prevalência de 25%) também foi simulada. Foram realizadas múltiplas simulações fenotípicas para gerar 200 réplicas dos conjuntos de dados de indivíduos relacionados. É importante notar que os dados genotípicos permaneceram constantes entre as réplicas, assim como a idade, sexo e configuração de parentesco. Nesse estudo, analisaremos apenas uma réplica que nos foi disponibilizada.

Neste estudo, o fenótipo quantitativo Q1 será empregado como a variável principal para a aplicação e avaliação das metodologias propostas. Dois dos SNPs que mais influenciam Q1 são oriundos do gene VEGF (fator de crescimento endotelial vascular).

As variantes funcionais incluíram alelos raros e comuns, exibindo uma gama diferente de tamanhos de efeito. Embora a maioria destas variantes exibam efeitos pequenos, algumas têm efeitos grandes que podem ser detectados na maioria das réplicas do conjunto de dados. Além disso, embora alguns genes tenham apenas uma variante funcional, outros contêm múltiplas variantes.

Os fatores de risco quantitativos Q1, Q2 e Q4 foram modelados como fenótipos normalmente distribuídos. A condição patológica foi simulada utilizando um modelo de limiar de responsabilidade, no qual os indivíduos situados nos 30% superiores da distribuição foram considerados afetados. Todos os efeitos dos SNPs eram aditivos na escala do traço quantitativo ou de responsabilidade, de modo que cada cópia do alelo menos frequente incrementava o valor médio do traço em uma quantidade uniforme.

Tabela 1 – Parâmetros do Modelo para Q1.

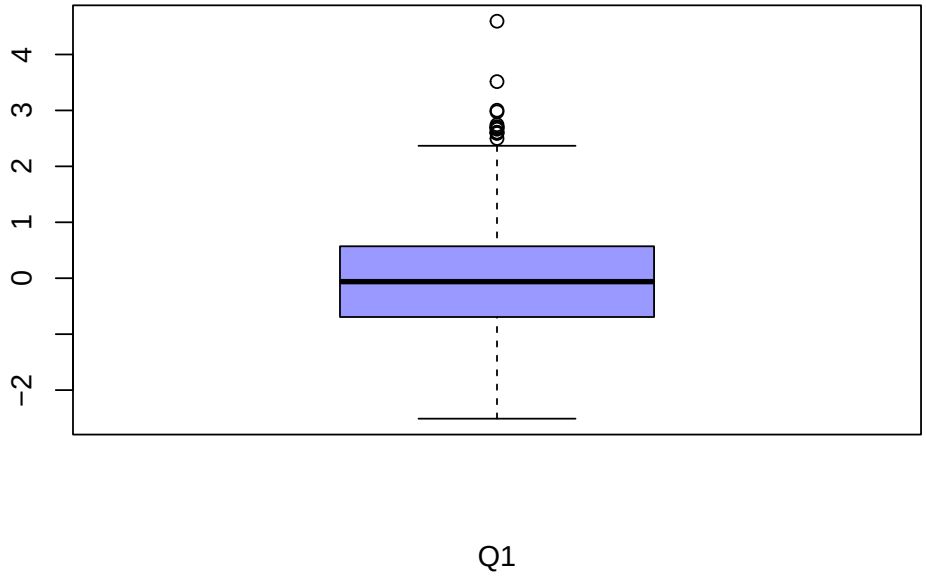
Gene	**SNP**	**MAF**	$\boldsymbol{\beta}$
ARNT	C1S6533	0.011478	0.589734
ARNT	C1S6537	0.000717	0.642689
ARNT	C1S6540	0.001435	0.323662
ARNT	C1S6542	0.002152	0.488219
ARNT	C1S6561	0.000717	0.625721
ELAVL4	C1S3181	0.000717	0.795093
ELAVL4	C1S3182	0.000717	0.328748
FLT1	C13S320	0.001435	0.18047
FLT1	C13S399	0.000717	0.457361
FLT1	C13S431	0.017217	0.732566
FLT1	C13S479	0.000717	0.839669
FLT1	C13S505	0.000717	0.38582
FLT1	C13S514	0.000717	0.549816
FLT1	C13S522	0.027977	0.623466
FLT1	C13S523	0.066714	0.653351
FLT1	C13S524	0.004304	0.596704
FLT1	C13S547	0.000717	0.549214
FLT1	C13S567	0.000717	0.0905862
FLT4	C5S5133	0.001435	0.120761
FLT4	C5S5156	0.000717	0.385374
HIF1A	C14S1718	0.000717	0.251622
HIF1A	C14S1729	0.002152	0.329088
HIF1A	C14S1734	0.012195	0.220448
HIF1A	C14S1736	0.000717	0.228202
HIF3A	C19S4799	0.000717	0.174668
HIF3A	C19S4815	0.000717	0.51468
HIF3A	C19S4831	0.000717	0.265181
KDR	C4S1861	0.002152	0.598271
KDR	C4S1873	0.000717	0.715613
KDR	C4S1874	0.000717	0.503025
KDR	C4S1877	0.000717	1.17194
KDR	C4S1878	0.164993	0.149975
KDR	C4S1879	0.000717	0.610938
KDR	C4S1884	0.020803	0.318125
KDR	C4S1887	0.000717	0.312058
KDR	C4S1889	0.000717	1.17194
KDR	C4S1890	0.002152	0.417977
VEGFA	C6S2981	0.002152	1.13045
VEGFC	C4S4935	0.000717	1.40529

O fenótipo Q1 é influenciado por 39 SNPs localizados em 9 genes distintos, conforme apresentado na Tabela 1[1]. Estes genes exibem entre 1 e 11 variantes funcionais, com as frequências alélicas mínimas (MAF) observadas nos dados do *Projeto 1000 Genomas* variando de

[1] As linhas destacadas na tabela são os SNPs que permaneceram raros, após as etapas de filtragem detalhadas a seguir.

0,07% (indicativo de uma única cópia do alelo menor) a 16,5%. Em todos os casos analisados, a presença do alelo menor estava associada a um incremento nos valores médios de Q1; a coluna β da tabela especifica o deslocamento médio em Q1 para cada cópia adicional do alelo menor. O boxplot do fator de risco quantitativo Q1 está disponível na Figura 1.

Figura 1 – Boxplot do fator de risco Q1.



Observamos, através da Figura 1, que a característica Q1 é uma variável contínua que assume valores no intervalo $(-3.0, 4.5)$ e apresenta diversos outliers, representados na figura por pontos vazios. Além disso, Q1 possui um comportamento simétrico (exceto pelos outliers) e 50% dos valores observados estão entre -1 e 1.

Os dados de genótipos para o GAW17 incluíam genótipos inferidos, de modo que todos os indivíduos possuíam genótipos para todas as posições de pares de bases, e os fenótipos foram simulados com base nesses dados. Os marcadores foram numerados sequencialmente em cada cromossomo e rotulados como *CxSx* de modo que, por exemplo, C1S254 é o 254º SNP no cromossomo 1. No total, 24487 SNPs autossômicos do genótipo foram, para fins de simulação, atribuídos a 3.205 genes com base na primeira interseção encontrada entre a localização do marcador e as coordenadas de pares de bases de todos os genes obtidos das anotações RefSeq36. SNPs que se sobrepunham a vários genes foram atribuídos a apenas um desses genes.

Além disso, consoante a simulação realizada, os cromossomos que possuem SNPs significativos (associados ao fenótipo em estudo) são os cromossomos 1, 4, 5, 6, 13, 14 e 19. Esses SNPs estão listados a seguir, organizados por cromossomo e gene:

- **Cromossomo 1:**

  - *gene ARNT*: C1S6533, C1S6537, C1S6540, C1S6542 e C1S6561;
  - *gene ELAVL4*: C1S3181 e C1S3182;

- **Cromossomo 4:**
  - *gene KDR*: C4S1861, C4S1873, C4S1874, C4S1877, C4S1878, C4S1879, C4S1884, C4S1887, C4S1889 e C4S1890;
  - *gene VEGFC*: C4S4935;

- **Cromossomo 5:**
  - *gene FLT4*: C5S5133 e C5S5156;

- **Cromossomo 6:**
  - *gene VEGFA*: C6S2981;

- **Cromossomo 13:**
  - *gene FLT1*: C13S320, C13S399, C13S431, C13S479, C13S505, C13S514, C13S522, C13S523, C13S524, C13S547 e C13S567;

- **Cromossomo 14:**
  - *gene HIF1A*: C14S1718, C14S1729, C14S1734 e C14S1736; e

- **Cromossomo 19:**
  - *gene HIF3A*: C19S4799, C19S4815 e C19S4831.

4.1.1 Tratamento do banco de dados

Originalmente, o banco de dados nos fornece as informações dos genótipos dos marcadores SNPs bialélicos através de 16 possíveis pares de bases nitrogenadas, sendo eles: A/A, T/T, C/C, G/G, A/T, A/C, A/G, T/A, T/C, T/G, C/A, C/T, C/G, G/A, G/T, G/C. Todavia, a fim de possibilitar este estudo, analisamos quais duas das quatro bases nitrogenadas ocorriam em cada marcador. A base menos frequente foi denominada de a e a mais frequente de A em cada marcador. Em outras palavras, os dados haplotípicos, que correspondem à codificação dos alelos do SNP, são representados por A ou a, como

$$\text{Haplótipo} = \begin{cases} A, \text{ se alelo mais frequente;} \\ a, \text{ se alelo menos frequente.} \end{cases}$$

Por outro lado, os dados genotípicos que representam a codificação dos alelos do SNP no par de cromossomos homólogos, foram codificados como:

$$\text{Genótipo} = \begin{cases} 0, \text{ se os dois alelos forem de maior frequência } (AA); \\ 1, \text{ se um dos alelos for de maior frequência e o outro de menor frequência } (Aa \text{ ou } aA); \\ 2, \text{ os dois alelos forem de menor frequência } (aa). \end{cases}$$

Dessa forma, para um determinado SNP, suponha que o alelo mais frequente seja representado por T e o alelo menos frequente por G. A codificação será realizada conforme Tabela 2.

