UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Construção de Redes Complexas de Interações entre SNPs: Uma Abordagem Computacional e Estatística

Rafaela Miglinski Lucca

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

ICMC USP SÃO CARLOS

Rafaela Miglinski Lucca

Construção de Redes Complexas de Interações entre SNPs: Uma Abordagem Computacional e Estatística

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. Andressa Cerqueira

USP – São Carlos
Maio de 2026

UFSCar

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciências Exatas

Programa Interinstitucional de Pós-Graduação em Estatística.

Folha de Aprovação

Defesa de Dissertação de Conclusão de Curso de Mestrado da candidata Rafaela Miglinski Lucca, realizada em 09/03/2026.

Comissão Julgadora:

Prof(a). Dr(a). Andressa Cerqueira (DES-UFSCAR)

Prof(a). Dr(a). Júlia Maria Pavan Soler (IME-USP)

Prof(a). Dr(a). Sabrina Luzia Caetano (UNESP)

Prof(a). Dr(a). Jaqueline Lopes Pereira (FSP-USP)

Prof(a). Dr(a). Daiane Aparecida Zuanetti (UFSCar)

Prof(a). Dr(a). Alex Rodrigo dos Santos Sousa (UNICAMP)

O Relatório de Defesa produzido pelos membros da Comissão Julgadora encontra-se arquivado junto ao Programa de Pós-Graduação.

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

Lucca, Rafaela
L934c Construção de Redes Complexas de Interações entre SNPs: Uma Abordagem Computacional e Estatística / Rafaela Lucca; orientador Andressa Cerqueira. -- São Carlos, 2026.
93 p.

Dissertação (Mestrado - Programa Interinstitucional de Pós-graduação em Estatística) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2026.

1. ESTATÍSTICA. 2. ESTATÍSTICA APLICADA. 3. GENÉTICA ESTATÍSTICA. I. Cerqueira, Andressa, 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

Rafaela Miglinski Lucca

Construction of Complex Interaction Networks among SNPs: A Computational and Statistical Approach

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. Andressa Cerqueira

USP – São Carlos
May 2026

Este trabalho é dedicado à minha família
por ser minha base, meu refúgio e minha maior inspiração. Cada conquista minha carrega o amor, o apoio e os valores que vocês me ensinaram.

AGRADECIMENTOS

Agradecimentos

Gostaria de expressar minha profunda gratidão a todas as pessoas e instituições que me apoiaram ao longo desta jornada acadêmica.

Aos meus pais, Priscila Lucca e Giancarlo Lucca, por todo amor, incentivo e apoio incondicional em cada etapa da minha vida. À minha irmã, Beatriz Lucca, e ao meu sobrinho, Ravi Lucca, por trazerem alegria e motivação ao meu dia a dia. Aos meus avós, por todo carinho e ensinamentos que moldaram quem sou. À minha afilhada, Stella Miglinski, por ser uma fonte de inspiração e felicidade. À minha orientadora, Andressa Cerqueira, pelo acompanhamento atento, orientação valiosa e incentivo à pesquisa, contribuindo significativamente para meu desenvolvimento acadêmico e científico. Aos meus amigos da pós-graduação, por estarem ao meu lado ao longo dessa jornada, pela parceria, apoio e compreensão nos momentos mais desafiadores, e por tornarem esse caminho mais leve e significativo.

À Universidade Federal de São Carlos (UFSCar), especialmente ao Centro de Estatística, e à Universidade de São Paulo (USP), por meio do Instituto de Ciências Matemáticas e de Computação (ICMC), por oferecerem um ambiente acadêmico de excelência e pelas oportunidades que me proporcionaram ao longo deste mestrado.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo suporte financeiro, essencial para a realização deste trabalho e para o avanço da pesquisa científica no Brasil.

A todos que, de alguma forma, contribuíram para a realização deste trabalho, deixo aqui meu mais sincero agradecimento.

“O que sabemos, sabemos; o que não sabemos, precisamos descobrir.”
(Carl Friedrich Gauss)

RESUMO

LUCCA, R. M. **Construção de Redes Complexas de Interações entre SNPs: Uma Abordagem Computacional e Estatística**. 2026. 93 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, 2026.

Resumo

Os polimorfismos de nucleotídeo único (SNPs) representam a forma mais comum de variação genética no genoma humano e têm sido amplamente investigados em razão de seu impacto em características fenotípicas e na susceptibilidade a doenças complexas. Apesar dos avanços proporcionados pelos estudos de associação genômica ampla (GWAS), que testam cada SNP de forma marginal, essa abordagem tende a ignorar dependências entre loci, fenômeno que é conhecido como epistasia, o que pode deixar parte da variância genética do fenótipo inexplicada.

Diante desse cenário, este trabalho propõe uma abordagem em duas etapas para identificar interações entre SNPs associadas aos níveis de glicose sanguínea (mg/dL). Na primeira etapa, modelos de regressão linear simples são ajustados para pré-selecionar os SNPs com maior evidência de associação marginal com o fenótipo. Na segunda etapa, Florestas de Interação são aplicadas ao subconjunto selecionado para ranquear pares de SNPs por seu efeito de interação, medido pela métrica *Effect Importance Measure* (EIM). Os pares priorizados são então validados por meio de regressão linear com termos de interação. SNPs ausentes foram imputados por meio de cadeias de Markov e frações de recombinação, e os dados provêm do estudo ISA-Capital, com 698 participantes genotipados para 324.471 SNPs após controle de qualidade.

Os resultados do estudo de simulação demonstraram que as Florestas de Interação são capazes de recuperar de forma consistente os pares com interações inseridas artificialmente, tanto em cenários quantitativos quanto qualitativos. Na aplicação aos dados reais, a abordagem com pré-seleção por regressão (Caso 3) produziu os menores p-valores de interação entre as três estratégias avaliadas, com pares apresentando $p < 10^{-15}$, e permitiu a construção de redes de interação entre SNPs com estruturas conectadas e ciclos identificáveis. Esses achados reforçam o potencial da metodologia proposta para a detecção de epistasia em dados genômicos de alta dimensionalidade.

Palavras-chave: SNPs, Redes de Interação Genética, Florestas de Interação,Epistasia.

ABSTRACT

LUCCA, R. M. **Construction of Complex Interaction Networks among SNPs: A Computational and Statistical Approach**. 2026. 93 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, 2026.

Single nucleotide polymorphisms (SNPs) represent the most common form of genetic variation in the human genome and have been extensively studied due to their impact on phenotypic traits and susceptibility to complex diseases. Although genome-wide association studies (GWAS) have yielded important discoveries by testing each SNP marginally, this approach tends to overlook dependencies between loci — a phenomenon known as epistasis, which may leave part of the phenotypic genetic variance unexplained.

To address this limitation, the present work proposes a two-step approach for identifying SNP–SNP interactions associated with blood glucose levels (mg/dL). In the first step, simple linear regression models are fitted to pre-select SNPs with the strongest evidence of marginal association with the phenotype. In the second step, Interaction Forests are applied to the selected subset to rank SNP pairs by their interaction effect, as quantified by the *Effect Importance Measure* (EIM). Prioritized pairs are subsequently validated using linear regression models with interaction terms. Missing SNPs were imputed using Markov chains and recombination fractions, and the data originate from the ISA-Capital study, comprising 698 genotyped participants with 324,471 SNPs after quality control.

Simulation results demonstrated that Interaction Forests consistently recover pairs with artificially introduced interactions, under both quantitative and qualitative scenarios. In the application to real data, the regression-based pre-selection strategy (Case 3) yielded the lowest interaction p-values among the three approaches evaluated, with pairs reaching $p < 10^{-15}$, and enabled the construction of SNP interaction networks with connected components and identifiable cycles. These findings reinforce the potential of the proposed methodology for detecting epistasis in high-dimensional genomic data.

Keywords: SNPs, Genetic Interaction Networks, Interaction Forests,Epistasis.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

EIM	Effect Importance Measure
GWAS	Estudos de Associação Genômica Ampla, do inglês, *Genome-Wide Association Studies*
MDS	*Multidimensional Scaling*
OOB	Out-of-Bag
PCA	*Principal Component Analysis*
PropRandom	Proportional Randomization
SNPs	*Single Nucleotide Polymorphisms*
VIM	Variable Importance Measures

SUMÁRIO

CAPÍTULO

1

INTRODUÇÃO

O genoma humano é formado por longas cadeias de Ácido Desoxirribonucleico, *Deoxyribonucleic Acid* (DNA), que armazenam a informação genética necessária para coordenar processos fundamentais do funcionamento biológico, como embriogênese, desenvolvimento, crescimento, metabolismo e reprodução, conforme descrito por Nussbaum, McInnes e Willard 2016. Em cada célula nucleada, encontra-se uma cópia completa desse genoma, organizada em cromossomos localizados no núcleo celular. Esses cromossomos abrigam os genes, unidades funcionais de informação dispostas linearmente ao longo da molécula de DNA.

Em um mesmo locus (plural loci), podem existir versões alternativas denominadas alelos, que resultam de alterações permanentes na sequência de nucleotídeos (mutações), como apresentado em Griffiths *et al.* 2015. Tais mutações podem ou não ter impacto funcional, dependendo de sua natureza e localização.

Entre as formas mais comuns de variação genética estão os polimorfismos de nucleotídeo único, do inglês (*Single Nucleotide Polymorphisms* (SNPs)), que consistem na substituição de uma única base no DNA, por exemplo, citosina versus timina em uma posição específica. Essas variações podem influenciar diretamente a expressão ou a função gênica, afetando a suscetibilidade a fenótipos clínicos complexos, como discutido por Nussbaum, McInnes e Willard 2007. A investigação dessas associações entre variação genética e características fenotípicas é conduzida por meio dos Estudos de Associação Genômica Ampla, do inglês, *Genome-Wide Association Studies* (GWAS), que testam centenas de milhares a milhões de SNPs em larga escala.

O fluxo analítico dos GWAS envolve etapas como genotipagem, controle de qualidade e modelagem estatística. Um desafio central nesse processo é o controle da *estratificação populacional*, que se refere a diferenças de ancestralidade entre indivíduos. Caso não sejam devidamente ajustadas, essas diferenças podem gerar associações espúrias. Para mitigar esse risco, técnicas como *Principal Component Analysis* (PCA) e *Multidimensional Scaling* (MDS)

são empregadas para resumir padrões de estrutura populacional em componentes principais, os quais são incorporados como covariáveis nos modelos estatísticos, conforme orientado por Marees *et al.* 2018.

Embora alguns métodos de GWAS tradicionais sejam eficazes na detecção de efeitos aditivos ao testar cada SNP de forma marginal, essa abordagem tende a ignorar dependências entre loci. Tais dependências são conhecidas como epistasia, fenômeno em que o efeito de um alelo em um locus depende do genótipo em outro, conforme definição de Cordell 2009. A epistasia reflete a complexa organização dos genes em vias e redes biológicas, e sua incorporação nas análises pode revelar parte da chamada "herdabilidade ausente" observada em fenótipos complexos, uma vez que interações gênicas não modeladas podem inflar a discrepância entre a herdabilidade explicada por modelos puramente aditivos e a herdabilidade aparente, como discutido por Zuk *et al.* 2012. Além disso, pode oferecer pistas mais próximas dos mecanismos biológicos subjacentes do que análises puramente marginais.

Contudo, detectar epistasia em escala genômica representa um desafio computacional significativo, dado que o número de combinações possíveis entre pares de SNPs cresce quadraticamente, em ordem de $O(p^2)$, em que p denota o número de SNPs disponíveis no banco de dados. Por exemplo, considerando 24.929 SNPs, temos $\binom{24\,929}{2} = 310{,}891{,}156$ pares possíveis de SNPs, o que inviabiliza abordagens ingênuas de varredura exaustiva. Para enfrentar esse cenário, a literatura científica tem avançado em três frentes principais: algoritmos exaustivos cada vez mais eficientes, esquemas em duas etapas e métodos não paramétricos.

Historicamente, um marco importante foi o desenvolvimento do método de Multifactor-Dimensionality Reduction (MDR), proposto por Ritchie *et al.* 2001, que colapsa combinações multilocus em classes de risco sem impor um modelo paramétrico. Essa abordagem foi posteriormente consolidada com a introdução de validação cruzada e aplicação prática em dados reais, conforme ilustrado por Hahn, Ritchie e Moore 2003. Em seguida, surgiram extensões como o Odds Ratio-based Multifactor-Dimensionality Reduction (OR-MDR), que utiliza a razão de chances como medida contínua de risco Chung e Park 2007; o Model-Based Multifactor-Dimensionality Reduction (MB-MDR), que adapta o MDR para um arcabouço baseado em modelo e permite a análise de traços quantitativos Calle *et al.* 2008; e o Ordinal Multifactor-Dimensionality Reduction (OMDR), voltado para fenótipos ordinais Kim, Cho e Park 2013.

Em paralelo, abordagens baseadas em regressão ganharam destaque, com ferramentas como o PLINK, que recomenda iniciar com uma triagem rápida em tabelas 3×3 (`--fast-epistasis`) e, posteriormente, aplicar testes de regressão mais rigorosos nos pares candidatos, conforme descrito por Chang e Purcell 2025. Para tornar a busca exaustiva viável em estudos caso–controle, foi desenvolvido o BOOST, que acelera o teste de interação por meio de operações booleanas e testes de razão de verossimilhança, conforme apresentado por Wan *et al.* 2010. Revisões recentes convergem para uma estratégia em duas etapas: triagem para reduzir o espaço de busca e validação paramétrica com controle de multiplicidade e codificação apropriada, como discutido

por Niel *et al.* 2015.

É importante destacar que o termo *epistasia* pode ser interpretado sob duas perspectivas complementares. A primeira é a *biológica*, que se refere à interação funcional entre genes, proteínas ou elementos regulatórios em uma via. A segunda é a *estatística*, que se manifesta nos dados como um desvio do modelo aditivo ao incluir um termo de interação. Essa distinção é fundamental para interpretar resultados de GWAS e discutir a herdabilidade ausente, evitando confundir padrões estatísticos com mecanismos biológicos. A epistasia estatística, por sua vez, é sensível ao modelo e à codificação adotados (por exemplo, dosagem 0/1/2 versus modelo genotípico), podendo alterar a magnitude e a significância do termo de interação sem necessariamente invalidar a hipótese de uma interação biológica real, como exemplificado por Cordell 2002 e Phillips 2008.

Diante desse cenário, propomos uma abordagem estatística para contornar a alta dimensionalidade e identificar interações entre SNPs utilizando um procedimento de duas etapas. A proposta apresentada nesta dissertação baseia-se em um algoritmo derivado das florestas aleatórias Breiman 2001. Especificamente, utilizamos as *Interaction Forests*, que estendem o paradigma das *Random Forests* ao explorar divisões bivariadas e ranquear pares de SNPs por meio da métrica *Effect Importance Measure* (EIM), conforme descrito por Hornung e Boulesteix 2022.

1.0.1 Objetivos

O objetivo central deste trabalho é investigar associações entre SNPs e os níveis de glicose sanguínea, com ênfase na identificação de efeitos de interação entre pares de SNPs que possam estar associados com essa variável fenotípica. Para alcançar esse objetivo, adota-se uma abordagem metodológica em etapas:

- **Etapa 1: Pré-seleção de SNPs.** Como já mencionamos o número de combinações possíveis para testar pares de interação é muito grade. Nessa etapa, selecionamos primeiramente um subconjunto de n de SNPs. Esse subconjunto pode ser baseado em SNPs presentes na literatura ou SNPs que são selecionados a partir de algum método estatístico. Propomos selecionar os n SNPs que apresentam maior associação com o fenótipo de interesse;

- **Etapa 2: Pré- Seleção de Interações.** Para identificar todos os possíveis pares de interação no subconjunto de n SNPs utilizaremos o método de Florestas de Interação, proposto por Hornung e Boulesteix 2022. Esse método será utilizado para ordenar os pares de SNPs com maior efeito de interação. Serão selecionados os m pares de SNPs com maior efeito de interação;

- **Etapa 3: Identificação de Interações.** Com base nos m pares de SNPs que apresentaram maior efeito de interação na Etapa 2 testaremos as interação utilizando regressão linear

com termos de interação. Faremos esse teste já que, como discutiremos mais adiante, a medida de efeito de interação das Florestas de Interação deve ser interpretada com cautela, não sendo possível tomar alguma decisão apenas baseada nela. Nessa etapa reduzimos um grande número de combinações possíveis se considerássemos todos os SNPs do cromossomo para m pares.

A Figura 1 a seguir apresenta um fluxograma esquemático que resume as etapas metodológicas descritas, desde a base de dados inicial até a seleção final dos SNPs com efeitos de interação significativos. Ressaltamos que para $n = 400$ o tempo médio de execução das florestas de interação foi de 2 minutos e 58 segundos.

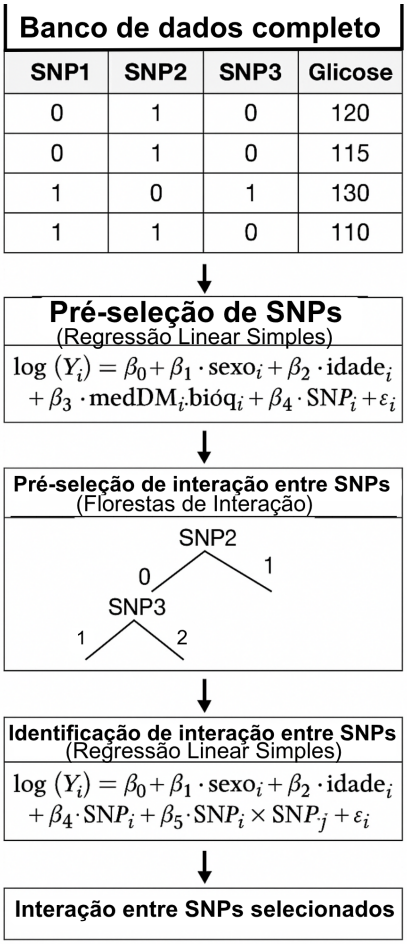

Figura 1 – Fluxograma do processo de análise e seleção de SNPs com base em regressão linear e Floresta de Interação.

Fonte: Elaborado pelo autor.

Para testarmos se essa metodologia proposta será capaz de identificar as interações entre SNPs propomos primeiramente um estudo de simulação. Neste estudo de simulação, a variável resposta foi construída artificialmente para incorporar interações genéticas previamente definidas entre pares de SNPs. Essa etapa teve por finalidade testar a capacidade do método de Florestas de Interação em encontrar associações significativas entre SNPs. Mais detalhes sobre o estudo de simulação podem ser encontrados no Capítulo 4.

1.0.2 Organização do trabalho

A estrutura deste trabalho está organizada de forma a conduzir o leitor desde os fundamentos teóricos até a aplicação prática da metodologia proposta. No Capítulo 2, são abordados os conceitos essenciais relacionados aos SNPs, sua origem, forma de codificação e influência sobre características fenotípicas. Ainda nesse capítulo, introduzem-se os GWAS, com uma explicação detalhada sobre seus objetivos, funcionamento e estratégias adotadas para identificar associações entre variantes genéticas e fenótipos de interesse. O Capítulo 3 é dedicado ao estudo das interações entre SNPs, discutindo sua importância biológica e os principais desafios metodológicos para sua detecção em contextos de alta dimensionalidade. Nesse mesmo capítulo, apresenta-se o algoritmo *Interaction Forests*, enfatizando seu funcionamento e o papel da métrica *Effect Importance Measure* na avaliação da relevância de efeitos univariados e bivariados.

No Capítulo 4, é desenvolvido um estudo de simulação com cenários controlados, construídos de modo a induzir interações entre SNPs e permitir a avaliação da capacidade do método em recuperar padrões epistáticos conhecidos. Em seguida, o Capítulo 5 apresenta a aplicação da metodologia proposta a dados reais, com foco na análise da associação entre variação genética e níveis de glicose. Por fim, o Capítulo 6 reúne as principais conclusões do estudo e discute possíveis extensões e perspectivas futuras da pesquisa.

CAPÍTULO

2

ESTUDO DE ASSOCIAÇÃO GENÔMICA

Neste capítulo, apresentamos os conceitos fundamentais relacionados aos SNPs, explicando sua origem, codificação e o impacto que podem exercer sobre fenótipos de interesse. Em seguida, introduzimos os Estudos de Associação Genômica Ampla (GWAS), detalhando seu funcionamento, objetivos e as abordagens utilizadas para identificar associações entre variantes genéticas e características fenotípicas.

2.1 Introdução aos SNPs

Os polimorfismos de nucleotídeo único (SNP) são a forma mais comum de variação genética no genoma humano. Um SNP ocorre quando um único nucleotídeo (A, C, G ou T) em uma posição específica da sequência de DNA apresenta variação entre indivíduos da mesma espécie. Essas variações, observadas entre os cromossomos homólogos, são essenciais para a diversidade genética e podem influenciar características fenotípicas, bem como a suscetibilidade a doenças. Para que uma variação seja considerada um SNP, ela deve estar presente em pelo menos 1% da população, diferenciando-se de mutações raras, que ocorrem esporadicamente em um pequeno número de indivíduos.

O DNA é a molécula fundamental da hereditariedade, responsável pelo armazenamento e transmissão da informação genética nos organismos vivos. Sua estrutura é composta por nucleotídeos, unidades formadas por três componentes principais: um grupo fosfato, um açúcar pentose (desoxirribose) e uma base nitrogenada. As bases nitrogenadas, que contêm átomos de nitrogênio em sua estrutura, são classificadas em dois grupos principais de acordo com sua composição química. As purinas, que possuem uma estrutura bicíclica, incluem a adenina (A) e a guanina (G), enquanto as pirimidinas, que possuem um único anel, englobam a citosina (C) e a timina (T). A organização dessas bases dentro da dupla hélice segue a regra de complementaridade, estabelecida por Chargaff 1950, onde a adenina sempre se emparelha com a timina por meio de duas ligações de hidrogênio, e a guanina se emparelha com a citosina através de três ligações

de hidrogênio. Essa complementaridade é essencial para garantir a replicação precisa do DNA durante a divisão celular e a manutenção da integridade genética ao longo das gerações. Os polimorfismos de nucleotídeo único correspondem a variações genéticas que ocorrem quando há a substituição de uma única base nitrogenada em uma posição específica da sequência de DNA. Por exemplo, em determinado *locus*, um indivíduo pode apresentar uma citosina (C), enquanto outro apresenta uma timina (T), caracterizando a presença de um SNP.

Cada SNP é composto por dois alelos possíveis: o alelo de referência, geralmente mais frequente na população, e o alelo alternativo, menos comum. Após a identificação dos SNPs, considera-se o par de bases presentes nos cromossomos homólogos de cada indivíduo, o que permite determinar seu genótipo naquele *locus* específico.

Para fins de análise estatística, os genótipos são codificados numericamente com base na combinação dos alelos nos dois cromossomos. Essa codificação segue a seguinte convenção: homozigose para o alelo de referência (aa) é representada por 0; heterozigose (Aa) por 1; e homozigose para o alelo alternativo (AA) por 2. Essa representação facilita a incorporação dos dados genéticos em modelos estatísticos, sendo amplamente utilizada nas análises subsequentes deste trabalho.

A Figura 2 ilustra, de forma esquemática, a ocorrência de diferentes SNPs ao longo de uma mesma região genômica.

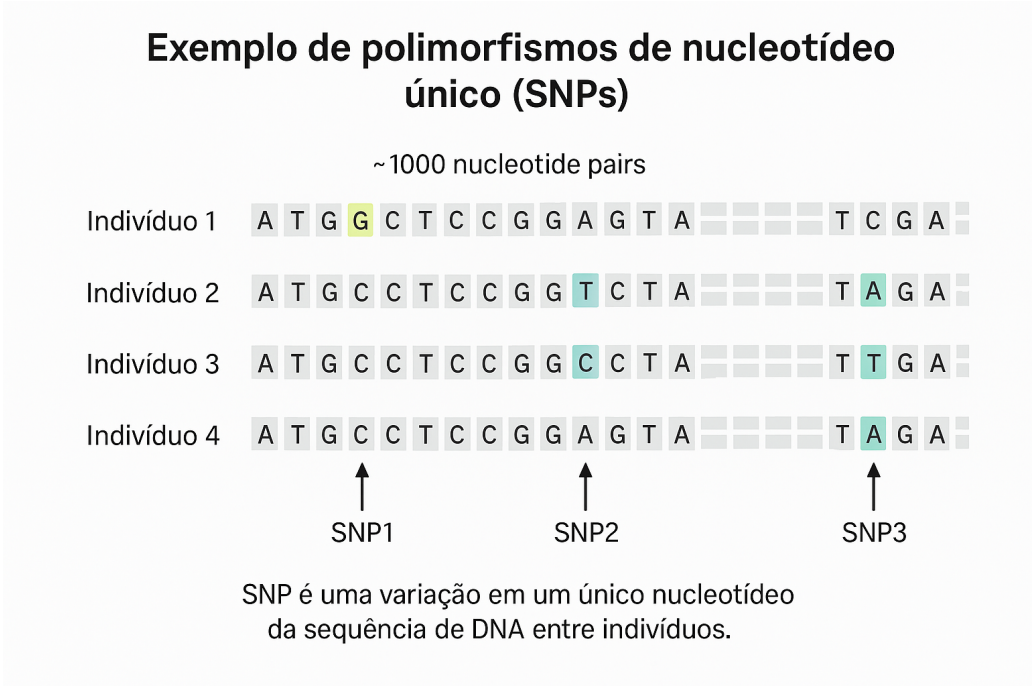

Figura 2 – Exemplo ilustrativo de três polimorfismos de nucleotídeo único (SNPs) ao longo de uma sequência de DNA. Fonte: Elaborado pela autora.

Antes de iniciar as análises com dados genômicos, é essencial aplicar um conjunto de filtros de controle de qualidade aos SNPs, com o objetivo de garantir a confiabilidade dos resultados estatísticos e evitar associações espúrias. Este processo segue um fluxo metodológico

rigoroso que abrange desde a filtragem de variantes raras até a reconstrução de dados perdidos. A Figura 3 sistematiza essa sequência de etapas, detalhando os critérios matemáticos e as justificativas para cada procedimento adotado.

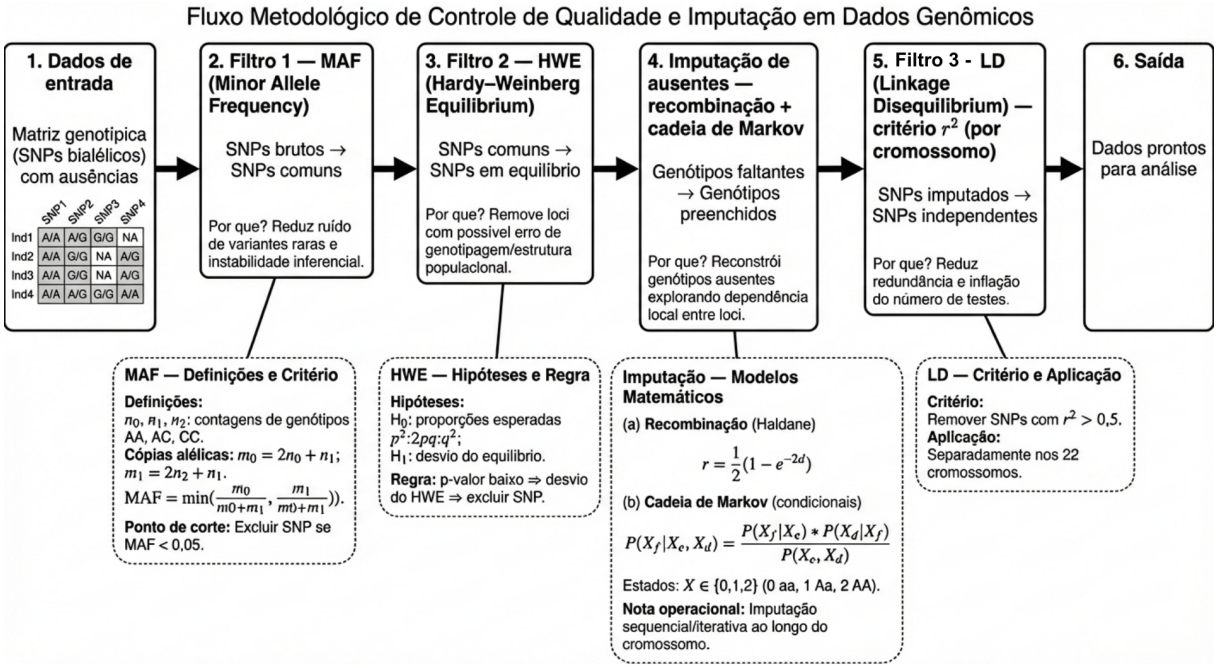

Figura 3 – Fluxo metodológico de controle de qualidade e imputação em dados genômicos. Fonte: Elaborado pela autora.