Tabela 2 – Codificação de SNPs.

SNP		**Genótipo**		**Genótipo**
TT	$\longrightarrow$	*AA*	$\longrightarrow$	0
TG	$\longrightarrow$	*Aa*	$\longrightarrow$	1
GG	$\longrightarrow$	*aa*	$\longrightarrow$	2

Sabemos que o fenótipo Q1 é associado originalmente por 39 SNPs localizados em 9 genes distintos (Tabela 1) dentre os 24465 SNPs. Destes, 37 são raros (MAF$\leq$ 0.05). O foco desse estudo está na identificação e seleção de variantes raras e, por essa razão, filtramos e analisamos apenas os SNPs com $0 <$ MAF ≤ 0.05. Como os dados são simulados, a frequência do alelo menor (MAF) na amostra pode diferir dos valores do modelo que originou os dados. Reforçamos também que apesar de variáveis de comportamento e ambientais, tais como idade, sexo e condição de fumante estarem disponíveis, elas não foram consideradas na análise.

O primeiro passo na preparação do banco de dados de SNPs para as análises foi, então, calcular a frequência do alelo menor para os SNPs e eliminar todas as variantes monomórficas, ou seja, SNPs sem variabilidade na amostra (MAF$=0$). O segundo passo consistiu em eliminar as variantes com MAF>0.05 (nas quais a frequência do alelo menos comum na amostra é superior a 5%).

Após as etapas de filtragem dos SNPs raros, restaram no banco de dados, considerando os 22 cromossomos, o total de 10652 SNPs, distribuídos em 2567 genes. Em relação aos SNPs conhecidos por influenciarem o fenótipo Q1, permaneceram 15 SNPs raros no banco de dados (22 SNPs raros foram descartados), distribuídos em 8 genes localizados em 6 cromossomos distintos (Tabela 3). Na Tabela 1, as linhas destacadas referem-se às 15 variantes raras, que na Tabela 4, são detalhadas e caracterizadas com informações adicionais.

Tabela 3 – Quantidade de SNPs raros influentes em Q1 antes e após tratamento do banco de dados.

Cromossomo	**1**	**4**	**5**	**6**	**13**	**14**	**19**	**Total**
n inicial	7	10	2	1	10	4	3	37
n final	3	5	0	1	4	1	1	15

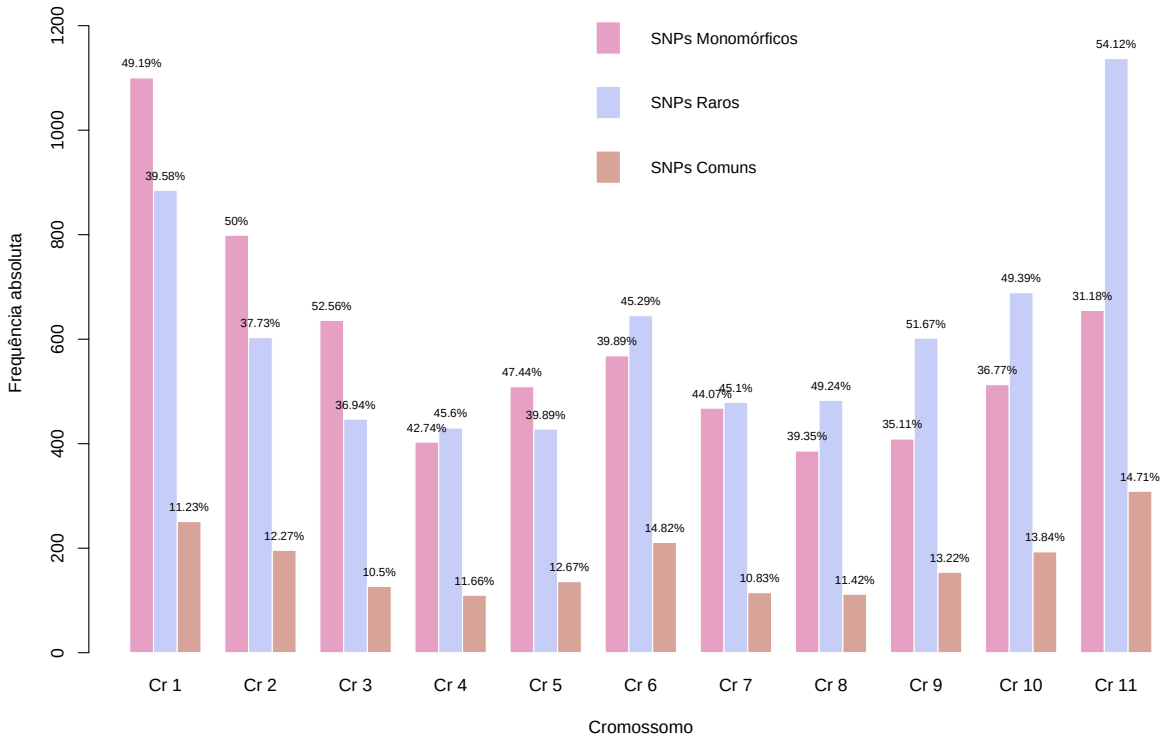
Tabela 4 – SNPs raros na amostra.

SNP	Cromossomo	Gene	β	MAF real	MAF dados
C1S6533	1	ARNT	0.589734	0.011478	0.0050215
C1S6540		ARNT	0.323662	0.001435	0.0028694
C1S3181		ELAVL4	0.795093	0.000717	0.0007174
C4S1861	4	KDR	0.598271	0.002152	0.0021521
C4S1873		KDR	0.715613	0.000717	0.0014347
C4S1884		KDR	0.318125	0.020803	0.0193687
C4S1890		KDR	0.417977	0.002152	0.0007174
C4S4935		VEGFC	1.405290	0.000717	0.0222382
C6S2981	6	VEGFA	1.130450	0.002152	0.0329986
C13S320	13	FLT1	0.180470	0.001435	0.0035868
C13S431		FLT1	0.732566	0.017217	0.0186514
C13S514		FLT1	0.549816	0.000717	0.0021521
C13S522		FLT1	0.623466	0.027977	0.0093257
C14S1734	14	HIF1A	0.220448	0.012195	0.0007174
C19S4831	19	HIF3A	0.265181	0.000717	0.0021521

O equilíbrio de Hardy-Weinberg é um princípio fundamental da genética populacional que descreve como as frequências alélicas e genotípicas permanecem constantes de geração em geração em uma população ideal, na ausência de forças evolutivas. Este princípio estabelece que, em uma população suficientemente grande, em que os acasalamentos ocorrem de maneira aleatória, sem mutação, migração ou seleção natural, as frequências alélicas de um locus bialélico serão estáveis. As frequências dos genótipos podem ser expressas pelas equações p^2, $2pq$ e q^2, em que p e q são as frequências dos alelos A e a, respectivamente. Este modelo nulo fornece uma base teórica para detectar desvios que possam indicar a presença de forças evolutivas atuando sobre a população ou erros de sequenciamento genético em população humana.

No contexto dos estudos de associação genética, a verificação do equilíbrio de Hardy-Weinberg é um passo crítico na análise de variantes genéticas. Através do teste qui-quadrado, é possível comparar as frequências genotípicas observadas com as esperadas sob a hipótese de equilíbrio. A rejeição da hipótese nula sugere que o locus não está em equilíbrio, possivelmente devido a fatores como seleção, deriva genética, mutação ou fluxo gênico. Neste estudo, utilizamos o pacote `genetics` no R para calcular a estatística do teste para cada SNP, com um nível de significância fixado em 10^{-6}.

Figura 2 – Gráfico de barras da frequência dos SNPs monomórficos, comuns e raros, presentes nos cromossomos de 1 a 11.



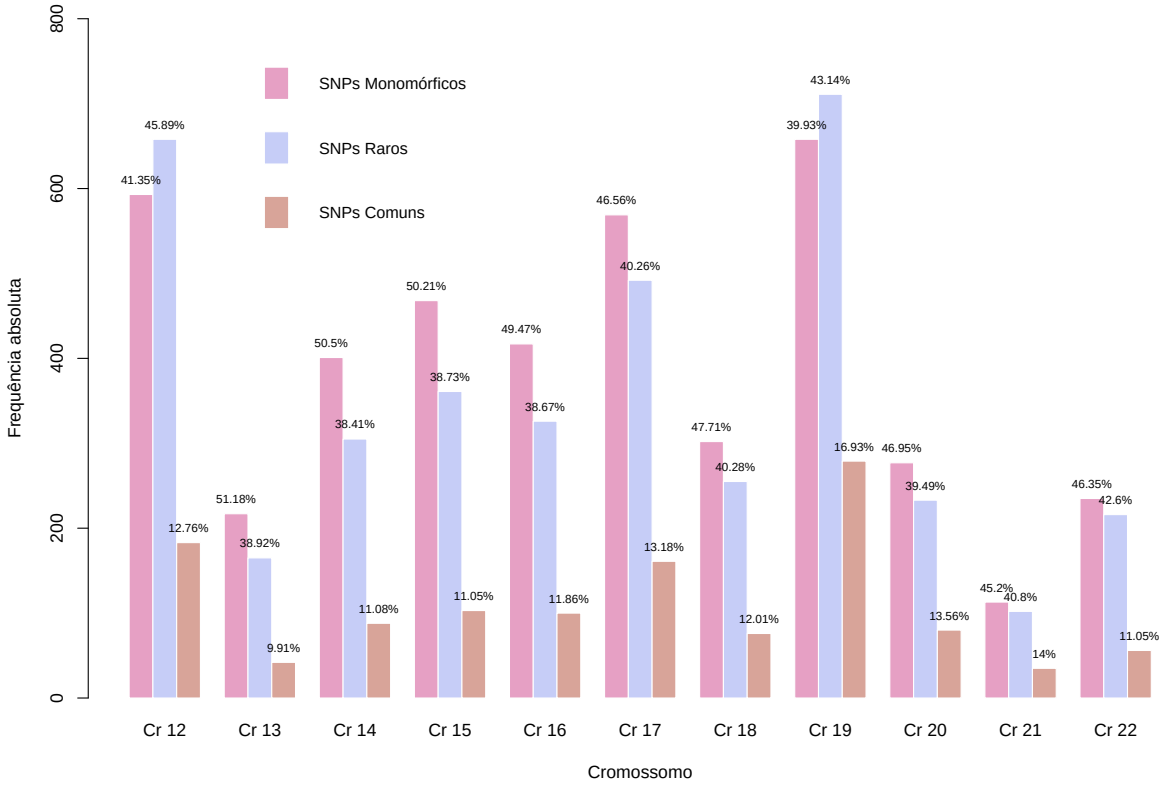
Após a filtragem dos SNPs raros da base de dados, testamos se cada SNP está em equilíbrio de Hardy-Weinberg (HWE), utilizando o teste de hipótese mencionado com o teste qui-quadrado. Na hipótese nula, assumimos que o SNP está em equilíbrio. Portanto, para *p*-valores pequenos, rejeitamos a hipótese nula. Observamos que para nenhum SNP rejeitamos a hipótese nula de equilíbrio de Hardy-Weinberg, permanecendo todos e assegurando a robustez dos resultados, minimizando vieses que poderiam comprometer a interpretação dos dados.