Conforme ilustrado na Figura 3, o processamento é dividido em seis estágios principais. Inicialmente, a Matriz Genotípica (1) é submetida ao Filtro de MAF (2), que remove SNPs com frequência do alelo menor inferior a 0,05, visando reduzir ruídos de variantes raras que poderiam causar instabilidade inferencial. Em seguida, o Filtro de HWE (3) elimina loci com desvios significativos do Equilíbrio de Hardy-Weinberg, o que previne a inclusão de variantes com possíveis erros de genotipagem ou problemas de estrutura populacional. A etapa de Imputação de ausentes (4) é crucial para lidar com as lacunas na matriz. Como detalhado no fluxograma, utiliza-se a função de recombinação de Haldane para calcular a distância genética e modelos de Cadeias de Markov para estimar os genótipos faltantes. Este processo explora a dependência local entre os loci vizinhos (esquerda X_e e direita X_d) para inferir probabilisticamente o estado do SNP ausente (X_f). Posteriormente, aplica-se o Filtro de LD (5), utilizando o critério de correlação r^2 por cromossomo. O objetivo desta etapa, conforme destacado na figura, é reduzir a redundância do banco de dados e mitigar a inflação do número de testes estatísticos, removendo SNPs altamente correlacionados. Por fim, obtém-se a Saída (6), composta pelos dados prontos para os testes de associação. A seguir, apresentamos de forma pormenorizada a fundamentação teórica de cada um desses critérios e os pontos de corte adotados.

O primeiro filtro corresponde à verificação da Frequência do Alelo Menor, que representa a proporção do alelo menos comum em uma população. Alelos muito raros tendem a ter baixa

capacidade de detectar associações confiáveis com fenótipos e são mais suscetíveis a erros de genotipagem, conforme discutido em Marees *et al.* 2018. Para um SNP bialélico com alelos A (referência) e C (alternativo), e contagens genotípicas n_0, n_1 e n_2 para os genótipos AA, AC e CC, respectivamente, podemos codificar os alelos como 0 (alelo de referência) e 1 (alelo variante). Nesse caso, n_0 corresponde ao número de indivíduos com genótipo $0/0$ (AA), n_1 ao número de indivíduos com genótipo $0/1$ (AC) e n_2 ao número de indivíduos com genótipo $1/1$ (CC). A partir desses genótipos, calculam-se os números de cópias de cada alelo na amostra:

$$m_0 = 2n_0 + 1n_1, \qquad m_1 = 2n_2 + 1n_1,$$

onde m_0 representa o número total de cópias do alelo de referência (A) e m_1 o número total de cópias do alelo alternativo (C). As frequências alélicas são então dadas por

$$\frac{m_0}{m_0+m_1} \quad \text{e} \quad \frac{m_1}{m_0+m_1},$$

e a frequência do alelo menos comum (Minor Allele Frequency, MAF) é definida como

$$\text{MAF} = \min\left\{\frac{m_0}{m_0+m_1}, \frac{m_1}{m_0+m_1}\right\}.$$

A literatura recomenda que o limiar de exclusão baseado em MAF seja adaptado ao tamanho amostral. Para amostras moderadas ($N \approx 10{,}000$), é comum utilizar um ponto de corte de 5%. Neste trabalho, SNPs com MAF inferior a 5% foram excluídos.

O segundo filtro avalia o Equilíbrio de Hardy–Weinberg, que verifica se as frequências genotípicas observadas estão em conformidade com aquelas esperadas sob ausência de seleção, mutação, migração ou deriva genética. A verificação é feita por meio de um teste qui-quadrado, com hipóteses dadas por:

$$H_0 : \begin{cases} \text{o SNP está em equilíbrio de Hardy–Weinberg;} \\ \text{as frequências genotípicas seguem as proporções } p^2 : 2pq : q^2. \end{cases}$$

$$H_1 : \begin{cases} \text{o SNP não está em equilíbrio de Hardy–Weinberg;} \\ \text{pelo menos uma das frequências genotípicas difere de } p^2, 2pq \text{ ou } q^2. \end{cases}$$

Valores baixos de p-valor correspondentes a este teste de hipóteses indicam desvio significativo do equilíbrio, podendo refletir erro de genotipagem ou estrutura populacional não modelada. Para traços quantitativos, recomenda-se excluir SNPs com p-valor inferior a 10^{-6} [Marees *et al.* 2018].

Além dos critérios de MAF e HWE, também foram removidos da base os SNPs e indivíduos com mais de 5% de dados ausentes, conforme práticas consolidadas de controle de qualidade em GWAS. Todos esses filtros foram aplicados utilizando a ferramenta *PLINK*, e seus efeitos estão resumidos na Figura 4.

A Figura 4 apresenta, para cada cromossomo, as contagens de SNPs em cada etapa do controle de qualidade: a barra cinza indica o total inicial de SNPs, as barras coloridas intermediárias indicam os SNPs removidos pelos filtros de MAF (laranja), HWE (amarelo) e dados ausentes (verde), e a barra vermelha ao final representa os SNPs com dados faltantes remanescentes. Observa-se que o filtro de MAF foi responsável pela maior redução no número de SNPs na maioria dos cromossomos, seguido pelo filtro de HWE. A proporção de dados ausentes foi relativamente uniforme entre os cromossomos, conforme também detalhado na Figura 5.

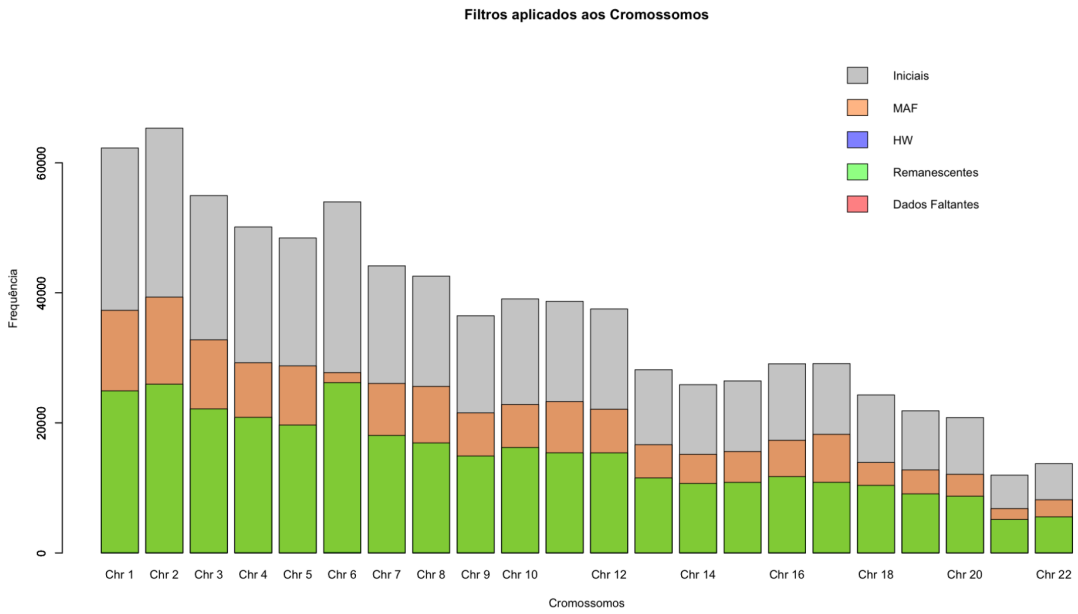

Figura 4 – Gráficos de barras representando, por cromossomo, o efeito sequencial dos filtros de controle de qualidade aplicados ao banco de dados: MAF (laranja), HWE (amarelo), dados ausentes (verde) e SNPs faltantes remanescentes (vermelho). As barras cinzas indicam o total inicial de SNPs antes da filtragem. Fonte: Elaborado pela autora.

Fonte: Elaborado pela autora.

2.2 Imputação de dados ausentes

A análise de dados genômicos frequentemente enfrenta o problema de dados ausentes, especialmente em estudos de larga escala que envolvem genotipagem de milhares ou milhões de SNPs. Esses dados ausentes podem surgir por diversas razões, como falhas na genotipagem, baixa cobertura de leitura em sequenciamentos de nova geração ou filtros de qualidade aplicados após o processamento inicial dos dados. A presença de SNPs ausentes compromete a integridade das análises estatísticas e pode reduzir a precisão dos modelos aplicados, tornando essencial a aplicação de técnicas de imputação. No presente estudo, os bancos de dados genômicos dos 22 cromossomos apresentavam uma quantidade significativa de dados ausentes. A ausência de informações em determinados loci prejudica a identificação de padrões de associação entre SNPs e a variável resposta. A Figura 5, apresentada a seguir, ilustra a proporção de dados ausentes por cromossomo.

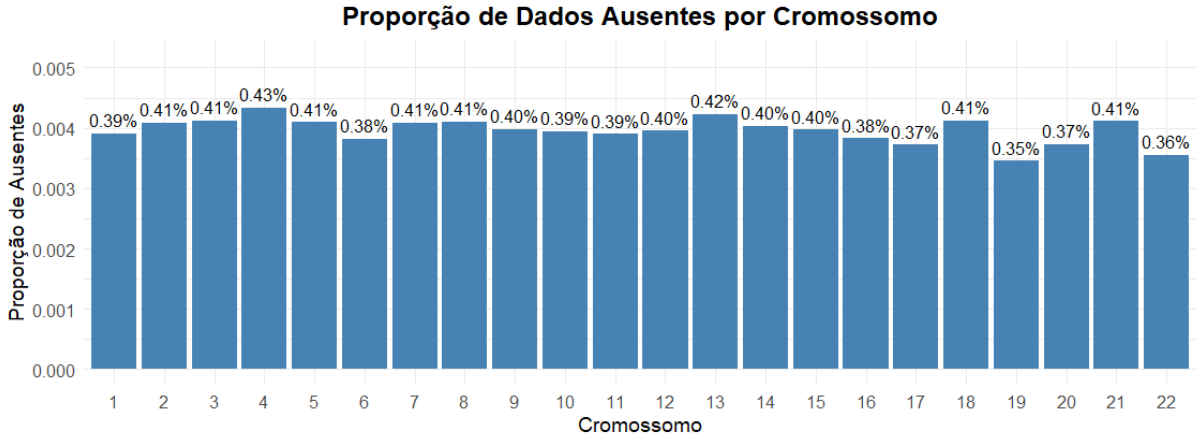

Figura 5 – Proporção de dados ausentes por cromossomo. O cálculo foi realizado dividindo a quantidade de dados ausentes pela multiplicação do número de linhas pelo número de colunas do banco de dados.

Para lidar com os dados ausentes, foi utilizada uma abordagem baseada na fração de recombinação e cadeias de Markov, conforme descrito por Stephens e Fisch 1998. O objetivo foi estimar os genótipos ausentes considerando a dependência entre SNPs vizinhos, respeitando a estrutura do genoma. A fração de recombinação (r) entre dois loci mede a probabilidade de que ocorra uma troca de material genético entre eles durante a meiose. Para calcular essa fração, foi utilizada a função de Haldane:

$$r = \frac{1}{2}(1 - e^{-2d})$$

onde:

- r representa a fração de recombinação entre os loci;
- d é a distância genética entre os dois loci, medida em unidades de Morgan (M).

Essa função assume que a recombinação ocorre de maneira independente entre as regiões do cromossomo. Como a fração de recombinação é proporcional à distância genética, SNPs próximos tendem a ser herdados juntos, enquanto SNPs mais distantes possuem maior probabilidade de sofrer recombinação. Após o cálculo da fração de recombinação, a imputação dos SNPs ausentes foi realizada por meio de probabilidades condicionais, modeladas por uma cadeia de Markov. Essa abordagem considera que o estado de um SNP ausente (X_f) pode ser estimado com base nos SNPs vizinhos observados (X_e e X_d), utilizando a seguinte equação:

$$P(X_f|X_e,X_d) = \frac{P(X_f|X_e)P(X_d|X_f)}{P(X_e,X_d)}$$

onde:

- X_f, X_e e X_d são variáveis aleatórias que representam os genótipos dos SNPs nas posições futura (ausente), esquerda e direita, respectivamente.

- Cada variável aleatória assume valores no espaço paramétrico $\{0,1,2\}$, correspondentes às codificações genotípicas: 0 para homozigose do alelo de referência (aa), 1 para heterozigose (Aa) e 2 para homozigose do alelo alternativo (AA).

Essa modelagem explora a dependência local entre os SNPs consecutivos, assumindo que a probabilidade de um genótipo ausente pode ser inferida com base nas relações probabilísticas com seus vizinhos genômicos, aproveitando a estrutura de correlação ao longo do cromossomo. Com essa modelagem, as probabilidades condicionais para um SNP ausente foram estimadas a partir dos genótipos conhecidos, levando em conta a fração de recombinação entre os loci adjacentes. Exemplos de cálculo incluem:

$$P(X_f = 0|X_e = 0, X_d = 0) = \frac{(1-r_1)^2(1-r_2)^2}{(1-r)^2}$$
$$P(X_f = 1|X_e = 0, X_d = 0) = \frac{2r_1(1-r_1)r_2(1-r_2)}{(1-r)^2}$$
$$P(X_f = 2|X_e = 0, X_d = 0) = \frac{r_1^2 r_2^2}{(1-r)^2}$$

onde:

- r_1 é a fração de recombinação entre o SNP ausente e o SNP à esquerda (X_e);

- r_2 é a fração de recombinação entre o SNP ausente e o SNP à direita (X_d).

A imputação foi aplicada de maneira sequencial, seguindo um processo iterativo. Inicialmente, para o primeiro SNP ausente, utilizou-se o SNP imediatamente anterior para estimar sua distribuição condicional. Nos casos em que o SNP ausente estava em uma posição intermediária, o algoritmo considerou tanto o SNP anterior quanto o seguinte para calcular as probabilidades condicionais. Para o último SNP ausente, a imputação foi realizada com base no penúltimo locus. Esse processo foi repetido sucessivamente até que todos os valores ausentes fossem imputados.

Após o controle de qualidade inicial, aplicamos o filtro de desequilíbrio de ligação aos 22 cromossomos, com o objetivo de reduzir a redundância entre SNPs altamente correlacionados. Em painéis genotípicos densos, variantes próximas podem apresentar correlação elevada, o que inflaciona o número de testes e compromete a independência das associações. Para mitigar esse problema, utilizamos o critério de correlação r^2, removendo SNPs com $r^2 > 0{,}5$. O procedimento foi aplicado separadamente a cada cromossomo, e os resultados estão apresentados na Figura 6.

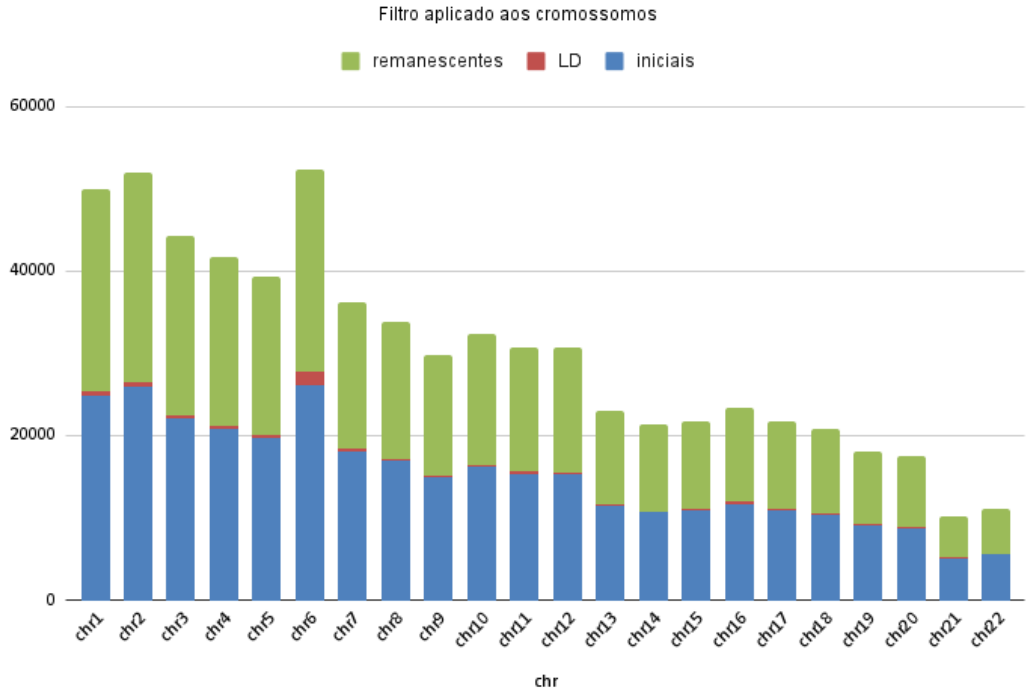

Figura 6 – Filtro de LD aplicado aos 22 cromossomos: contagens iniciais de SNPs (azul), SNPs removidos por desequilíbrio de ligação (vermelho) e SNPs remanescentes após o filtro (verde). O filtro foi aplicado com limiar $r^2 > 0{,}5$, separadamente para cada cromossomo.

2.3 Estudos de Associação Genômica

A relevância dos SNPs está diretamente ligada à sua influência na suscetibilidade a doenças. Estudos genômicos têm demonstrado que essas variações podem afetar a expressão de genes, modificar a funcionalidade de proteínas e impactar vias biológicas essenciais. Um exemplo clássico de associação entre SNPs e doenças é a variante do gene *APOE*, onde o alelo *APOE* $\varepsilon 4$ está fortemente relacionado a um risco aumentado para a doença de Alzheimer [Corder *et al.* 1993]. Da mesma forma, variantes nos genes *BRCA1* e *BRCA2* aumentam significativamente a suscetibilidade ao câncer de mama e ovário [Jara *et al.* 2017]. Além dessas condições, os SNPs também foram identificados como fatores-chave na genética da doença de Crohn, um distúrbio inflamatório intestinal crônico. Os estudos de associação genômica ampla frequentemente focam na busca de SNPs individuais associados ao risco da doença; no entanto, evidências recentes indicam que as interações entre SNPs podem ser fundamentais para entender a predisposição genética à doença de Crohn. Um estudo utilizando regressão lógica para modelar interações SNP-SNP identificou combinações específicas de variantes genéticas que contribuem para o risco da doença, revelando genes de susceptibilidade como *IL23R*, *NOD2* e *CYLD* [Dinu *et al.* 2012]. Tais descobertas foram possíveis graças ao desenvolvimento de métodos sistemáticos de análise genômica, como os GWAS, os quais serão detalhados a seguir. Os GWAS constituem uma estratégia amplamente utilizada para investigar relações entre variantes genéticas, como os

SNPs, e características fenotípicas de interesse ou doenças complexas. Esses estudos analisam grandes volumes de dados genéticos, geralmente envolvendo milhões de SNPs genotipados em milhares de indivíduos, com o objetivo de identificar variantes associadas de forma significativa a um determinado fenótipo. Uma possível metodologia empregada no GWAS baseia-se em ajustes de modelos de regressão, nos quais os efeitos dos SNPs são testados individualmente. Devido à grande quantidade de testes realizados simultaneamente, é fundamental aplicar correções para múltiplas comparações, como a correção de Bonferroni ou o controle da taxa de descobertas falsas, a fim de evitar associações espúrias. Além disso, o rigor no controle de qualidade dos dados genéticos e fenotípicos, bem como na imputação de genótipos ausentes, é essencial para garantir a confiabilidade dos resultados. A seguir, descreve-se cada etapa envolvida na condução de um estudo GWAS, utilizando como referência a figura adaptada de Uffelmann *et al.* 2021, que sintetiza graficamente esse processo.

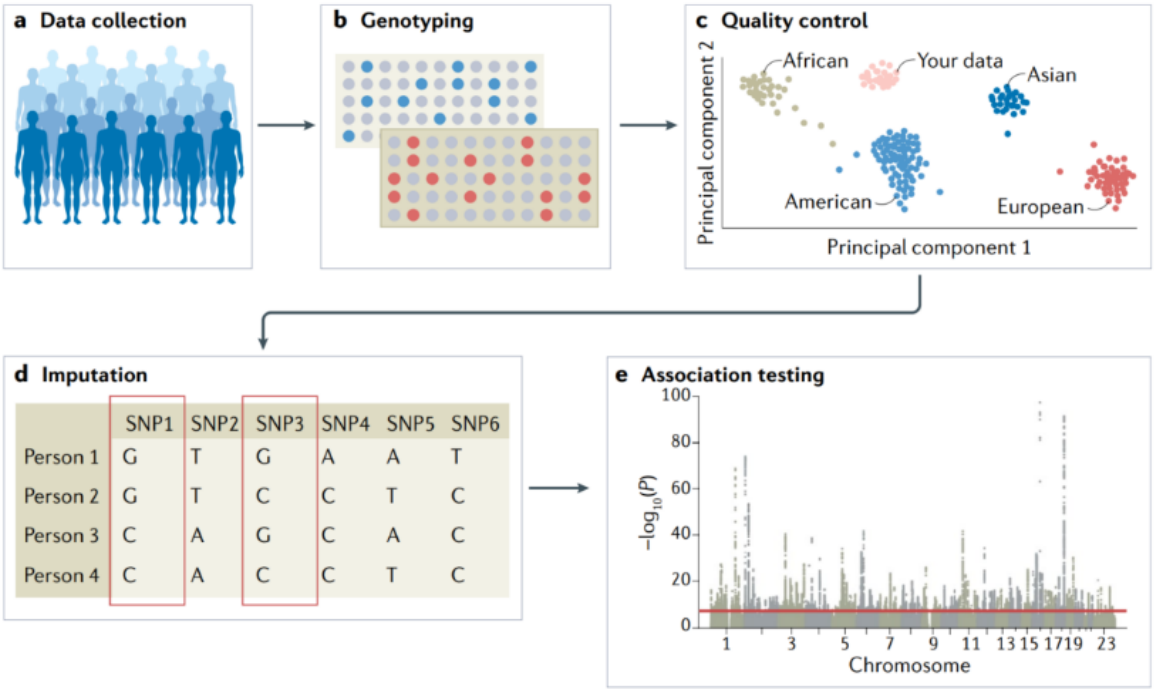

Figura 7 – Fluxograma esquemático de um estudo de associação genômica ampla (GWAS), incluindo coleta de dados, genotipagem, controle de qualidade, imputação genotípica e testes de associação. Adaptado de Uffelmann *et al.* 2021.

O primeiro passo para GWAS é a definição da população de estudo. É essencial contar com amostras de tamanho suficientemente grande para garantir poder estatístico adequado. As amostras podem ser obtidas de estudos prospectivos, coortes populacionais, biobancos ou bases de dados já existentes. Nesse processo, estratégias de recrutamento devem ser cuidadosamente planejadas para evitar viés de seleção, como o viés de colisor, e a estratificação populacional precisa ser considerada, especialmente quando diferentes etnias estão presentes na amostra.

A etapa seguinte envolve a genotipagem, que pode ser realizada por microarranjos, método mais utilizado devido ao seu custo-benefício, ou por técnicas mais abrangentes, como o

sequenciamento de exoma (WES) ou genoma completo (WGS). A escolha do método depende dos objetivos do estudo e da disponibilidade de recursos. O WGS tende a oferecer maior cobertura, permitindo identificar variantes raras além das comuns, mas seu custo ainda é um fator limitante em muitos contextos.

Após a genotipagem, os dados genéticos brutos passam por uma etapa essencial de controle de qualidade. Esse processo envolve uma série de verificações e filtros aplicados tanto aos SNPs quanto aos indivíduos da amostra. Entre os primeiros critérios aplicados está a filtragem baseada no MAF. SNPs com MAF muito baixo são frequentemente removidos porque oferecem pouca informação estatística, aumentam o risco de resultados espúrios e podem comprometer a estabilidade dos modelos. Outro critério importante é a verificação do equilíbrio de Hardy-Weinberg. SNPs que apresentam desvios significativos desse equilíbrio podem indicar problemas como erros de genotipagem, estrutura populacional não considerada ou pressões seletivas, e são geralmente excluídos da análise.

A etapa de imputação tem como objetivo estimar genótipos ausentes com base em dados de referência. No presente estudo, conforme descrito anteriormente, foi utilizada uma abordagem baseada na fração de recombinação e Cadeias de Markov, proposta por Stephens e Fisch 1998. Essa estratégia considera a dependência entre SNPs vizinhos para prever genótipos faltantes, utilizando a função de Haldane para estimar a fração de recombinação entre loci.

Além disso, SNPs com alta taxa de falha na genotipagem (ou seja, com muitos dados ausentes em diferentes indivíduos) são descartados, bem como indivíduos que apresentam proporção excessiva de genótipos ausentes, genótipos duplicados, ou outros indicadores de má qualidade. Um aspecto crucial do controle de qualidade é o tratamento da estratificação populacional, que ocorre quando a amostra analisada contém subgrupos de indivíduos com diferentes origens ancestrais. Esses subgrupos podem apresentar frequências alélicas distintas por razões não relacionadas ao fenótipo de interesse, gerando associações espúrias. Para mitigar esse problema, utiliza-se a análise de componentes principais. Os primeiros componentes principais (geralmente os dois ou três primeiros) capturam as maiores diferenças genéticas dentro da amostra, frequentemente refletindo variações de ancestralidade (como entre europeus, africanos e asiáticos). Esses componentes são então incluídos como covariáveis nos modelos de regressão do GWAS, funcionando como variáveis de ajuste que controlam os efeitos da estrutura populacional.

A etapa central do GWAS é o teste de associação estatística entre os SNPs e o desfecho de interesse. Uma possível abordagem se baseia em testar cada SNP individualmente quanto à sua associação com o fenótipo de interesse, utilizando modelos de regressão. Se o desfecho for contínuo (como altura ou glicemia), utiliza-se a regressão linear, se for binário (como presença ou ausência de uma doença), utiliza-se a regressão logística. Covariáveis como idade, sexo, e os componentes principais de ancestralidade são incluídas no modelo para controlar potenciais confundidores. Em alguns casos, modelos mistos com efeitos aleatórios também são aplicados para considerar o grau de parentesco entre os indivíduos. Nessa abordagem é utilizado o seguinte

modelo de regressão linear.

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 \text{sexo}_i + \beta_3 \text{idade}_i + \beta_4 \text{medDM}_i + \beta_5 \text{PC1}_i + \beta_6 \text{PC2}_i + \varepsilon_i, \qquad (2.1)$$

em que Y_i representa o fenótipo de interesse (neste caso, os níveis de glicose medidos em mg/dL); X_i corresponde ao genótipo do SNP avaliado, codificado numericamente como uma variável discreta que assume valores $\{0,1,2\}$, sendo 0 para homozigose do alelo de referência (aa), 1 para heterozigose (Aa) e 2 para homozigose do alelo alternativo (AA); sexo_i e idade_i referem-se, respectivamente, ao sexo e à idade do indivíduo; medDM_i indica o uso de medicamento para diabetes; PC1_i e PC2_i representam os dois primeiros componentes principais derivados da análise de ancestralidade genética, utilizados para ajustar a estratificação populacional; e ε_i é o termo de erro aleatório, assumido como normalmente distribuído, com média zero e variância constante. Para avaliar a associação entre o SNP e a variável resposta, testa-se a significância do coeficiente de regressão β_1. O teste de hipótese realizado é:

$$H_0 : \beta_1 = 0 \quad \text{vs} \quad H_1 : \beta_1 \neq 0$$

Esse teste verifica se há evidência estatística de que o SNP contribui para explicar a variação na variável fenotípica. Sob a hipótese nula, assume-se que o SNP não exerce efeito algum sobre o fenótipo. A estatística de teste utilizada é baseada na razão entre o valor estimado do coeficiente $\hat{\beta}_1$ e seu erro padrão:

$$t = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}$$

O p-valor pequeno indica forte evidência contra H_0, sugerindo que o SNP está significativamente associado ao fenótipo. Em síntese, a modelagem por regressão linear simples permite avaliar o efeito marginal de cada SNP sobre o fenótipo, mas não contempla possíveis interações entre marcadores (epistasia). No próximo capítulo, passaremos a considerar explicitamente essas interações entre SNPs por meio de métodos específicos.

CAPÍTULO

3

INTERAÇÃO ENTRE SNPS

Este capítulo discute a importância da identificação de interações entre SNPs no contexto da análise genômica, e apresenta os fundamentos teóricos e metodológicos relacionados ao funcionamento das Florestas de Interação, método aplicado para detectar interações entre SNPs.

3.1 Interações entre SNPs

A análise genômica tradicional frequentemente se limita à avaliação da associação entre um único SNP e a variável fenotípica de interesse. Embora essa abordagem possa identificar efeitos marginais importantes, ela negligencia a possibilidade de que combinações de SNPs interajam entre si, modulando conjuntamente o fenótipo analisado. Em muitas doenças, como o diabetes, os efeitos genéticos não são determinados apenas por variantes isoladas, mas emergem da interação entre múltiplos loci genéticos. Assim, para capturar de maneira mais precisa os mecanismos biológicos subjacentes, é necessário ir além da análise univariada e considerar a interação entre os próprios SNPs.