Desse modo, analisamos os SNPs presentes em cada um dos 22 cromossomos de acordo com seus MAFs, com o intuito de verificar a distribuição dessas frequências e a proporção de SNPs monomórficos, comuns e raros em cada um dos cromossomos do conjunto de dados GAW17. Os gráficos de barras das frequências de SNPs monomórficos, raros e comuns por cromossomo, estão disponíveis nas Figuras 2 e 3. Os boxplots das frequências do alelo menor dos SNPs raros que permaneceram após a filtragem, por cromossomo, estão disponíveis nas Figuras 4 e 5 e suas medidas resumo podem ser encontradas na Tabela 6.

Conforme ilustrado nas Figuras 2 e 3, 43.54% dos marcadores são raros (10652 SNPs), ou seja, apresentam frequência do alelo menor igual ou inferior a 5% e superior a 0%. Mais de 56% dos marcadores são monomórficos (43.72%) ou comuns (12.74%). Portanto, metodologias tradicionais de seleção de variáveis geralmente não apresentam bons resultados nesse contexto.

Desse modo, métodos de colapsagem de variantes raras foram utilizados neste conjunto de dados e seus desempenhos foram comparados.

Figura 3 – Gráfico de barras da frequência dos SNPs monomórficos, comuns e raros, presentes nos cromossomos de 12 a 22.



A Tabela 5 apresenta a distribuição dos SNPs por cromossomo no conjunto de dados do GAW17. A tabela categoriza os SNPs em monomórficos, raros e comuns, e fornece o total de SNPs para cada cromossomo. Observa-se que o cromossomo 1 possui o maior número de SNPs monomórficos (1100), enquanto o cromossomo 11 apresenta a maior quantidade de SNPs raros (1137). A distribuição total de SNPs revela que existem 10696 SNPs monomórficos, 10652 SNPs raros e 3117 SNPs comuns, totalizando 24465 SNPs analisados. Esta distribuição é fundamental para entender a variabilidade genética presente em cada cromossomo e para direcionar as análises subsequentes de seleção de variantes raras, foco desse trabalho.

Ao analisarmos as Figuras 4 e 5, notamos que, para cada um dos 22 cromossomos autossômicos, a frequência do alelo menor assume valores extremos (em geral, acima de 2%), porém todos abaixo de 5%, por serem raros. Todos apresentam assimetria positiva e a mediana varia entre 0.36% e 0.79%. Através da Tabela 6, nota-se que 75% dos SNPs raros apresentam MAF abaixo de 1.36% e que 50% das variantes raras apresentam MAF abaixo de 0.57%. A média de MAF é de 1.02%.

Tabela 5 – Distribuição dos SNPs por Cromossomo.

Cromossomo	**Mono**	**Raros**	**Comuns**	**Total**
1	1100	885	251	2236
2	799	603	196	1598
3	636	447	127	1210
4	403	430	110	943
5	509	428	136	1073
6	568	645	211	1424
7	468	479	115	1062
8	386	483	112	981
9	409	602	154	1165
10	513	689	193	1395
11	655	1137	309	2101
12	593	658	183	1434
13	217	165	42	424
14	401	305	88	794
15	468	361	103	932
16	417	326	100	843
17	569	492	161	1222
18	302	255	76	633
19	658	711	279	1648
20	277	233	80	590
21	113	102	35	250
22	235	216	56	507
Total	10696	10652	3117	24465

Figura 4 – Boxplot da Frequência do Alelo Menor (MAF) para os cromossomos de 1 a 11.

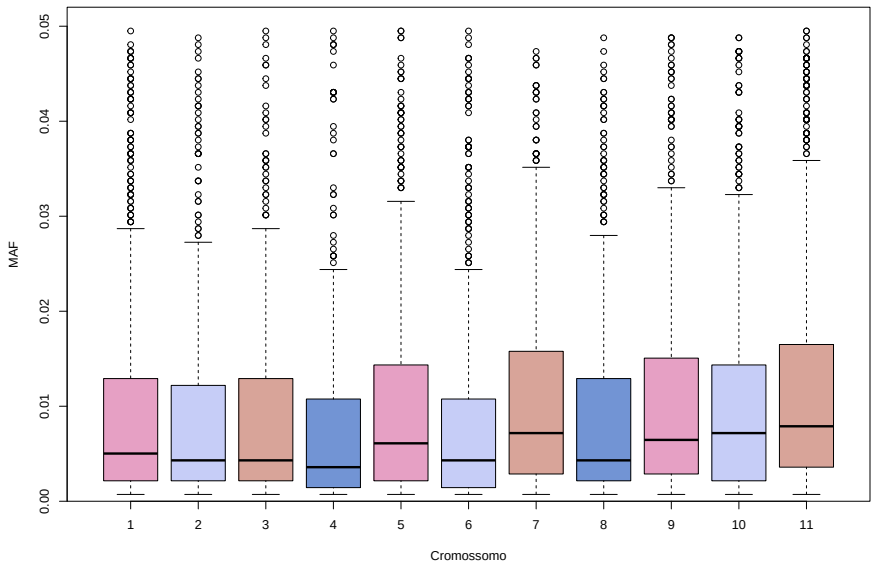


Figura 5 – Boxplot da Frequência do Alelo Menor (MAF) para os cromossomos de 12 a 22.

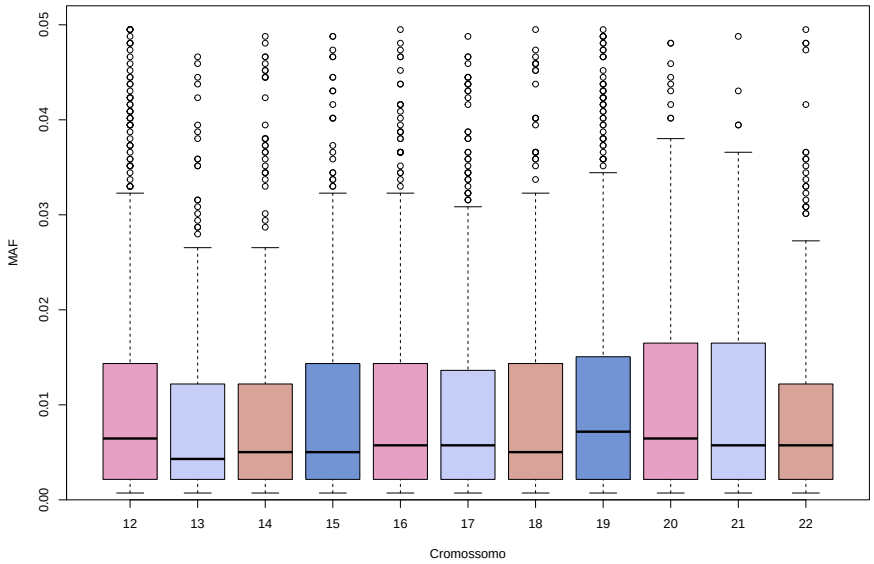


Tabela 6 – Resumo da Frequência do Alelo Menor (MAF) dos 10652 SNPs Raros.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.0007174	0.0021521	0.0057389	0.0102416	0.0136298	0.0494979

4.2 Métrica de desempenho para seleção de SNPs

Para avaliar e comparar a eficácia das metodologias empregadas na seleção de SNPs associados, foram utilizadas métricas de desempenho denominadas acurácia, especificidade e sensibilidade (CHUN; KELES, 2010).

A matriz de confusão é uma ferramenta essencial para avaliar o desempenho de um modelo de seleção, especialmente no contexto da seleção de SNPs ou genes significativos. Ela permite a visualização das seleções feitas pelo método em comparação com os valores reais, facilitando a análise dos resultados. No contexto desse estudo, utilizaremos tais métricas na seleção dos genes significativos, descritos na Tabela 4.

Tabela 7 – Matriz de Confusão.

	Real Significativo (8 genes)	**Real Não Significativo (2559 genes)**
Predito Significativo	Verdadeiro Positivo (VP)	Falso Positivo (FP)
Predito Não Significativo	Falso Negativo (FN)	Verdadeiro Negativo (VN)

Neste contexto, a matriz de confusão, exemplificada na Tabela 7, pode ser organizada da seguinte maneira:

- **Verdadeiros Positivos (VP)**: O número de genes que foram corretamente identificados como significativos pelo método.
- **Falsos Positivos (FP)**: O número de genes que foram incorretamente identificados como significativos pelo método.
- **Verdadeiros Negativos (VN)**: O número de genes que foram corretamente identificados como não significativos pelo método.
- **Falsos Negativos (FN)**: O número de genes que foram incorretamente identificados como não significativos pelo método.