Uma abordagem estatística comum para modelar essas interações é a inclusão de termos de interação em modelos de regressão como citado em Liu *et al.* 2012. Supondo dois SNPs X_1 e X_2, e uma variável resposta contínua Y, o modelo com interação é representado por:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i} + \varepsilon_i \tag{3.1}$$

onde Y_i representa a variável fenotípica de interesse, neste caso, os níveis de glicose medidos em mg/dL. Os SNPs X_{1i} e X_{2i} são variáveis aleatórias que assumem valores discretos em $\{0,1,2\}$, correspondendo à codificação genotípica: 0 para homozigose do alelo de referência, 1 para heterozigose e 2 para homozigose do alelo alternativo, β_0 é o intercepto, β_1 e β_2 representam os efeitos principais dos SNPs X_1 e X_2, respectivamente, β_3 é o coeficiente de interação que quantifica o efeito combinado dos dois SNPs sobre o fenótipo, e ε_i é o erro aleatório, assumido com distribuição normal de média zero e variância constante. A identificação de uma interação

significativa entre SNPs é realizada por meio do teste de hipótese $H_0 : \beta_3 = 0$ contra $H_1 : \beta_3 \neq 0$, sendo o p-valor associado ao coeficiente de interação o principal indicador de sua relevância.

Estatisticamente, diz-se que há uma interação entre dois efeitos fixos X_1 e X_2 quando o efeito de X_1 sobre Y não é constante ao longo dos níveis de X_2, ou, equivalentemente, quando o efeito de X_2 sobre Y varia conforme os valores de X_1. No modelo da Equação (5.1), essa condição se traduz em $\beta_3 \neq 0$: o coeficiente de interação captura justamente a parcela do efeito de X_1 que depende do valor de X_2, e vice-versa.

Vale ressaltar que, nesta formulação, tanto X_1 quanto X_2 entram no modelo com **codificação aditiva** (dosagem 0/1/2), o que implica testar especificamente a componente *linear* $\times$ *linear* da interação. Nessa codificação, assume-se que o efeito do genótipo heterozigoto (Aa) está exatamente no ponto médio entre os dois homozigotos (aa e AA), de modo que o modelo traça uma relação linear perfeita entre os três valores de dosagem. O termo de interação β_3 testa, portanto, apenas se as retas de X_1 mudam de inclinação conforme os valores de X_2. Essa abordagem consome apenas 1 grau de liberdade, o que é vantajoso para a manutenção do poder estatístico, mas representa uma simplificação biológica relevante, pois ignora possíveis desvios do heterozigoto em relação à média esperada dos homozigotos, fenômeno conhecido como **dominância**.

A **codificação nominal** (ou categórica) constitui uma alternativa mais flexível: ao tratar cada SNP como uma variável com três categorias distintas por meio de variáveis *dummy*, o modelo não impõe nenhuma estrutura linear e calcula a média do fenótipo para cada uma das nove combinações possíveis de pares de genótipos (3×3), sendo capaz de capturar interações, incluindo efeitos de dominância e padrões não lineares. O custo dessa flexibilidade é o consumo de 4 graus de liberdade para o termo de interação, o que exige tamanhos amostrais consideravelmente maiores para manter o poder estatístico [Aschard *et al.* 2016, Cockerham 1954, Vitezica *et al.* 2017].

Biologicamente, a **dominância** ocorre quando um alelo domina o outro, de modo que o indivíduo heterozigoto (Aa) apresenta fenótipo semelhante ao homozigoto dominante (AA). Estatisticamente, o efeito de dominância corresponde ao desvio do heterozigoto em relação ao ponto médio esperado entre os dois homozigotos. Quando se utiliza a codificação aditiva (0/1/2), o modelo é cego para esse desvio e o incorpora no erro residual, o que pode resultar em perda de poder para detectar certas formas de epistasia [Cockerham 1954, Vitezica *et al.* 2017].

Apesar da utilidade teórica dessa modelagem, sua aplicação direta em estudos genômicos de larga escala é inviável. Considerando um banco com aproximadamente 24.929 SNPs, o número total de combinações de pares possíveis é dado por $\binom{24.929}{2} = 310.891.156$. Avaliar todas essas interações duas a duas em um modelo de regressão se torna não apenas computacionalmente intensivo, mas também estatisticamente ineficiente, devido ao alto risco de erros tipo I e à necessidade de correções severas para múltiplos testes. Adicionalmente, muitos SNPs não apresentam efeitos marginais significativos quando analisados isoladamente, mas podem ter

influência relevante no fenótipo quando interagem com outros loci. Esse comportamento desafia métodos tradicionais de aprendizado de máquina, como as florestas aleatórias convencionais, que tendem a privilegiar variáveis com fortes efeitos marginais e falham em detectar interações complexas. Diante dessas limitações, surgem abordagens mais sofisticadas, como as Florestas de Interação, especialmente desenvolvidas para identificar efeitos interativos em cenários de alta dimensionalidade, tema que será aprofundado a seguir.

3.2 Epistasia

No capítulo anterior, discutimos como a análise marginal de SNPs tende a capturar predominantemente efeitos aditivos. No entanto, essa abordagem ignora um aspecto fundamental da arquitetura genética de fenótipos complexos, o fato de que o efeito de um alelo pode depender do contexto genético em outro locus. Esse padrão de dependência é conhecido como epistasia, e sua incorporação é relevante tanto do ponto de vista biológico quanto estatístico.

Do ponto de vista biológico, a epistasia reflete a organização funcional dos genes em vias e redes regulatórias, nas quais múltiplos elementos interagem para modular a expressão e a função gênica. Já na perspectiva estatística, a inclusão de termos de interação entre SNPs pode recuperar parte da variância fenotípica não explicada por modelos puramente aditivos, contribuindo para a elucidação da chamada herdabilidade ausente [Manolio *et al.* 2009, Moore e Williams 2009].

A literatura costuma distinguir dois sentidos complementares do termo epistasia. A epistasia biológica refere-se a interações funcionais entre genes ou proteínas, como ocorre quando um enhancer regula a expressão de um gene alvo. Já a epistasia estatística diz respeito ao desvio, nos dados, em relação ao modelo aditivo: formalmente, ela se manifesta quando o coeficiente β_3 do termo de interação $X_1 \times X_2$ na Equação (5.1) é significativamente diferente de zero, indicando que o efeito conjunto dos dois SNPs não pode ser explicado apenas pela soma de seus efeitos individuais. Em outras palavras, a epistasia estatística existe sempre que o modelo puramente aditivo for insuficiente para descrever a relação entre os SNPs e o fenótipo, e a inclusão do termo de interação resultar em melhora significativa do ajuste. Embora as duas definições se informem mutuamente, elas não são equivalentes: a ausência de significância estatística para um termo de interação não exclui a existência de um mecanismo biológico subjacente, e o inverso também é verdadeiro [Cordell 2002, Phillips 2008].

Quando se fala em epistasia, é importante reconhecer que a interação entre dois SNPs não precisa se restringir à componente aditiva de ambos. O arcabouço de decomposição genética de Cockerham permite distinguir diferentes tipos de epistasia de acordo com os componentes envolvidos [Cockerham 1954, Vitezica *et al.* 2017]. A componente aditivo $\times$ aditivo corresponde à interação entre os efeitos lineares de dosagem dos dois loci e é justamente o que o coeficiente β_3 testa na codificação 0/1/2. A componente aditivo $\times$ dominante descreve uma situação em que

a presença de um alelo adicional no SNP_1 altera a forma como a dominância atua no SNP_2, ou seja, o efeito de heterozigose em um locus depende da dosagem do outro. Por fim, a componente dominante $\times$ dominant captura a interação entre os desvios de dominância de ambos os loci, correspondendo à situação em que o desvio do heterozigoto em relação ao ponto médio dos homozigotos em um SNP é amplificado ou revertido pelo desvio de dominância do outro SNP. Essas componentes de ordem superior são invisíveis ao modelo com codificação puramente aditiva e exigem a especificação explícita de contrastes de dominância, o que tem implicações diretas para a escolha do modelo e para a interpretação dos resultados.

A forma mais direta de testar interações entre SNPs é por meio da inclusão de termos de interação em modelos de regressão. Como discutido na Equação (5.1), o termo $X_1 \times X_2$ representa a componente epistática entre dois SNPs codificados numericamente. A escolha da codificação influencia diretamente o tipo de teste estatístico aplicado. A codificação aditiva, que utiliza dosagens 0/1/2, leva ao teste *linear* $\times$ *linear* (L$\times$L), com um grau de liberdade, sendo parcimoniosa e geralmente potente quando a relação dose-resposta é aproximadamente linear. A codificação nominal, que trata cada SNP como um fator com três níveis por meio de variáveis *dummy*, resulta em um teste geral 3×3, com quatro graus de liberdade para a interação, sendo mais flexível por acomodar dominância e padrões não lineares, embora apresente menor poder em amostras moderadas [Aschard *et al.* 2016, Cockerham 1954, Vitezica *et al.* 2017]. Uma abordagem intermediária consiste em incluir contrastes de dominância explícitos para um ou ambos os SNPs, permitindo testar componentes específicas da epistasia, como aditivo $\times$ dominante ou dominante $\times$ dominante, com consumo de graus de liberdade proporcional à complexidade do modelo especificado.

Para formalizar a estrutura dos dados, Blumenthal *et al.* 2021 propõem uma definição matemática clara de epistasia estatística. Consideremos S o conjunto de todos os SNPs, I o conjunto de indivíduos e F o conjunto de fenótipos. Para fenótipos quantitativos temos $F = \mathbb{R}$ (e.g., níveis de glicose), enquanto para fenótipos categóricos com C categorias temos $F = \{1, \ldots, C\}$.

Uma *instância de epistasia* é definida pelo par (G, y), em que

$$G = (g_{s,i}) \in \{0,1,2\}^{|S| \times |I|}, \qquad y = (y_i)_{i \in I} \in F^{|I|}.$$

A matriz G é a matriz genotípica de dimensão $|S| \times |I|$: cada **linha** corresponde a um SNP $s \in S$ e cada **coluna** a um indivíduo $i \in I$. O elemento $g_{s,i}$ representa a dosagem do alelo menor do SNP s no indivíduo i, usualmente codificada como

$$g_{s,i} \in \{0,1,2\},$$

com 0 para homozigose do alelo de referência, 1 para heterozigose e 2 para homozigose do alelo

alternativo. De forma explícita, podemos escrever

$$G = \begin{pmatrix} g_{s_1,i_1} & g_{s_1,i_2} & \cdots & g_{s_1,i_{|I|}} \\ g_{s_2,i_1} & g_{s_2,i_2} & \cdots & g_{s_2,i_{|I|}} \\ \vdots & \vdots & \ddots & \vdots \\ g_{s_{|S|},i_1} & g_{s_{|S|},i_2} & \cdots & g_{s_{|S|},i_{|I|}} \end{pmatrix}, \qquad y = \begin{pmatrix} y_{i_1} \\ y_{i_2} \\ \vdots \\ y_{i_{|I|}} \end{pmatrix}.$$

Para um subconjunto de SNPs $S' \subseteq S$ e um indivíduo i, o vetor

$$g_{S'}(i) = (g_{s,i})_{s \in S'} \in \{0,1,2\}^{|S'|}$$

representa o genótipo multilocus do indivíduo i restrito a S'. Um *modelo estatístico de epistasia* é então definido como

$$M = (f_{G,y}, \delta),$$

em que $f_{G,y} : \mathscr{P}(S) \to \mathbb{R}$ é uma função de escore que, dados os dados (G,y), atribui a cada subconjunto de SNPs $S' \subseteq S$ um valor real $f_{G,y}(S')$ que mede o grau de associação entre S' e o fenótipo, e $\delta \in \{\text{MIN}, \text{MAX}\}$ indica se esse escore deve ser minimizado ou maximizado.

Intuitivamente, $f_{G,y}$ quantifica quão "bom"é um conjunto de SNPs para explicar a variação fenotípica; o papel de δ é apenas dizer se conjuntos "melhores"têm escores menores ou maiores. Dois exemplos ilustram essa ideia:

- **Exemplo 1 – teste qui-quadrado**: define-se $f_{G,y}(S')$ como o p-valor de um teste χ^2 de associação entre os genótipos em S' e o fenótipo. Como conjuntos de SNPs mais interessantes produzem p-valores menores, escolhemos $\delta = \text{MIN}$.
- **Exemplo 2 – floresta de interação**: no caso das florestas de interação, o escore é o *Interaction Effect Measure* (EIM) associado ao conjunto de SNPs S'. Aqui, conjuntos com interações mais fortes recebem EIM maior, e, portanto, escolhemos $\delta = \text{MAX}$ e $f_{G,y}(S') = \text{EIM}(S')$.

A partir dessa estrutura, diferentes métodos de busca (exaustivos ou heurísticos) podem ser utilizados para encontrar subconjuntos S' com valores de $f_{G,y}(S')$ particularmente extremos, isto é, com evidência de epistasia estatística.

Casos bem documentados ajudam a ilustrar o conceito. Na coloração de olhos, a interação entre *HERC2* e *OCA2* modula a expressão gênica e determina a tonalidade azul ou castanho [Sturm *et al.* 2008]. Em doenças autoimunes, como espondilite anquilosante e psoríase, variantes em *ERAP1* elevam o risco apenas na presença de alelos específicos do complexo HLA [Evans *et al.* 2011, Strange *et al.* 2010]. Já na doença celíaca, o risco está associado à formação do heterodímero HLA-DQ2.5/DQ8, evidenciando que a combinação alélica é mais determinante que marcadores isolados [Megiorni e Pizzuti 2012].

Detectar epistasia em larga escala impõe desafios significativos. O número de pares cresce como $O(p^2)$, em que p denota o número de SNPs (marcadores genéticos) considerados na análise, o que torna a busca exaustiva inviável em bancos com centenas de milhares de SNPs. Além disso, a baixa frequência alélica de muitos marcadores gera esparsidade nas tabelas de contingência, reduzindo o poder estatístico dos testes [Niel *et al.* 2015, Blumenthal *et al.* 2021]. Para contornar essas limitações, consolidou-se uma abordagem em duas etapas: uma triagem inicial para reduzir o espaço de busca, baseada em critérios rápidos e computacionalmente leves, e uma validação paramétrica posterior, com codificação apropriada e controle de múltiplos testes.

Essa lógica de triagem e refinamento conduz naturalmente à próxima seção, dedicada às Interaction Forests, uma abordagem desenvolvida com o propósito de identificar interações epistáticas mesmo quando os efeitos marginais são pouco expressivos.