Com base nesses valores, podemos calcular várias métricas de desempenho:

- **Acurácia**: A proporção de todas as seleções corretas (tanto verdadeiros positivos quanto verdadeiros negativos) em relação ao número total de genes existentes. É calculada como:

$$\text{Acurácia} = \frac{VP+VN}{VP+FP+VN+FN}.$$

A acurácia nos fornece uma visão geral do desempenho do modelo, indicando a proporção de previsões corretas.

- **Sensibilidade**: A proporção de verdadeiros positivos em relação ao número total de genes que são realmente significativos. É calculada como:

$$\text{Sensibilidade} = \frac{VP}{VP+FN}.$$

A sensibilidade mede a capacidade do modelo de identificar corretamente os genes significativos, ou seja, a capacidade de detectar verdadeiros positivos.

- **Especificidade**: A proporção de verdadeiros negativos em relação ao número total de genes que são realmente não significativos. É calculada como:

$$\text{Especificidade} = \frac{VN}{VN+FP}.$$

A especificidade mede a capacidade do modelo de identificar corretamente os genes não significativos, ou seja, a capacidade de evitar falsos positivos.

Essas métricas são cruciais para entender o equilíbrio entre identificar corretamente os genes significativos e evitar a identificação errônea de genes não significativos. Nesse estudo, essas métricas nos permitem avaliar a eficácia das metodologias aplicadas na seleção de SNPs raros significativos, auxiliando na interpretação e validação dos resultados obtidos.

A acurácia é descrita como a proporção de acertos, em relação ao total de elementos existentes, ou seja, é a soma dos valores na diagonal principal da matriz de confusão, dividido pela soma de todos os valores na matriz. Matematicamente, isso é expresso por:

$$\text{Acurácia} = \frac{\text{Número total de genes corretamente classificados}}{\text{Número total de genes}}.$$

Definimos a sensibilidade como a proporção de genes corretamente identificados, em relação ao total de elementos verdadeiros que são efetivamente significativos. Matematicamente, isso é expresso por:

$$\text{Sensibilidade} = \frac{\text{Número de genes corretamente selecionados}}{\text{Número total de genes com efeito em Q1}}.$$

Por outro lado, a especificidade é calculada como a proporção de genes corretamente identificadas como não significativos dentre os genes que realmente não exercem influência em Q1, refletindo a capacidade do teste de identificar corretamente as não associações. Essa métrica é quantificada pela fórmula:

$$\text{Especificidade} = \frac{\text{Número de genes corretamente não selecionados}}{\text{Número total de genes sem efeito em Q1}}.$$

Essas métricas são particularmente aplicáveis em contextos de dados simulados, como o dataset GAW17 utilizado neste estudo, em que a verdadeira natureza dos genes que abrigam SNPs significativos é conhecida *a priori*. No contexto presente, sabemos que o número total de genes é 2567, dos quais 8 representam genes com efeito em Q1 e 2559 representam genes sem efeito em Q1.

Dado que, naturalmente, nesses casos temos dados desbalanceados, a acurácia pode não ser uma métrica de avaliação adequada nesse cenário. Portanto, torna-se necessário dar maior importância à análise das outras duas métricas: sensibilidade e especificidade. A seleção ideal de variáveis ocorre quando ambas, sensibilidade e especificidade, atingem o valor unitário. No entanto, frequentemente observa-se um *trade-off* entre essas medidas; aumentos em especificidade geralmente resultam em reduções correspondentes em sensibilidade, e vice-versa. Para SNPs comuns, essa avaliação de desempenho é comumente realizada individualmente. Como o foco desse trabalho envolve os SNPs raros, o desempenho será avaliado para conjuntos de SNPs, agrupados por genes aos quais pertencem, analisando se tais conjuntos foram corretamente identificados como significativos ou não.

4.3 Criação da variável de colapsagem

Conforme apresentado na Seção 3.1.1, a variável que utiliza a ideia central do método de Burden consiste em agregar a informação dos SNPs raros dentro da unidade de análise através da contagem dos alelos raros para cada unidade amostral, utilizando essa nova pontuação de variação genética como variável comum nos modelos MLM ou *Grid-LMM*. Dessa forma, criamos

a variável D_i conforme a Equação (3.5), em que, para o i-ésimo indivíduo e k-ésima variante:

$$w_{ik} = \begin{cases} 0, & \text{se os dois alelos forem de maior frequência } (AA); \\ 1, & \text{se um dos alelos for de maior frequência e o outro de menor frequência } (Aa \text{ ou } aA); \\ 2, & \text{os dois alelos forem de menor frequência } (aa). \end{cases}$$

Na abordagem escolhida nesse estudo, essa pontuação é somada dentro de cada gene ao nível de unidade amostral e será substituída na matriz que corresponde aos efeitos fixos de cada modelo específico na Seção 5.2.

CAPÍTULO

5

RESULTADOS

Nesta seção, apresentamos os resultados obtidos ao aplicarmos as diferentes metodologias estudadas na análise dos dados familiares do GAW17. Como o foco de nosso estudo é a seleção de variantes raras, conduzimos quatro análises distintas. A primeira consiste na aplicação do *FamSKAT-RC*, que realiza um teste de hipótese sobre a variância comum dos coeficientes aleatórios. As demais análises consideram a colapsagem do genótipo (aproximação do teste de Burden) e são ajustadas utilizando três abordagens diferentes: *Grid-LMM*, *Grid-LMM* com LASSO e, finalmente, o Modelo Linear Misto (MLM) com LASSO.

5.1 *FamSKAT-RC*

Essa metodologia não nos permite testar se os SNPs raros são ou não significativos individualmente, pois o teste de hipóteses realizado é sobre a variância comum dos coeficientes aleatórios. Sendo assim, para cada cromossomo autossômico, realizamos os testes dos SNPs separadamente, tendo por blocos os genes. Ao rejeitarmos a hipótese nula, sabemos que pelo menos uma das variantes do respectivo gene possui coeficiente de regressão diferente de zero. Caso um gene seja selecionado, performamos novo teste dividindo o gene em blocos na tentativa que identificar blocos menores de SNPs raros associados ao fenótipo. Para os SNPs que são únicos em seus respectivos genes, agrupamos todos eles e os dividimos em novos blocos de cinco.

Consideramos que, ao rejeitar a hipótese nula nos sub-blocos, todos os SNPs considerados no teste possuem efeito relevante no fenótipo de interesse, Q1. Apesar disso, em relação ao cálculo da especificidade e sensibilidade, o faremos ao nível do gene. Aqui ressaltamos que se o foco for identificar SNPs raros associados ao fenótipo e não apenas grupos de SNPs raros, as metodologias apresentadas nas Seções 3.1.2, 3.1.2.1 e 3.1.4 podem ser aplicadas depois que grupos de SNPs são selecionados.

Na versão mais recente da função descrita por Lee (2023), que utilizamos para implementar o método *FamSKAT* no software R, o argumento que se refere aos pesos v_{ks} apresenta como valor default a densidade da distribuição $Beta(1,25)$ calculada no MAF do SNP em análise.

Nesse estudo, como realizamos o teste para os SNPs raros ao nível de cada gene, consideramos o nível de significância α de 0.05 corrigido por Bonferroni para verificar quais são significantes. A correção foi realizada da seguinte forma para cada um dos 22 cromossomos autossômicos:

$$\alpha_{BF} = \frac{\alpha}{\text{número de genes do respectivo cromossomo testado}}. \tag{5.1}$$

Para garantir a robustez dos resultados obtidos em nossos testes, aplicamos um nível de significância de $\alpha = 5\%$, corrigido por Bonferroni, nos demais testes de significância. Esta correção é essencial para controlar a taxa de falsos positivos quando múltiplas comparações são realizadas simultaneamente. A correção de Bonferroni ajusta o nível de significância padrão dividindo-o pelo número de testes realizados, reduzindo assim a probabilidade de detectar associações espúrias. Essa abordagem é particularmente importante em estudos genômicos, nos quais o grande número de variantes testadas aumenta o risco de resultados falsamente positivos. Portanto, ao adotar essa correção rigorosa, asseguramos que os SNPs e genes identificados como significativos em nossas análises sejam verdadeiramente relevantes para o fenótipo estudado.

O modelo identificou, para $\phi = 1$ (considerando apenas SNPs raros), apenas um gene no cromossomo autossômico 6, o gene VEGFA, que abriga 4 SNPs (C6S2958, C6S2970, C6S2981 e C6S3010). Ao dividirmos o gene em dois blocos e ajustarmos o *FamSKAT* novamente, o segundo bloco de SNPs foi selecionado (C6S2981 e C6S3010), subgrupo este que realmente contém uma das variantes significativas conhecidas: C6S2981 (Tabela 4).

Tabela 8 – Matriz de Confusão para a abordagem *FamSKAT*.

	Real Significativo	**Real Não Significativo**
Selecionado Significativo	VP = 1	FP = 0
Selecionado Não Significativo	FN = 7	VN = 2559

Dos 2567 genes, sabemos que apenas 8 deles abrigam SNPs raros que possuem efeito significativo sobre Q1 e apenas 1 deles fora selecionado nesse método (verdadeiro positivo); logo 2559 genes não abrigam nenhum SNP raro com efeito significativo em Q1. Desse modo, nenhum gene foi identificado incorretamente como significativo pela metodologia utilizada (falso positivo). Essas informações são melhor sintetizadas na matriz de confusão (Tabela 8).

Para o nível de significância de 5% corrigido por Bonferroni, a acurácia (A), a sensibilidade (S) e a especificidade (E) são, respectivamente:

$$A = \frac{2560}{2567} = 0.9973, \quad S = \frac{1}{8} = 0.125 \quad \text{e} \quad E = \frac{2559}{2559} = 1.$$

Os valores acima possuem a seguinte interpretação: o método classificou corretamente 99.73% do total de genes existentes, 12.5% dos genes que são realmente significativos são identificados pelo método, enquanto 100% dos genes que não possuem efeito importante sobre o fenótipo Q1 não são identificados pelo método.