3.3 Florestas de Interação

As Interaction Forests, propostas por Hornung e Boulesteix 2022, constituem uma extensão das Florestas Aleatórias e foram desenvolvidas com o objetivo de identificar e explorar interações quantitativas e qualitativas entre variáveis. Essa metodologia se diferencia dos métodos tradicionais que realizam apenas divisões univariadas, os quais frequentemente negligenciam interações complexas entre preditores. Em contraste, as Interaction Forests adotam divisões bivariadas, possibilitando a identificação de efeitos combinados entre pares de variáveis. Essa abordagem mostra-se particularmente eficaz na análise de dados genômicos, pois é capaz de detectar interações relevantes mesmo quando os efeitos marginais individuais são fracos ou inexistentes. O algoritmo das Interaction Forests mantém a estrutura geral das Florestas Aleatórias, sendo baseado na construção de múltiplas árvores a partir de amostras bootstrap dos dados originais, o que garante variabilidade entre as árvores e reduz o risco de sobreajuste. No entanto, apresenta adaptações importantes em relação à estrutura das florestas aleatórias, além das tradicionais divisões baseadas em um único preditor, são permitidas divisões bivariadas, baseadas em pares de variáveis, para capturar diretamente os efeitos de interação. A escolha da melhor divisão em cada nó da árvore pode seguir critérios como a redução da variância (em variáveis contínuas) ou o índice de Gini (para desfechos categóricos), entre outros. Dado que o número de pares possíveis de SNPs cresce quadraticamente com o número de variáveis $O(p^2)$, o algoritmo também implementa estratégias de pré-seleção de variáveis, a fim de reduzir a complexidade computacional e focar em interações mais promissoras. Um dos principais diferenciais das Interaction Forests está na possibilidade de realizar dois tipos de divisões nos nós das árvores, univariadas e bivariadas como apresentado a seguir.

1. **Divisões Univariadas**
   As divisões univariadas seguem o formato tradicional das árvores de decisão:

$$x_j < p_u \quad \text{vs.} \quad x_j \geq p_u$$

onde x_j é uma variável e p_u é um ponto de corte escolhido com base nos valores únicos da variável.

2. **Divisões Bivariadas**
   As divisões bivariadas buscam capturar interações entre duas variáveis (x_i, x_j), sendo classificadas em:

   - **Interações Quantitativas**: Uma interação é classificada como quantitativa quando o efeito da variável x_j sobre a variável resposta Yij depende da magnitude de outra variável x_i, sem alteração na direção desse efeito. Em outras palavras, o impacto de x_j pode variar em intensidade conforme os valores de x_i, porém o sinal da associação (positivo ou negativo) permanece constante.

     As divisões ocorrem com base nas seguintes regiões:

     – Região 1: x_i pequeno e x_j pequeno

     $$\text{Região: } \{x_i < p_i\} \cap \{x_j < p_j\}$$

     $$\text{Complementar: } \{x_i \geq p_i\} \cup \{x_j \geq p_j\}$$

     – Região 2: x_i pequeno e x_j grande

     $$\text{Região: } \{x_i < p_i\} \cap \{x_j \geq p_j\}$$

     $$\text{Complementar: } \{x_i \geq p_i\} \cup \{x_j < p_j\}$$

     – Região 3: x_i grande e x_j pequeno

     $$\text{Região: } \{x_i \geq p_i\} \cap \{x_j < p_j\}$$

     $$\text{Complementar: } \{x_i < p_i\} \cup \{x_j \geq p_j\}$$

     – Região 4: x_i grande e x_j grande

     $$\text{Região: } \{x_i \geq p_i\} \cap \{x_j \geq p_j\}$$

     $$\text{Complementar: } \{x_i < p_i\} \cup \{x_j < p_j\}$$

   - **Interações Qualitativas**: Interações qualitativas ocorrem quando o efeito da variável x_j sobre a resposta depende dos valores de x_i, de forma que a direção desse efeito se inverte conforme a magnitude de x_i. Nessas situações, x_j pode estar positivamente associado à resposta em determinados níveis de x_i, e negativamente associado em outros. Esse tipo de interação reflete uma dependência no sinal do efeito e é modelado no algoritmo Interaction Forest por divisões que contrastam regiões diagonais do espaço definido pelas variáveis.

Divisão:

$$\big(\{x_i < p(i)\} \cap \{x_j < p(j)\}\big) \cup \big(\{x_i \geq p(i)\} \cap \{x_j \geq p(j)\}\big)$$

vs.

$$\big(\{x_i < p(i)\} \cap \{x_j \geq p(j)\}\big) \cup \big(\{x_i \geq p(i)\} \cap \{x_j < p(j)\}\big)$$

Divisão Complementar:

$$\big[\big(\{x_i < p(i)\} \cap \{x_j < p(j)\}\big) \cup \big(\{x_i \geq p(i)\} \cap \{x_j \geq p(j)\}\big)\big]^c$$

vs.

$$\big[\big(\{x_i < p(i)\} \cap \{x_j \geq p(j)\}\big) \cup \big(\{x_i \geq p(i)\} \cap \{x_j < p(j)\}\big)\big]^c$$

Um exemplo de interação quantitativa ocorre quando o efeito de uma variável, como um medicamento (x_i), varia em intensidade dependendo da idade do paciente (x_j), mantendo, no entanto, a mesma direção (por exemplo, sempre positiva). Já as interações qualitativas se caracterizam por uma inversão no sinal do efeito: por exemplo, o SNP_1 pode estar associado a um maior risco de diabetes quando o SNP_2 está no genótipo homozigoto variante (2), mas pode exercer efeito protetor quando o SNP_2 está no genótipo de referência (0). Nesse caso, a direção do efeito de SNP_1 sobre o desfecho depende do estado de SNP_2, configurando uma dependência qualitativa. As divisões descritas são ilustradas na Figura 8.

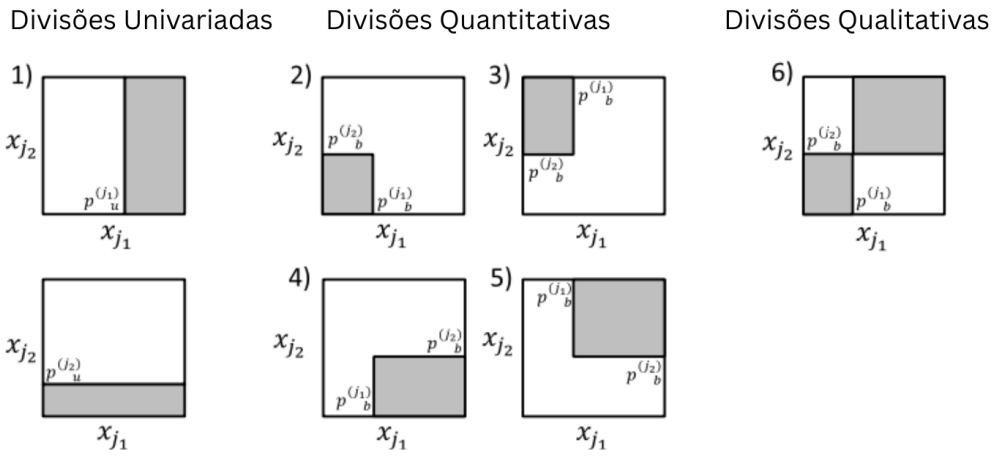

Figura 8 – Representação dos tipos de divisões utilizadas nas Interaction Forests: divisões univariadas (à esquerda), interações quantitativas (ao centro) e interações qualitativas (à direita).

Fonte: adaptado de Hornung e Boulesteix 2022.

A construção de árvores nas Interaction Forests segue um processo iterativo, em que cada nó pode ser dividido univariadamente ou bivariadamente. O objetivo central é selecionar a divisão que maximize a pureza dos nós filhos, ou seja, aquela que melhor separa os indivíduos com base na variável resposta. Esse mecanismo assegura que as divisões mais informativas sejam priorizadas durante o crescimento da árvore. O processo de seleção das divisões se inicia com a

geração de um conjunto de divisões candidatas. Para isso, sorteia-se aleatoriamente um número fixo de pares de variáveis, denotado por n_{pairs}. Esse parâmetro é análogo ao *mtry* utilizado nas *Random Forests*, que representa o número de variáveis a serem consideradas em cada nó. Em vez de considerar todas as variáveis, essa amostragem torna o processo computacionalmente viável, especialmente em cenários de alta dimensionalidade. Para cada par (x_i, x_j) selecionado, são geradas duas divisões univariadas (uma para cada variável) por meio do sorteio de pontos de corte entre os valores únicos ordenados, utilizando-se os pontos médios entre observações adjacentes. Em seguida, são construídas cinco divisões bivariadas distintas: quatro correspondentes a diferentes formas de interações quantitativas e uma qualitativa, todas utilizando o mesmo ponto de corte bivariado $(p(i), p(j))$. O valor de $p(i)$ é obtido como a média entre dois valores únicos sorteados da variável x_i. Já $p(j)$ é sorteado de forma a garantir que os quatro quadrantes definidos pelas divisões contenham pelo menos uma observação, assegurando validade estatística para todas as cinco divisões.

Após a geração do conjunto de splits candidatos, cada divisão é avaliada quanto ao critério de divisão apropriado, para o caso da nossa aplicação, será utilizado o índice de Gini. A divisão que apresentar o melhor desempenho segundo esse critério é selecionada para dividir o nó. Esse processo é repetido recursivamente até que os critérios de parada sejam atingidos, como número mínimo de observações por nó ou profundidade máxima da árvore. A escolha dos pares de variáveis a serem avaliados como candidatos à divisão é crucial, visto que o número total de combinações cresce quadraticamente com o número de variáveis p. Para contornar a inviabilidade computacional de testar todas as combinações, aplica-se uma estratégia de pré-seleção baseada na metodologia descrita em Hornung e Boulesteix 2022. Para conjuntos com $p \leq 100$, todos os pares possíveis são considerados. No presente estudo, como o número de SNPs analisados foi inferior a 100, foi possível testar todas as combinações possíveis de pares de variáveis, sem a necessidade de pré-seleção. No entanto, de forma geral, quando $100 < p \leq 447$, testa-se cada par de variáveis por meio de modelos lineares para detectar indícios de interação, e os 5000 pares com menores valores de p-valor são selecionados. Se $447 < p \leq 30000$, adota-se o método BOLT-SSI, que permite uma seleção eficiente de interações relevantes em contextos de alta dimensionalidade. Como o BOLT-SSI tem limitação de selecionar no máximo p pares, ele é aplicado repetidamente a subconjuntos aleatórios de variáveis até que se obtenha um total de 5000 pares promissores. Por fim, para $p > 30000$, executa-se uma regressão linear univariada para cada variável, retendo as 30000 mais associadas à variável resposta.

É importante observar que as divisões bivariadas implementadas nas Interaction Forests assumem, em sua formulação padrão, que os efeitos de interação se manifestam por meio de pontos de corte lineares no espaço dos preditores, o que corresponde, na prática, à detecção de interações do tipo *linear* $\times$ *linear* quando os preditores são variáveis de dosagem genotípica (0/1/2). Interações que envolvam componentes de dominância, como *aditivo* $\times$ *dominante* ou *dominante* $\times$ *dominante*, não são capturadas diretamente por esse mecanismo, pois exigiriam que as variáveis X_1 e X_2 fossem recodificadas para contrastes de dominância antes de serem

inseridas no algoritmo.

3.3.1 Medida de Importância de Efeito

A avaliação da importância das variáveis é uma etapa crucial em modelos baseados em árvores, especialmente em contextos de alta dimensionalidade, como em estudos genômicos com SNPs. As chamadas Variable Importance Measures (VIM) têm como objetivo quantificar a contribuição de cada variável para o desempenho preditivo do modelo. No algoritmo Random Forests, uma das medidas mais utilizadas é a permutation importance, proposta por Breiman 2001. Essa abordagem consiste em calcular a acurácia preditiva de uma árvore utilizando os dados fora da amostra Out-of-Bag (OOB) e, em seguida, embaralhar os valores de uma variável específica nos dados OOB, rompendo sua associação com a variável resposta. A diferença entre a acurácia original e a acurácia com a variável permutada reflete sua importância para o modelo. No entanto, esse procedimento pode ser sensível à presença de valores ausentes e exigir procedimentos adicionais. Para contornar esse problema, Hapfelmeier *et al.* 2014 propuseram uma alternativa denominada Proportional Randomization (PropRandom). Esta técnica simula a ausência de uma variável no momento da avaliação sem a necessidade de manipular diretamente seus valores. O funcionamento se da, da seguinte forma, ao calcular a acurácia preditiva com dados OOB, sempre que a árvore encontra um nó em que a divisão foi feita com base na variável de interesse, o algoritmo desativa essa divisão. Em vez de usar o valor da variável para decidir o caminho a seguir, o indivíduo é direcionado aleatoriamente para um dos ramos filhos. Essa aleatoriedade não é uniforme, a probabilidade de ir para cada ramo é proporcional ao número de observações que originalmente seguiram para ele durante a construção da árvore. Com isso, a dependência da divisão em relação à variável é neutralizada, respeitando a estrutura original da árvore. Com base nesse mecanismo, o Effect Importance Measure (EIM), adotado no algoritmo Interaction Forests, permite calcular separadamente a importância dos efeitos univariados, interações quantitativas e interações qualitativas. O procedimento ocorre da seguinte forma:

1. Inicialmente, calcula-se a acurácia preditiva de cada árvore utilizando suas observações OOB.

2. Em seguida, aplica-se o método *PropRandom* para simular a ausência de um determinado efeito (univariado ou bivariado) e recalcula-se a acurácia da árvore.

3. A diferença entre a acurácia original e a obtida com o efeito desativado indica o quanto esse efeito contribui para a predição.

4. Essa diferença é calculada para cada árvore da floresta e, ao final, as médias dessas diferenças são obtidas, gerando três vetores distintos de EIM: um para efeitos univariados, um para interações quantitativas e outro para interações qualitativas.

Para o caso das interações quantitativas, o algoritmo considera quatro tipos distintos de divisão e calcula separadamente o EIM para cada um deles. Como cada par de variáveis pode exibir, no máximo, um tipo de interação quantitativa, os valores são ajustados de modo a refletir apenas interações genuínas, excluindo efeitos marginais. O maior valor ajustado entre os quatro tipos é então considerado como o EIM final da interação quantitativa. Para exemplificar como os valores de EIM podem ser interpretados na prática, utilizaremos um exemplo apresentado no Material Suplementar 1 do artigo de Hornung e Boulesteix 2022. Nesse estudo, os autores aplicaram o algoritmo Interaction Forests ao conjunto de dados zoo, disponível na plataforma OpenML. Esse banco contém informações sobre 101 espécies biológicas diferentes, descritas por 16 atributos, sendo 15 variáveis binárias (como presença de pelos, ovos, veneno etc.) e uma variável métrica (número de pernas). O desfecho analisado é binário, distinguindo entre mamífero e outros. O objetivo da análise era identificar quais variáveis e interações contribuíam para classificar corretamente os mamíferos com base nos atributos fornecidos. Ao construir a floresta com esse conjunto de dados e calcular a medida de importância, os autores observaram que a variável com maior valor de EIM univariado foi `milk`, que indica se a espécie produz leite. Seu valor de EIM foi aproximadamente 0,1748, o mais elevado entre todas as variáveis analisadas. Esse achado é coerente com a biologia das espécies, pois todas as 41 classificadas como mamíferos produzem leite, enquanto nenhuma das outras 60 o faz. Assim, essa única variável permitiria classificar corretamente todas as espécies da amostra. A imagem com a distribuição dos valores de EIM para os efeitos univariados, interações quantitativas e qualitativas é apresentada a seguir na Figura 9.