5.2 MLM com Burden adaptado

Nessa nova abordagem, que consiste numa aproximação do Burden (Eq. 3.5), utilizamos uma nova variável com a colapsagem do genótipo das variantes raras que representa o número de alelos raros que apareceram nos SNPs de cada gene. Serviremo-nos do *Grid-LMM* e os testes de significância associados para selecionar os grupos de SNPs raros mais relevantes (*p-valor* e LASSO), assim como o MLM combinado com o LASSO.

5.2.1 Grid-LMM com p-valor

Nessa metodologia, foram realizados testes de hipótese para verificar a significância do coeficiente associado ao genótipo colapsado de cada gene em modelo linear misto conforme Seção 2.2.4, de Runcie e Crawford (2019). Os resultados obtidos nesses testes são expostos a seguir e comparados com os 8 genes que são realmente significativos para o fenótipo de interesse, Q1. O método *Grid-LMM* está disponível no software R através da biblioteca `GridLMM`[1].

Ao rejeitarmos a hipótese nula, sabemos que pelo menos uma das variantes do respectivo gene possui coeficiente de regressão diferente de zero. O nível de significância utilizado em cada um dos testes nessa primeira análise foi 0.05, corrigido por Bonferroni como no método anterior.

Empregando o nível de significância de 5%, o modelo identificou apenas um gene significativo no cromossomo autossômico 22 com *p-valor* de 0.000478, SEC14L3, que abriga 10 SNPs em sua totalidade. Trata-se de falso positivo, visto que é de nosso conhecimento que nenhum gene do cromossomo autossômico 22 tem efeito em Q1.

Dos 2567 genes, sabemos que apenas 8 deles abrigam SNPs raros que possuem efeito significativo sobre Q1 e nenhum deles fora selecionado nesse método. Logo, 2559 genes não abrigam nenhum SNP raro com efeito significativo em Q1. Um gene foi identificado incorretamente como significativo pela metodologia utilizada (falso positivo). Essas informações são melhor sintetizadas na matriz de confusão para a abordagem com *Grid-LMM* (Tabela 9).

Tabela 9 – Matriz de confusão para a abordagem *Grid-LMM*.

	Real Significativo	**Real Não Significativo**
Selecionado Significativo	VP = 0	FP = 1
Selecionado Não Significativo	FN = 8	VN = 2558

[1] Disponível no link: https://github.com/deruncie/GridLMM

Para o nível de significância de 5% corrigido por Bonferroni, a acurácia (A), a sensibilidade (S) e a especificidade (E) são, respectivamente:

$$A = \frac{2558}{2567} = 0.9965, \quad S = \frac{0}{8} = 0 \quad \text{e} \quad E = \frac{2558}{2559} = 0.9996.$$

Os valores acima possuem a seguinte interpretação: o método classificou corretamente 99.65% do total de genes existentes, 0% dos genes que são realmente significativos são identificados pelo método, enquanto 99.96% dos genes que não possuem efeito importante sobre Q1, não são identificados pelo método.

5.2.2 *Grid-LMM com LASSO*

Os resultados obtidos através do *Grid-LMM* combinado com LASSO para cada cromossomo são apresentados a seguir e comparados com os genes que realmente influenciam Q1.

Inicialmente, testamos todos os genes conjuntamente por cromossomo, resultando na seleção de 476 genes. Todos os genes escolhidos foram então analisados novamente em uma segunda etapa, restando 74 deles.

Outra abordagem empregada foi a análise dos genes por cromossomo em blocos de 50, o que resultou na seleção de 313 genes. Estes genes foram testados novamente por cromossomo, e apenas 10 permaneceram no ajuste final.

Observamos que, ao ajustarmos o modelo com todos os genes de cada cromossomo, um número maior de genes é selecionado. A Tabela 10 apresenta a quantidade de genes selecionados em ambas as abordagens e em cada etapa.

No cromossomo 1, que contém 2 genes significativos, a primeira etapa da abordagem que considera todos os genes selecionou um dos genes significativos (ARNT, que contém 2 SNPs significativos). No entanto, o gene ELAVL4, que abriga um SNP significativo, não foi selecionado; na segunda etapa, nenhum gene foi selecionado. Na abordagem por blocos, 25 genes foram selecionados na primeira etapa, mas nenhum deles tem efeito em Q1.

No cromossomo 2, a abordagem que considera todos os genes selecionou os mesmos 22 genes em ambas as etapas. Contudo, sabe-se que não há nenhum SNP associado ao fenótipo Q1 nesse cromossomo.

No cromossomo 4, a abordagem que considera todos os genes selecionou 3 genes na primeira etapa, dois dos quais abrigam SNPs significativos: KDR e VEGFC. O gene VEGFC contém o SNP C4S4935, enquanto o gene KDR abriga outros 4 SNPs, todos com efeito em Q1. No entanto, nenhum gene foi selecionado na segunda etapa dessa abordagem. Na abordagem por blocos, apenas 1 gene não significativo foi selecionado na primeira etapa.

Tabela 10 – Quantidade de genes selecionados pelo *Grid-LMM* com LASSO.

Cr	**genes**	**Todos os genes**		**Em blocos de 50**	
		Etapa 1	**Etapa 2**	**Etapa 1**	**Etapa 2**
1	184	51	0	25	0
2	107	22	22	13	0
3	89	23	0	11	0
4	72	3	0	1	0
5	90	17	0	10	0
6	85	25	0	26	0
7	154	25	0	20	0
8	156	0	0	9	9
9	228	41	0	32	0
10	216	37	0	18	0
11	397	77	51	47	0
12	172	32	0	19	0
13	27	1	1	1	1
14	56	22	0	22	0
15	45	0	0	0	0
16	60	16	0	0	0
17	98	22	0	20	0
18	35	0	0	0	0
19	191	42	0	19	0
20	41	12	0	12	0
21	20	0	0	0	0
22	44	8	0	8	0
Total	2567	476	74	313	10

No cromossomo 6, onde está localizado um SNP significativo, o gene que o abriga (VEGFA) foi selecionado na primeira etapa em ambas as abordagens. Entretanto, nenhum gene foi selecionado na segunda etapa de ambas as abordagens.

No cromossomo 13, que contém 27 genes, o gene FLT1, que abriga 4 SNPs significativos, foi selecionado já no primeiro passo em ambas as abordagens.

No cromossomo 14, que contém um SNP significativo, 22 genes foram selecionados na primeira etapa em ambas as abordagens, mas nenhum deles corresponde ao gene alvo. O mesmo ocorre no cromossomo 19.

No cromossomo 15, que não tem efeito em Q1, nenhum gene foi selecionado em todas as abordagens, assim como nos cromossomos 18 e 21.

A acurácia (A), a sensibilidade (S) e especificidade (E) atingidas por esses métodos também são expostas a seguir para ambas abordagens e suas etapas, bem como as respectivas tabelas (11, 12, 13 e 14) e interpretações:

- **Como todos os genes:**
  - *Etapa 1:*

$$A = \frac{2093}{2567} = 0.8153, \quad S = \frac{5}{8} = 0.625 \quad \text{e} \quad E = \frac{2088}{2559} = 0.816;$$

ou seja, o método classificou corretamente 81.53% do total de genes existentes; 62.5% dos genes que são realmente significativos são identificados pelo método, enquanto 81.6% dos genes que não possuem efeito importante sobre Q1 não são identificados pelo método.

Tabela 11 – Matriz de confusão para *Grid-LMM* com LASSO: Etapa 1 com todos os genes.

	Real Significativo	**Real Não Significativo**
Selecionado Significativo	VP = 5	FP = 471
Selecionado Não Significativo	FN = 3	VN = 2088

  - *Etapa 2:*

$$A = \frac{2489}{2567} = 0.9696, \quad S = \frac{2}{8} = 0.25 \quad \text{e} \quad E = \frac{2487}{2559} = 0.972;$$

ou seja, o método classificou corretamente 96.96% do total de genes existentes; 25% dos genes que são realmente significativos são identificados pelo método, enquanto 97.2% dos genes que não possuem efeito importante sobre Q1 não são identificados pelo método.

Tabela 12 – Matriz de confusão para *Grid-LMM* com LASSO: Etapa 2 com todos os genes.

	Real Significativo	**Real Não Significativo**
Selecionado Significativo	VP = 2	FP = 72
Selecionado Não Significativo	FN = 6	VN = 2487

- **Em blocos de 50 genes:**
  - *Etapa 1:*

$$A = \frac{2250}{2567} = 0.8765, \quad S = \frac{2}{8} = 0.25 \quad \text{e} \quad E = \frac{2248}{2559} = 0.878;$$

ou seja, o método classificou corretamente 87.65% do total de genes existentes; 25% dos genes que são realmente significativos são identificados pelo método, enquanto 87.8% dos genes que não possuem efeito importante sobre Q1 não são identificados pelo método.

Tabela 13 – Matriz de confusão para *Grid-LMM* com LASSO: Etapa 1 em blocos de 50 genes.

	Real Significativo	**Real Não Significativo**
Selecionado Significativo	VP = 2	FP = 311
Selecionado Não Significativo	FN = 6	VN = 2248

– *Etapa 2:*

$$A = \frac{2553}{2567} = 0.9945, \quad S = \frac{2}{8} = 0.25 \quad \text{e} \quad E = \frac{2551}{2559} = 0.997;$$

ou seja, o método classificou corretamente 99.45% do total de genes existentes; 25% dos genes que são realmente significativos são identificados pelo método, enquanto 99.7% dos genes que não possuem efeito importante sobre Q1 não são identificados pelo método.

Tabela 14 – Matriz de confusão para *Grid-LMM* com LASSO: Etapa 2 em blocos de 50 genes.

	Real Significativo	**Real Não Significativo**
Selecionado Significativo	VP = 2	FP = 8
Selecionado Não Significativo	FN = 6	VN = 2551

Observamos que a abordagem que ajusta todos os genes por cromossomo apresenta maior sensibilidade na primeira etapa, com a seleção de 5 genes influentes. No entanto, demonstra menor especificidade se comparada à segunda etapa. Em contraste, a abordagem com blocos de 50 genes apresenta melhor desempenho na segunda etapa, mantendo a sensibilidade enquanto aumenta a especificidade, devido à menor quantidade de falsos positivos.

5.2.3 MLM com LASSO

Diante dos ajustes anteriores, principalmente com abordagem LASSO, que não se apresentaram muito estáveis, foi necessário adotar uma nova abordagem para a seleção de SNPs associados e predição de efeitos aleatórios, conforme apresentado na Seção 2.2.6. Utilizamos o Modelo Linear Misto (MLM) com o Burden adaptado, empregando a biblioteca `glmnet` do R para realizar o ajuste do LASSO. Esta abordagem foi aplicada sem a estrutura do *Grid-LMM*, tanto para 5 *folds* quanto para 10 *folds* (o valor padrão para o *Grid-LMM* é 10 *folds*). A utilização da penalização LASSO permite lidar com a alta dimensionalidade dos dados, promovendo a seleção de variáveis e com essa nova abordagem visamos melhorar a estabilidade das estimativas.

Os resultados obtidos através do ajuste do MLM combinado com LASSO para cada cromossomo são apresentados a seguir e comparados com os genes que realmente influenciam o fenótipo Q1.

Realizamos o ajuste de todos os genes conjuntamente para cada cromossomo autossômico, variando apenas o parâmetro *nfolds*. Genes foram selecionados em todos os cromossomos autossômicos. Para *nfolds*= 5, foram selecionados 83 genes, enquanto que, para *nfolds*= 10, foram selecionados 87 genes, sendo que estes incluem os mesmos 83 genes identificados no ajuste com *nfolds*= 5. A Tabela 15 apresenta a quantidade de genes selecionados em ambas as abordagens por cromossomo autossômico.

Tabela 15 – Quantidade de genes selecionados pelo MLM com LASSO.

Cr	**genes**	***nfolds*=5**	***nfolds*=10**
1	184	9	9
2	107	3	5
3	89	4	4
4	72	4	4
5	90	1	1
6	85	3	3
7	154	5	5
8	156	4	4
9	228	6	7
10	216	1	1
11	397	8	9
12	172	4	4
13	27	1	1
14	56	1	1
15	45	4	4
16	60	4	4
17	98	2	2
18	35	2	2
19	191	9	9
20	41	2	2
21	20	1	1
22	44	5	5
Total	2567	83	87

Em ambas as abordagens, apenas um dos genes influentes em Q1 foi selecionado: o gene KDR, localizado no cromossomo 4, que contém 4 SNPs significativos (C4S1861, C4S1873, C4S1884 e C4S1890). Este gene foi consistentemente identificado como significativo tanto na abordagem com *nfolds*= 5 quanto com *nfolds*= 10, com os coeficientes estimados em 0.2349 e 0.2453, respectivamente. No entanto, os demais genes selecionados consistem em falsos positivos, uma vez que não apresentam associação real com o fenótipo Q1.

A acurácia (A), a sensibilidade (S) e especificidade (E) atingidas por esses métodos também são expostas a seguir para ambas abordagens, bem como as respectivas tabelas (16 e 17) e interpretações:

- ***nfolds*= 5:**

$$A = \frac{2478}{2567} = 0.9653, \quad S = \frac{1}{8} = 0.125 \quad \text{e} \quad E = \frac{2477}{2559} = 0.968;$$

ou seja, o método classificou corretamente 96.53% do total de genes existentes; 12.5% dos genes que são realmente significativos são identificados pelo método, enquanto 96.8% dos genes que não possuem efeito importante sobre o fator de risco quantitativo Q1 não são identificados pelo método.

Tabela 16 – Matriz de confusão para MLM com LASSO: *nfolds*= 5.

	Real Significativo	**Real Não Significativo**
Selecionado Significativo	VP = 1	FP = 82
Selecionado Não Significativo	FN = 7	VN = 2477

- ***nfolds***= 10**:**

$$A = \frac{2474}{2567} = 0.9638, \quad S = \frac{1}{8} = 0.125 \quad \text{e} \quad E = \frac{2473}{2559} = 0.966;$$

ou seja, o método classificou corretamente 96.38% do total de genes existentes; 12.5% dos genes que são realmente significativos são identificados pelo método, enquanto 96.6% dos genes que não possuem efeito importante sobre o fator de risco quantitativo Q1 não são identificados pelo método.

Tabela 17 – Matriz de confusão para MLM com LASSO: *nfolds*= 10.

	Real Significativo	**Real Não Significativo**
Selecionado Significativo	VP = 1	FP = 86
Selecionado Não Significativo	FN = 7	VN = 2473

Através dessa abordagem, observamos um melhor desempenho no ajuste do Modelo Linear Misto (MLM) combinado com LASSO, utilizando o parâmetro *nfolds*=5. Esse ajuste apresenta a mesma sensibilidade que a configuração com *nfolds*=10 (que é o valor utilizado no *Grid-LMM* com LASSO), mas com especificidade superior.

Uma estratégia potencial para aumentar a especificidade do Modelo Linear Misto (MLM) combinado com LASSO seria implementar uma segunda etapa de seleção, considerando apenas os genes identificados na fase inicial, similar ao procedimento realizado com o *Grid-LMM* com LASSO. Outra abordagem alternativa, também utilizada no *Grid-LMM* com LASSO, consiste em analisar inicialmente os genes em blocos de menor tamanho, em vez de considerar todos os genes de um mesmo cromossomo simultaneamente, e conduzir a seleção final em uma etapa subsequente. Essas estratégias poderão aprimorar a precisão e a eficiência da seleção de SNPs influentes em características de interesse.

5.3 Comparação dos resultados obtidos

O tempo computacional requerido para a execução de cada método variou consideravelmente. O método *FamSKAT-RC* apresentou o maior tempo médio de processamento, com aproximadamente 23 minutos por cromossomo. Em contraste, o método GridLMM com cálculo de p-valor demonstrou ser significativamente mais eficiente, exigindo cerca de 2 minutos por cromossomo. Os demais métodos avaliados consumiram menos de 1 minuto por cromossomo,

destacando-se como opções computacionalmente mais viáveis para a análise de grandes volumes de dados genômicos.

Com o propósito de comparar o desempenho das metodologias na seleção dos genes que influenciam o fenótipo Q1 considerando dados de família provenientes do banco de dados GAW17 e SNPs raros, analisamos o número de genes identificados como significativos, bem como os valores de acurácia, sensibilidade e especificidade obtidos a partir de cada um dos métodos estudados anteriormente. Os resultados são apresentados na Tabela 18.

Tabela 18 – Comparação de métodos para seleção de genes influentes.

Método	**FamSKAT**	**Grid-LMM**	**Grid-LMM com LASSO**				**MLM com LASSO**	
		com p-valor	**Todos os genes**		**Blocos de 50**		***nfolds*=5**	***nfolds*=10**
			Etapa 1	**Etapa 2**	**Etapa 1**	**Etapa 2**		
Acurácia	0.9973	0.9965	0.8153	0.9696	0.8765	0.9945	0.9653	0.9638
Sensibilidade	0.1250	0.0000	0.6250	0.2500	0.2500	0.2500	0.1250	0.1250
Especificidade	1.0000	0.9996	0.8160	0.9720	0.8780	0.9970	0.9680	0.9660
Nº genes Selecionados	8	1	476	74	310	10	83	87

Em todos os métodos que envolvem testes de hipótese, utilizamos um nível de significância de 5%, corrigido por Bonferroni. Através da Tabela 18, observamos que o *FamSKAT-RC* apresenta as melhores métricas de acurácia e especificidade, dado que não apresenta nenhum falso positivo. Em contrapartida, a sensibilidade deste método é relativamente baixa, indicando que o modelo não identifica como significativos muitos dos genes que realmente influenciam a característica Q1. O gene identificado pelo *FamSKAT-RC* é o VEGFA, localizado no cromossomo 6, que contém um único SNP (C6S2981) com um efeito elevado (1.13045), sendo o segundo maior valor de efeito. Além disso, este SNP possui a maior frequência alélica menor (MAF) na amostra entre os SNPs associados.

O *Grid-LMM* baseado em p-valor apresentou um desempenho insatisfatório, pois foi a única abordagem com sensibilidade nula, uma vez que selecionou apenas um gene e de forma incorreta (falso-positivo). Entre as abordagens que utilizam p-valor, o *FamSKAT-RC* demonstrou um desempenho superior.

O método com maior sensibilidade foi o *Grid-LMM* com LASSO ajustado com todos os genes por cromossomo autossômico, considerando apenas o primeiro ajuste (etapa 1). Selecionou corretamente 5 genes que conjuntamente abrigam 12 dos 15 SNPs com efeito no fenótipo Q1, porém apresentou a menor especificidade de todas as abordagens que utilizam o LASSO devido à grande quantidade de falsos positivos. A abordagem do *Grid-LMM* com LASSO ajustado em blocos de 50 em sua segunda etapa apresenta a segunda menor sensibilidade, porém com a maior especificidade dentre todas abordagens com o LASSO. O Modelo Linear Misto (MLM) combinado com LASSO apresentou melhor desempenho com *nfolds*=5, embora tenha demonstrado limitações em termos de sensibilidade. Esta metodologia selecionou apenas o gene

KDR, localizado no cromossomo autossômico 4, que contém quatro SNPs significativos. Estes SNPs, em conjunto, representam um valor de efeito total de 2.05, contribuindo significativamente para a variação observada no fenótipo Q1.

Devido à simplicidade de implementação em comparação com o *Grid-LMM* com LASSO, uma alternativa promissora para futuras análises, como dito anteriormente, é conduzir estudos adicionais utilizando o Modelo Linear Misto (MLM) combinado com LASSO. Essa abordagem permitirá uma avaliação mais detalhada de seu desempenho, incluindo a realização de múltiplas etapas de seleção e a aplicação de blocos com um número reduzido de SNPs raros, de forma análoga ao que foi executado para o *Grid-LMM* com LASSO. Essas análises adicionais poderão fornecer insights valiosos sobre a eficiência e robustez do MLM com LASSO em diferentes cenários de seleção de SNPs raros.