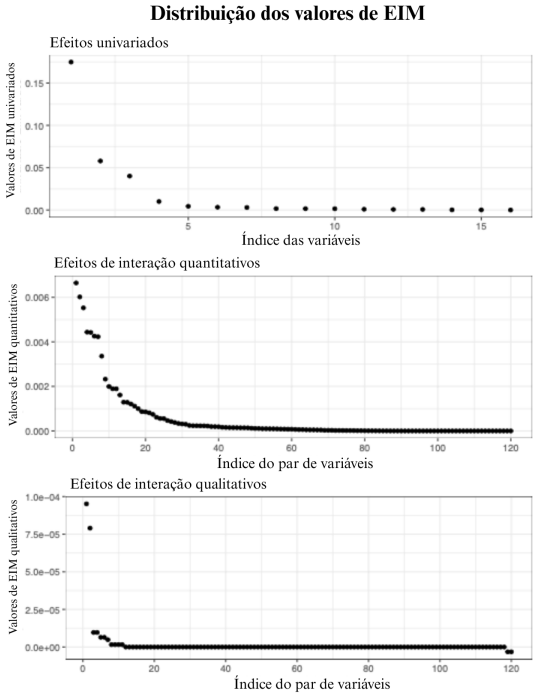

Figura 9 – Representação dos tipos de divisões utilizadas nas Interaction Forests: divisões univariadas (à esquerda), interações quantitativas (ao centro) e interações qualitativas (à direita).

Fonte: adaptado de Hornung e Boulesteix 2022.

Apesar desse resultado quase perfeito, os autores destacam que o valor de EIM para milk não atinge um valor elevado. Isso demonstra que, mesmo quando há uma associação forte, os valores de EIM não são necessariamente elevados. Essa característica decorre da natureza da métrica, o EIM é baseado na diferença de acurácia preditiva estimada por simulação, e pode ser influenciado por diversos fatores, como o número de árvores, a estrutura das árvores, a frequência de uso das variáveis nos nós e a distribuição dos dados. Por isso, os valores de EIM devem ser interpretados com cautela e não utilizados como único critério de decisão.

CAPÍTULO

4

ESTUDO DE SIMULAÇÃO

Este capítulo apresenta as análises iniciais conduzidas com o algoritmo, com o objetivo de avaliar sua eficácia na identificação de interações entre SNPs. As investigações envolveram um estudo de simulação em que interações genéticas específicas foram inseridas artificialmente na variável resposta, repetido em 100 réplicas independentes por cenário. Esse delineamento permite avaliar não apenas se o método é capaz de recuperar o par verdadeiro, mas também com que frequência o faz e quantos falsos-positivos são produzidos no processo.

4.1 Descrição do Banco de Dados

Os dados utilizados neste estudo são provenientes do projeto temático "Estilo de vida, marcadores bioquímicos e genéticos como fatores de risco cardiometabólico: inquérito de saúde na cidade de São Paulo" (processo FAPESP 17/05125-7), baseado no estudo populacional "Inquérito de Saúde do Município de São Paulo" (ISA-Capital) e em seu complemento "Inquérito de Saúde de São Paulo com foco em Nutrição" (ISA-Nutrição). O ISA-Capital é um inquérito transversal em ciclos quinquenais que investiga condições de vida, situação de saúde e acesso a serviços na população paulistana, enquanto o ISA-Nutrição aprofunda aspectos relacionados à alimentação e estilo de vida.

No ciclo de 2015, 4043 indivíduos foram entrevistados, dos quais 901 participaram da coleta de dados biológicos. Para a análise genética, amostras de DNA de 846 participantes foram genotipadas utilizando o *Axiom Precision Medicine Research Array – Axiom 2.0 Assay* (Thermo Fisher Scientific), cobrindo 902.527 SNPs. Após controle de qualidade (*call rate* > 0,98) e seleção de 720 indivíduos independentes com base na matriz de parentesco genômico, aplicaram-se filtros de MAF mínima de 5%, p-valor do teste de Hardy–Weinberg inferior a 10^{-6} e exclusão de SNPs e indivíduos com mais de 5% de dados ausentes. Em seguida, a remoção de valores ausentes nas covariáveis sexo e idade resultou em um total de 698 participantes incluídos nas análises.

Os SNPs foram organizados por cromossomo (1 a 22), e valores ausentes em alguns marcadores foram imputados por meio de Cadeias de Markov, com base na fração de recombinação, conforme descrito no Capítulo 2 e em Stephens e Fisch 1998. Para reduzir a dimensionalidade e selecionar SNPs mais informativos para os níveis de glicose (variável resposta contínua), ajustou-se um modelo de regressão linear múltipla, considerando a glicemia transformada pelo logaritmo natural. O modelo incluiu como covariáveis fixas sexo, idade, uso de medicamento para diabetes (*medDM_bioq*) e duas componentes principais (*PC1* e *PC2*), derivadas de uma análise de componentes principais aplicada aos dados e utilizadas para ajustar a estratificação genética Pereira *et al.* 2024. Em linhas gerais, *PC1* captura variações relacionadas à ancestria Europeia e *PC2* à ancestria Africana, contribuindo para reduzir associações espúrias entre SNPs e fenótipo. As covariáveis de natureza antropométrica como peso e altura não foram incorporadas ao modelo neste momento, ainda que o IMC seja um conhecido modulador da glicemia; trata-se de uma limitação reconhecida e discutida nas considerações finais (Capítulo **??**).

No cromossomo 1, cada um dos 24.930 SNPs foi inserido individualmente no modelo, mantendo-se constantes as covariáveis clínicas e populacionais. Dessa forma, estimou-se o efeito marginal de cada marcador sobre os níveis de glicose e obtiveram-se os respectivos p-valores. Os 10 SNPs com menores p-valores foram selecionados para análises posteriores. Ressalta-se que, caso a variável resposta fosse binária (por exemplo, presença ou ausência de diabetes), o modelo estatístico adequado seria uma regressão logística, permitindo modelar a probabilidade de ocorrência do evento em função dos preditores genéticos e clínicos.

4.2 Análise de associação por regressão linear simples

Neste estudo, o objetivo inicial é selecionar os marcadores genéticos mais promissores com base na sua associação com um desfecho quantitativo contínuo, os níveis de glicose sanguínea. Para isso, aplicamos um modelo de regressão linear simples com múltiplas covariáveis ajustadas, de modo a obter o p-valor correspondente a cada SNP e, assim, priorizar aqueles com maior evidência estatística de associação.

Seja Y_i o nível de glicose observado para o i-ésimo indivíduo da amostra de tamanho n. Consideramos o seguinte modelo:

$$\log(Y_i) = \beta_0 + \beta_1 \cdot \text{sexo}_i + \beta_2 \cdot \text{idade}_i + \beta_3 \cdot \text{medDM\_bioq}_i + \beta_4 \cdot \text{PC1}_i + \beta_5 \cdot \text{PC2}_i + \beta_j \cdot \text{SNP}_{ij} + \varepsilon_i \tag{4.1}$$

em que:

- sexo_i, idade_i, medDM_bioq_i, PC1_i e PC2_i são covariáveis fixas, incluídas em todos os modelos;
- SNP_{ij} representa o j-ésimo marcador genético do indivíduo i;

- $\beta_0, \beta_1, \ldots, \beta_5, \beta_j$ são os coeficientes do modelo;
- $\varepsilon_i \sim N(0, \sigma^2)$ é o termo de erro aleatório, assumido com distribuição normal e variância constante.

A transformação logarítmica da variável resposta foi aplicada para atender aos pressupostos do modelo, especialmente a normalidade dos resíduos. Para cada SNP do cromossomo 1 (totalizando 24.930 marcadores), foi ajustado individualmente um modelo conforme descrito acima. Isso significa que, para cada iteração, testou-se a significância do coeficiente β_j associado ao SNP avaliado, mantendo constantes as demais covariáveis. A inferência é feita por meio do teste de hipótese:

$$H_0: \ \beta_j = 0 \quad \text{(nenhuma associação entre o SNP e o nível de glicose)}$$
$$H_1: \ \beta_j \neq 0 \quad \text{(existe associação entre o SNP e o nível de glicose).}$$

O p-valor obtido para cada teste quantifica a evidência contra a hipótese nula. SNPs com valores abaixo de um nível de significância ajustado são considerados estatisticamente associados ao desfecho. A Figura 10 apresenta a distribuição da variável glicose em sua forma original e após a transformação logarítmica. Observa-se que a transformação melhora a simetria da distribuição, aproximando-se de uma distribuição normal.

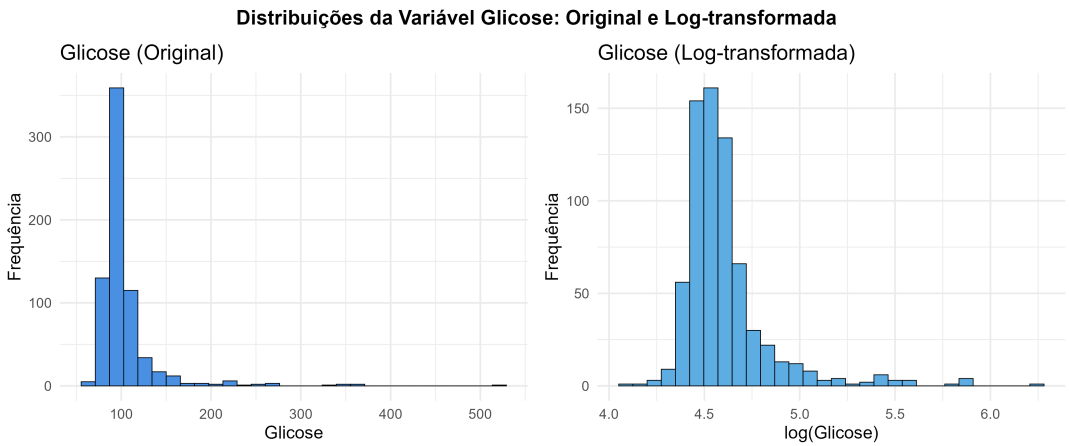

Figura 10 – Distribuição da variável glicose antes (esquerda) e após (direita) a transformação logarítmica. A transformação busca estabilizar a variância e aproximar a distribuição da normalidade.

4.2.1 Diagnóstico do modelo basal

Antes de incorporar os SNPs como preditores, é importante avaliar a adequação do chamado *modelo basal*, isto é, a regressão de $\log(\text{glicose})$ sobre apenas as covariáveis fixas (sexo, idade, *PC1*, *PC2* e *medDM_bioq*), sem qualquer SNP. Esse exame permite verificar em que medida os pressupostos clássicos da regressão linear são razoáveis para os dados antes da inclusão de marcadores genéticos.

A Figura 11 apresenta quatro painéis diagnósticos do modelo basal. O QQ-plot dos resíduos padronizados revela desvio claro da normalidade na cauda direita, com observações cujos quantis amostrais excedem em muito os quantis teóricos. O gráfico de resíduos versus valores ajustados exibe duas concentrações de pontos, em torno de $\log(\text{glicose}) \approx 4{,}5$ e $\log(\text{glicose}) \approx 4{,}9$, correspondentes aos grupos de indivíduos com e sem uso de medicamento para diabetes. O histograma dos resíduos é assimétrico à direita, com cauda longa, e o gráfico escala-localização sugere aumento da dispersão nos valores ajustados mais altos, característica de heterocedasticidade.

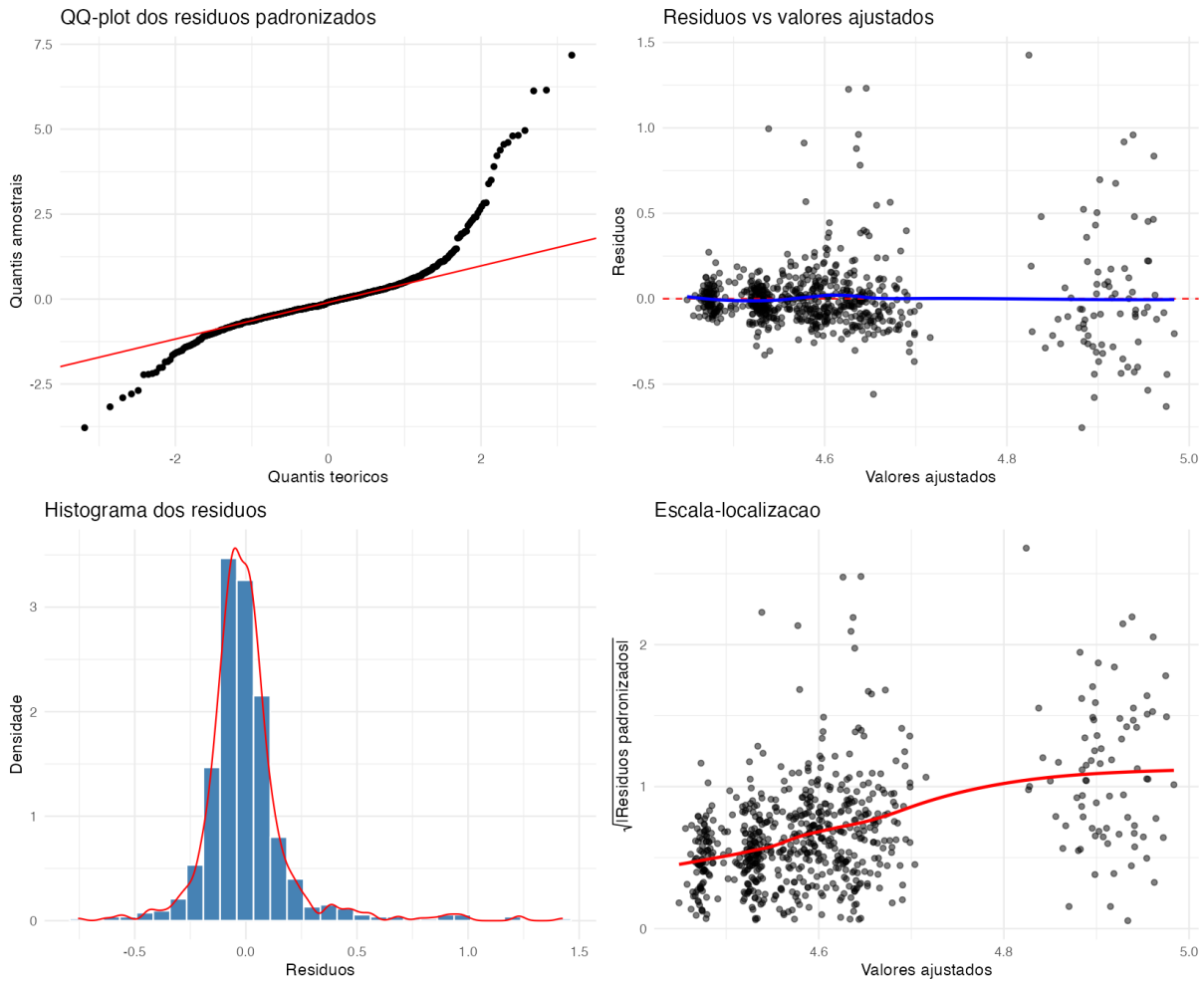

Figura 11 – Diagnóstico do modelo basal $\log(\text{glicose}) \sim \text{sexo} + \text{idade} + \text{PC1} + \text{PC2} + \text{medDM\_bioq}$, ajustado em $n = 698$ indivíduos. Painéis: QQ-plot dos resíduos padronizados (superior esquerdo), resíduos versus valores ajustados (superior direito), histograma dos resíduos com curva de densidade (inferior esquerdo) e gráfico escala-localização (inferior direito).