CAPÍTULO

6

CONCLUSÃO E DISCUSSÃO

Através dos estudos GWAS existentes (com base em variantes comuns) verifica-se que inúmeras metodologias estatísticas foram utilizadas com satisfatório êxito. Porém, quando consideramos as variantes raras, os mesmos métodos não se mostraram tão eficientes para as análises, o que pode ser consequência de vários fatores: tamanhos de amostra menores resultantes do maior custo de sequenciamento, anotação incompleta de variantes fora da parte codificadora dos genes e falta de poder dos testes. Por tal razão, nos últimos anos, uma grande quantidade de pesquisas foi realizada no desenvolvimento de novos testes para análise de associação ampla com variantes raras.

Mais pesquisas são necessárias sobre identificação e seleção de variantes. Estamos no meio de significativo aumento de dados de sequenciamento de todo o genoma e sabemos pouco sobre a construção de conjuntos coesos de variantes raras contendo uma grande proporção de variantes causais (os conjuntos são muito esparsos). Ademais, não existe uma estratégia clara para analisar os dados da sequência do genoma inteiro, e esta é uma questão importante para pesquisas futuras (aliar variantes raras e comuns, por exemplo).

Iniciamos esse trabalho com a introdução do Modelo Linear Misto (MLM) que tem uma característica importante para trabalhar com dados que apresentem correlação ou outro tipo de dependência entre as unidades, como medidas obtidas de indivíduos relacionados (familiares), caso muito frequente nos dados de sequenciamento genético.

Verificamos na Seção 2.2.3 a metodologia em que se baseia o *FamSKAT-RC*. Tal modelo abrange conjuntamente variantes raras e variantes comuns com base em MLM, trazendo um grande diferencial em comparação com as demais técnicas em que temos que trabalhar com as classes de variantes separadamente. Ainda, utiliza-se do teste de componente de variância SKAT para colapsagem dos SNPs. Em suma, trata-se de uma extensão do SKAT para dados familiares com a possibilidade de testar SNPs comuns e raros conjuntamente.

O *Grid-LMM* (Seção 2.2.4) traz uma otimização do algoritmo para o ajuste repetido

de MLM complexo, permitindo inclusive interações genótipo-ambiente, e ainda possibilitando sua aplicação em diferentes ferramentas e técnicas de seleção, das quais destacamos o LASSO. Diferencia-se das demais metodologias principalmente devido à flexibilidade do número de termos nos efeitos aleatórios.

Além deles, também descrevemos um dos modelos mais completos em termos de controle de estratificação populacional em MLM, o *fastGWA* (Seção 2.2.2), bastante robusto e eficiente em termos computacionais. Interessante notar que o modelo proposto pelos autores (JIANG *et al.*, 2019) tem por foco estimar $\hat{\beta}_{snp}$, mas $\mathbf{X}_{snp}$ não traz a colapsagem de variantes raras. Por esse motivo e por fazer uma análise variante a variante, não analisamos seu desempenho nesse estudo, mas um ajuste interessante seria feito ao trazer a informação genotípica dos SNPs raros já colapsada para a matriz $\mathbf{X}_{snp}$ ao nível de gene, função ou outra característica agregadora, como propusemos para a adaptação do MLM com base na ideia do teste de Burden.

No contexto das metodologias de colapsagem, apresentamos uma adaptação dos Modelos Lineares Mistos (MLMs) baseada no teste de Burden, cujo objetivo é agregar informações de SNPs raros dentro de unidades de análise, utilizando a contagem de alelos raros para cada unidade amostral. Este método foi combinado com o *Grid-LMM* e o LASSO, proporcionando uma abordagem robusta para a identificação de SNPs associados em dados familiares.

Através das diversas metodologias aplicadas, comparou-se o número de SNPs identificados como significativos e os valores de sensibilidade e especificidade. Sob os dados analisados do GAW17 com dados familiares, observamos que o *FamSKAT-RC* apresentou as melhores acurácias e especificidade, devido à ausência de falsos positivos, mas com sensibilidade relativamente baixa. Em contrapartida, o *Grid-LMM* com p-valor teve desempenho inferior, sendo a única abordagem com sensibilidade nula, indicando um poder muito baixo na identificação de SNPs significativos.

A abordagem *Grid-LMM* com LASSO, ajustada com todos os genes por cromossomo autossômico, apresentou maior sensibilidade na etapa 1, selecionando corretamente cinco genes influentes. No entanto, essa abordagem apresentou a menor especificidade devido à grande quantidade de falsos positivos. A abordagem *Grid-LMM* com LASSO ajustada em blocos de 50 genes apresentou maior especificidade na segunda etapa, embora com sensibilidade reduzida. O MLM com LASSO teve melhor desempenho com *nfolds* = 5, mas com sensibilidade inferior às demais abordagens.

Concluímos que, embora as metodologias aplicadas tenham apresentado variações no desempenho, a combinação de modelos mistos com LASSO mostrou-se promissora na identificação e seleção de SNPs raros. A aplicação do LASSO em algumas etapas e considerando análise de blocos com menos variantes em cada etapa podem resultar em melhorias significativas.

Futuras pesquisas podem explorar outras alternativas para a agregação da informação genotípica das variantes raras, que não se limitem apenas à soma do número de alelos raros,

mas que também considerem diferentes pesos e direções dos seus efeitos. Por exemplo, a agregação genotípica pode ser realizada por meio de componentes principais ou análise fatorial dos SNPs raros (JIA, 2015). Em dados familiares, componentes ou fatores poderiam ser ajustados utilizando uma base de indivíduos pseudo-independentes (um subgrupo de indivíduos que podem ser considerados independentes dentro da amostra total) e, posteriormente, aplicados à base completa. Essas abordagens têm o potencial de proporcionar uma compreensão mais profunda da arquitetura genética de fenótipos complexos e de melhorar a robustez e a precisão das análises genéticas.

Além disso, metodologias que considerem interações entre SNPs e fatores ambientais podem fornecer uma compreensão mais profunda da arquitetura genética de fenótipos complexos. A integração de novas técnicas estatísticas e a combinação de abordagens existentes são caminhos promissores para avanços futuros na análise de dados genéticos.

REFERÊNCIAS

ADZHUBEI, I. A.; SCHMIDT, S.; PESHKIN, L.; RAMENSKY, V. E.; GERASIMOVA, A.; BORK, P.; KONDRASHOV, A. S.; SUNYAEV, S. R. A method and server for predicting damaging missense mutations. **Nature Methods**, Nature Publishing Group US New York, v. 7, n. 4, p. 248–249, 2010. Citado na página 43.

ALLEN, A. S.; BELLOWS, S. T.; BERKOVIC, S. F.; BRIDGERS, J.; BURGESS, R.; CAVALLERI, G.; CHUNG, S.-K.; COSSETTE, P.; DELANTY, N.; DLUGOS, D. *et al.* Ultra-rare genetic variation in common epilepsies: a case-control sequencing study. **The Lancet Neurology**, Elsevier, v. 16, n. 2, p. 135–143, 2017. Citado na página 21.

ALMASY, L.; BLANGERO, J. Multipoint quantitative-trait linkage analysis in general pedigrees. **The American Journal of Human Genetics**, Elsevier, v. 62, n. 5, p. 1198–1211, 1998. Citado na página 25.

ALMASY, L.; DYER, T. D.; PERALTA, J. M.; KENT, J. W.; CHARLESWORTH, J. C.; CURRAN, J. E.; BLANGERO, J. Genetic analysis workshop 17 mini-exome simulation. In: SPRINGER. **BMC proceedings**. [S.l.], 2011. v. 5, p. 1–9. Citado na página 51.

AMADEU, R. R.; GARCIA, A. A. F.; MUNOZ, P. R.; FERRÃO, L. F. V. AGHmatrix: genetic relationship matrices in R. **Bioinformatics**, Oxford University Press, v. 39, n. 7, p. btad445, 2023. Citado na página 27.

ASIMIT, J.; ZEGGINI, E. Rare variant association analysis methods for complex traits. **Annual Review of Genetics**, Annual Reviews, v. 44, p. 293–308, 2010. Citado na página 21.

BENAGLIA, T.; CHAUVEAU, D.; HUNTER, D. R. An em-like algorithm for semi-and non-parametric estimation in multivariate mixtures. **Journal of Computational and Graphical Statistics**, Taylor & Francis, v. 18, n. 2, p. 505–526, 2009. Citado na página 46.

BOMBA, L.; WALTER, K.; SORANZO, N. The impact of rare and low-frequency genetic variants in common disease. **Genome Biology**, BioMed Central, v. 18, n. 1, p. 1–17, 2017. Citado na página 21.

CAPANU, M.; BEGG, C. B. Hierarchical modeling for estimating relative risks of rare genetic variants: properties of the pseudo-likelihood method. **Biometrics**, Wiley Online Library, v. 67, n. 2, p. 371–380, 2011. Citado na página 48.

CHUN, H.; KELES, S. Simultaneous dimension reduction and variable selection with sparse partial least squares. **Journal of Royal Statistical Society B**, v. 72, n. 1, p. 3–25, 2010. Citado na página 61.

CIRULLI, E. T.; WHITE, S.; READ, R. W.; ELHANAN, G.; METCALF, W. J.; SCHLAUCH, K. A.; GRZYMSKI, J. J.; LU, J.; WASHINGTON, N. L. Genome-wide rare variant analysis for thousands of phenotypes in 54,000 exomes. **BioRxiv**, Cold Spring Harbor Laboratory, p. 692368, 2019. Citado na página 21.

CIRULLI, E. T.; WHITE, S.; READ, R. W.; ELHANAN, G.; METCALF, W. J.; TANUDJAJA, F.; FATH, D. M.; SANDOVAL, E.; ISAKSSON, M.; SCHLAUCH, K. A. *et al.* Genome-wide rare variant analysis for thousands of phenotypes in over 70,000 exomes from two cohorts. **Nature communications**, Nature Publishing Group UK London, v. 11, n. 1, p. 542, 2020. Citado na página 21.

DAVIES, R. B. Hypothesis testing when a nuisance parameter is present only under the alternative. **Biometrika**, Oxford University Press, v. 64, n. 2, p. 247–254, 1977. Citado na página 44.

DAVYDOV, E. V.; GOODE, D. L.; SIROTA, M.; COOPER, G. M.; SIDOW, A.; BATZOGLOU, S. Identifying a high fraction of the human genome to be under selective constraint using gerp++. **PLoS Computational Biology**, Public Library of Science San Francisco, USA, v. 6, n. 12, p. e1001025, 2010. Citado na página 41.

DEMIDENKO, E. **Mixed models: theory and applications with R**. [S.l.]: John Wiley & Sons, 2013. Citado na página 28.

GRIFFITHS, A.; MILLER, J.; SUZUKI, D.; LEWONTIN, R.; GELBART, W.; WESSLER, S. Introdução à genética. 9ª edição. **Guanabara Koogan**, 2009. Citado na página 23.

HENDERSON, C. R. Best linear unbiased estimation and prediction under a selection model. **Biometrics**, JSTOR, p. 423–447, 1975. Citado na página 29.

IONITA-LAZA, I.; CAPANU, M.; RUBEIS, S. D.; MCCALLUM, K.; BUXBAUM, J. D. Identification of rare causal variants in sequence-based studies: methods and applications to vps13b, a gene involved in cohen syndrome and autism. **PLoS genetics**, Public Library of Science San Francisco, USA, v. 10, n. 12, p. e1004729, 2014. Citado nas páginas 21, 42, 45, 46, 47 e 48.

IONITA-LAZA, I.; LEE, S.; MAKAROV, V.; BUXBAUM, J. D.; LIN, X. Family-based association tests for sequence data, and comparisons with population-based association tests. **European Journal of Human Genetics**, Nature Publishing Group, v. 21, n. 10, p. 1158–1162, 2013. Citado na página 44.

____. Sequence kernel association tests for the combined effect of rare and common variants. **The American Journal of Human Genetics**, Elsevier, v. 92, n. 6, p. 841–853, 2013. Citado na página 32.

IZBICKI, R.; SANTOS, T. M. dos. **Aprendizado de máquina: uma abordagem estatística**. [S.l.]: Rafael Izbicki, 2020. Citado nas páginas 37 e 38.

JIA, J. **Association analysis between binary traits and common or rare genetic variants on family-based data**. Tese (Doutorado) — University of Pittsburgh, 2015. Citado na página 79.

JIANG, J.; NGUYEN, T. **Linear and generalized linear mixed models and their applications**. [S.l.]: Springer, 2007. v. 1. Citado nas páginas 20 e 25.

JIANG, L.; ZHENG, Z.; QI, T.; KEMPER, K. E.; WRAY, N. R.; VISSCHER, P. M.; YANG, J. A resource-efficient tool for mixed model association analysis of large-scale data. **Nature Genetics**, Nature Publishing Group, v. 51, n. 12, p. 1749–1755, 2019. Citado nas páginas 20, 25, 29, 30 e 78.

KING, C. R.; RATHOUZ, P. J.; NICOLAE, D. L. An evolutionary framework for association testing in resequencing studies. **PLoS Genetics**, Public Library of Science San Francisco, USA, v. 6, n. 11, p. e1001202, 2010. Citado na página 42.

KIRCHER, M.; WITTEN, D. M.; JAIN, P.; O'ROAK, B. J.; COOPER, G. M.; SHENDURE, J. A general framework for estimating the relative pathogenicity of human genetic variants. **Nature genetics**, Nature Publishing Group US New York, v. 46, n. 3, p. 310–315, 2014. Citado na página 42.

KUNJI, K.; SAAD, M. famskatrc: Family sequence kernel association test for rare and common variants. **R package version**, v. 1, n. 0, 2017. Citado na página 33.

LEE, S.; WU, M. C.; LIN, X. Optimal tests for rare variant effects in sequencing association studies. **Biostatistics**, Oxford University Press, v. 13, n. 4, p. 762–775, 2012. Citado na página 43.

LEE, S. S. Skat package. **Disponıvel em: https://cran. r-project. org/web/packages/SKAT/-vignettes/SKAT. pdf. Acessado: 2024-02-17**, p. 08–06, 2023. Citado na página 66.

LETTRE, G. Rare and low-frequency variants in human common diseases and other complex traits. **Journal of Medical Genetics**, BMJ Publishing Group Ltd, v. 51, n. 11, p. 705–714, 2014. Citado na página 21.

LI, B.; LEAL, S. M. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. **The American Journal of Human Genetics**, Elsevier, v. 83, n. 3, p. 311–321, 2008. Citado na página 42.

LO, S.-H.; ZHENG, T. A demonstration and findings of a statistical approach through reanalysis of inflammatory bowel disease data. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 101, n. 28, p. 10386–10391, 2004. Citado na página 46.

MÄGI, R.; ASIMIT, J. L.; DAY-WILLIAMS, A. G.; ZEGGINI, E.; MORRIS, A. P. Genome-wide association analysis of imputed rare variants: application to seven common complex diseases. **Genetic Epidemiology**, Wiley Online Library, v. 36, n. 8, p. 785–796, 2012. Citado na página 21.

MORGENTHALER, S.; THILLY, W. G. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (cast). **Mutation Research/-Fundamental and Molecular Mechanisms of Mutagenesis**, Elsevier, v. 615, n. 1-2, p. 28–56, 2007. Citado na página 42.

NEALE, B. M.; RIVAS, M. A.; VOIGHT, B. F.; ALTSHULER, D.; DEVLIN, B.; ORHO-MELANDER, M.; KATHIRESAN, S.; PURCELL, S. M.; ROEDER, K.; DALY, M. J. Testing for an unusual distribution of rare variants. **PLoS Genetics**, Public Library of Science San Francisco, USA, v. 7, n. 3, p. e1001322, 2011. Citado na página 42.

NELSON, M. R.; WEGMANN, D.; EHM, M. G.; KESSNER, D.; JEAN, P. S.; VERZILLI, C.; SHEN, J.; TANG, Z.; BACANU, S.-A.; FRASER, D. *et al.* An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. **Science**, American Association for the Advancement of Science, v. 337, n. 6090, p. 100–104, 2012. Citado na página 47.

NICOLAE, D. L. Association tests for rare variants. **Annual Review of Genomics and Human Genetics**, Annual Reviews, v. 17, p. 117–130, 2016. Citado nas páginas 20, 21, 41 e 42.

PAN, W. Asymptotic tests of association with multiple snps in linkage disequilibrium. **Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society**, Wiley Online Library, v. 33, n. 6, p. 497–507, 2009. Citado na página 42.

PETROVSKI, S.; TODD, J. L.; DURHEIM, M. T.; WANG, Q.; CHIEN, J. W.; KELLY, F. L.; FRANKEL, C.; MEBANE, C. M.; REN, Z.; BRIDGERS, J. *et al.* An exome sequencing study to assess the role of rare genetic variation in pulmonary fibrosis. **American Journal of Respiratory and Critical Care Medicine**, American Thoracic Society, v. 196, n. 1, p. 82–93, 2017. Citado na página 21.

POVYSIL, G.; PETROVSKI, S.; HOSTYK, J.; AGGARWAL, V.; ALLEN, A. S.; GOLDSTEIN, D. B. Rare-variant collapsing analyses for complex traits: guidelines and applications. **Nature Reviews Genetics**, Nature Publishing Group UK London, v. 20, n. 12, p. 747–759, 2019. Citado nas páginas 19, 21, 24 e 42.

RUNCIE, D. E.; CRAWFORD, L. Fast and flexible linear mixed models for genome-wide genetics. **PLoS Genetics**, Public Library of Science San Francisco, CA USA, v. 15, n. 2, p. e1007978, 2019. Citado nas páginas 20, 25, 33, 35, 36, 37 e 67.

SAAD, M.; WIJSMAN, E. M. Combining family-and population-based imputation data for association analysis of rare and common variants in large pedigrees. **Genetic Epidemiology**, Wiley Online Library, v. 38, n. 7, p. 579–590, 2014. Citado nas páginas 20, 25, 31 e 32.

SIM, N.-L.; KUMAR, P.; HU, J.; HENIKOFF, S.; SCHNEIDER, G.; NG, P. C. Sift web server: predicting effects of amino acid substitutions on proteins. **Nucleic Acids Research**, Oxford University Press, v. 40, n. W1, p. W452–W457, 2012. Citado na página 43.

SINGER, J. M.; ANDRADE, D. d. Análise de dados longitudinais. **Simpósio Nacional de Probabilidade e Estatística**, Embrapa São Paulo, v. 7, 1986. Citado na página 29.

SUN, Y. V.; SUNG, Y. J.; TINTLE, N.; ZIEGLER, A. Identification of genetic association of multiple rare variants using collapsing methods. **Genetic Epidemiology**, Wiley Online Library, v. 35, n. S1, p. S101–S106, 2011. Citado nas páginas 21 e 41.

SVISHCHEVA, G. R.; BELONOGOVA, N. M.; AXENOVICH, T. I. FFBSKAT: fast family-based sequence kernel association test. **PloS One**, Public Library of Science San Francisco, USA, v. 9, n. 6, p. e99407, 2014. Citado na página 31.

TIBSHIRANI, R. Regression shrinkage and selection via the lasso. **Journal of the Royal Statistical Society: Series B (Methodological)**, Wiley Online Library, v. 58, n. 1, p. 267–288, 1996. Citado nas páginas 37 e 38.

WU, M. C.; LEE, S.; CAI, T.; LI, Y.; BOEHNKE, M.; LIN, X. Rare-variant association testing for sequencing data with the sequence kernel association test. **The American Journal of Human Genetics**, Elsevier, v. 89, n. 1, p. 82–93, 2011. Citado nas páginas 32 e 42.


ICMC
USP
SÃO CARLOS