Esses padrões são corroborados por testes formais. Os testes de Shapiro-Wilk ($W = 0{,}79$; $p < 2 \times 10^{-16}$) e Anderson-Darling ($A = 33{,}25$; $p < 2 \times 10^{-16}$) rejeitam a hipótese de normalidade dos resíduos, e o teste de Breusch-Pagan ($BP = 53{,}23$; $p = 3{,}0 \times 10^{-10}$) rejeita a hipótese de homocedasticidade. O modelo basal explica $R^2_{\text{aj}} = 0{,}271$ da variabilidade de $\log(\text{glicose})$, com efeito amplamente dominante de *medDM_bioq* ($p < 2 \times 10^{-16}$), seguido por

idade ($p = 1{,}5 \times 10^{-11}$) e sexo ($p = 4 \times 10^{-4}$); as componentes principais *PC1* e *PC2* não foram individualmente significativas neste subconjunto.

A violação dos pressupostos não invalida as análises subsequentes, mas exige cautela. O uso de medicamento para diabetes é a principal fonte de heterogeneidade nos resíduos: indivíduos sob tratamento apresentam glicemia mais elevada e mais variável, o que produz a mistura visível no painel de resíduos versus valores ajustados. Nos modelos com SNP e termos de interação, mantemos a transformação logarítmica e o ajuste por *medDM_bioq*, e interpretamos os p-valores das componentes genéticas como medidas de evidência relativa, em vez de probabilidades estritamente calibradas sob normalidade.

4.2.2 Identificação de SNPs mais associados

A partir da aplicação da regressão linear simples, foram identificados os SNPs com os menores p-valores no cromossomo 1, os quais apresentaram associação estatisticamente significativa com os níveis de glicose transformados. Esses marcadores foram selecionados como base para o estudo de simulação descrito na próxima seção. A distribuição dos p-valores obtidos é visualizada por meio de um gráfico do tipo Manhattan, apresentado na Figura 12.

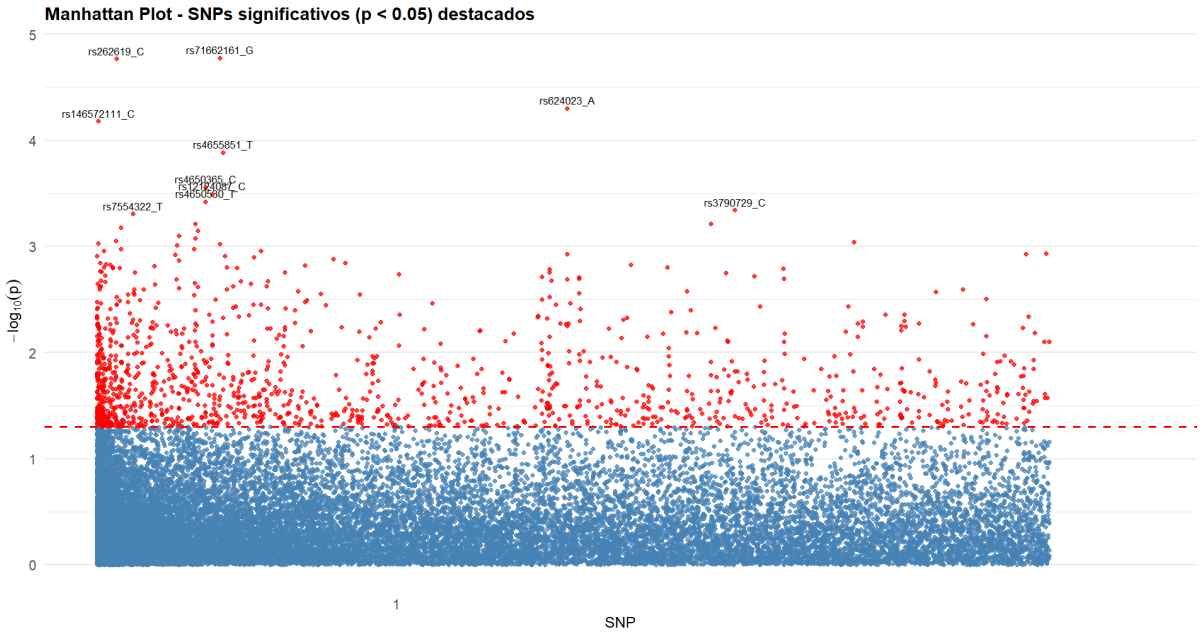

Figura 12 – Gráfico Manhattan representando os $p-$valores obtidos na regressão linear simples entre os SNPs do cromossomo 1 e os níveis transformados de glicose. Cada ponto representa um SNP, e os valores de $-\log_{10}(p)$ indicam a força da associação estatística com a variável resposta.

Os 10 SNPs com os menores p-valores utilizados nas análises seguintes foram listados na Tabela 1.

SNP	pvalue
rs71662161_G	1.666622×10^{-5}
rs262619_C	1.700867×10^{-5}
rs624023_A	4.977104×10^{-5}
rs146572111_C	6.585245×10^{-5}
rs4655851_T	1.300204×10^{-4}
rs4650365_C	2.755933×10^{-4}
rs12124087_C	3.200288×10^{-4}
rs4650530_T	3.786899×10^{-4}
rs3841330_AT	4.102078×10^{-4}
rs3790729_C	4.548439×10^{-4}

Tabela 1 – SNPs com maior nível de associação ao nível de glicose

4.3 Estudo de simulação

Com o objetivo de avaliar a capacidade das Florestas de Interação (IF) em detectar diferentes tipos de interação entre SNPs, conduzimos um estudo de simulação a partir do subconjunto de dez SNPs com menores *p*-valores marginais identificados na Tabela 1. Essa escolha visa controle e reprodutibilidade; nas aplicações com dados reais (Capítulo 5), empregamos um conjunto maior, com 400 SNPs na etapa de pré-seleção. Em todas as simulações, as covariáveis sexo, idade, *PC1*, *PC2* e *medDM_bioq* foram incluídas como efeitos principais, com coeficientes estimados a partir do modelo basal descrito na Subseção 4.2.1.

4.3.1 Desenho do estudo

Quatro cenários foram organizados em duas vertentes: (i) interações quantitativas, caracterizadas por mudança de magnitude do efeito sem inversão de sinal, com versões de *betas altos* e *betas baixos*; e (ii) interações qualitativas, caracterizadas por inversão de sinal do efeito, igualmente com versões de *betas altos* e *betas baixos*. Nos cenários quantitativos, são introduzidas duas interações simultâneas, $\text{SNP}_1 \times \text{SNP}_2$ e $\text{SNP}_1 \times \text{SNP}_4$; nos cenários qualitativos, apenas $\text{SNP}_1 \times \text{SNP}_2$. Os coeficientes utilizados em cada cenário estão consolidados na Tabela 2.

Tabela 2 – Coeficientes utilizados nos quatro cenários de simulação. Os efeitos principais e covariáveis (sexo, idade, *PC1*, *PC2*, *medDM_bioq*) são comuns a todos os cenários e seguem os valores estimados no modelo basal. γ_{12} e γ_{14} denotam, respectivamente, os coeficientes das interações $\text{SNP}_1 \times \text{SNP}_2$ e $\text{SNP}_1 \times \text{SNP}_4$.

Cenário	β_{SNP_1}	β_{SNP_2}	β_{SNP_4}	γ_{12}	γ_{14}	σ
Quantitativo — betas altos	0,0415	0,0384	0,0503	2,000	3,000	2,5
Quantitativo — betas baixos	0,0415	0,0384	0,0503	0,700	1,200	2,5
Qualitativo — betas altos	0,600	0,600	0,000	−1,200	0,000	2,5
Qualitativo — betas baixos	0,400	0,400	0,000	−0,800	0,000	2,5

Para cada cenário, o procedimento é repetido em $N_{\text{rep}} = 100$ réplicas independentes. Em

cada réplica, mantemos fixos os genótipos observados e os coeficientes do cenário, e geramos um novo vetor de erros $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, produzindo uma realização nova de $\log(Y)$. Para essa realização, ajusta-se uma Floresta de Interação a partir do banco contendo os dez SNPs e as cinco covariáveis (totalizando $\binom{15}{2} = 105$ pares possíveis), com num.trees $= 500$. A escolha desse número de árvores, inferior ao padrão de 5000 do pacote `diversityForest`, é uma decisão de viabilidade computacional.

A partir do ajuste, extraem-se os rankings de EIM (componentes quantitativa e qualitativa) e selecionam-se os 20 pares com maiores valores em cada componente para validação por regressão linear com termo de interação. Em cada réplica, registram-se três conjuntos de informações: (a) a posição do(s) par(es) verdadeiro(s) no ranking de EIM da componente correspondente; (b) o p-valor do termo de interação dos pares verdadeiros, quando estes aparecem entre os 20 priorizados; e (c) o número de pares *não verdadeiros* no top-20 cujo p-valor de interação é inferior a 0,05, contabilizados como falsos-positivos.

A apresentação dos resultados é organizada em torno de três métricas: a posição do par verdadeiro no ranking de EIM (Figura 13), o $-\log_{10}(p_{\text{int}})$ do par verdadeiro (Figura 14) e o número de falsos-positivos no top-20 (Figura 15). A síntese quantitativa do desempenho do método em recuperar o par verdadeiro está consolidada na Tabela 3, na forma da taxa de recuperação *Recall@K*, definida como a proporção de réplicas em que o par verdadeiro aparece até a posição K do ranking de EIM.

4.3.2 Resultados

A Figura 13 apresenta a posição do(s) par(es) verdadeiro(s) no ranking de EIM através das 100 réplicas de cada cenário. Cada ponto representa uma simulação, e as linhas horizontais sobre cada coluna indicam a média (preta, sólida) e a mediana (tracejada, na cor do par). Como o ranking começa em 1 (topo), o eixo vertical foi invertido, de modo que pontos no alto do gráfico correspondem a melhor recuperação.

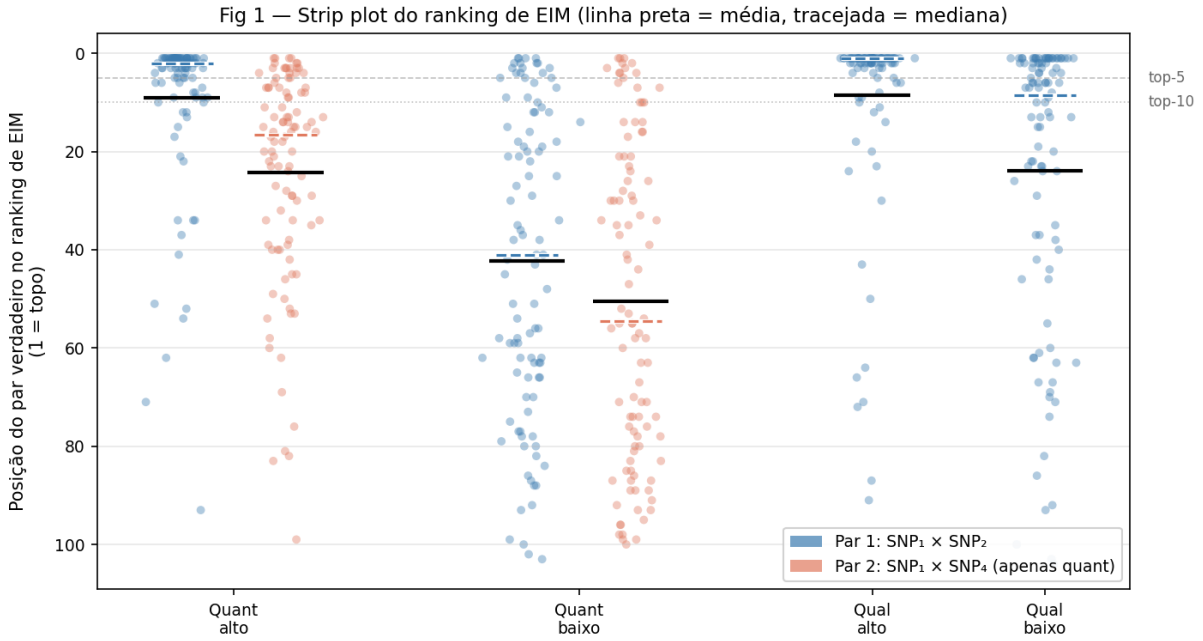

Figura 13 – Posição do par verdadeiro no ranking de EIM em 100 réplicas, por cenário. Cada ponto é uma simulação; linha preta sólida = média, linha colorida tracejada = mediana. As linhas cinzas tracejadas demarcam as posições *top-5* e *top-10*.

Três padrões emergem da Figura 13. Primeiro, nos cenários de *betas altos* (Quant alto e Qual alto), o Par 1 ($\text{SNP}_1 \times \text{SNP}_2$) é consistentemente posicionado no topo do ranking, com mediana 2 e 1 respectivamente, e a maioria das réplicas situadas dentro do top-10. Segundo, nos cenários de *betas baixos*, a dispersão é muito maior: o Par 1 do cenário Quant baixo apresenta mediana na posição 41, com pontos espalhados ao longo de praticamente todo o intervalo $[1, 100]$. Terceiro, o cenário Quant alto, que possue coeficiente de interação de $\gamma_{14} = 3$, apresenta o Par 2 ($\text{SNP}_1 \times \text{SNP}_4$) em posições medianas em torno de 16. Esse resultado sugere que as Florestas de Interação têm dificuldade em priorizar simultaneamente dois pares que compartilham um SNP em comum: a divisão repetida da árvore em torno de SNP_1 acaba favorecendo um único parceiro (SNP_2) em detrimento do outro (SNP_4), mesmo quando ambos têm efeitos verdadeiros relevantes. A Figura 14 mostra a evidência inferencial associada aos pares verdadeiros, expressa como $-\log_{10}(p_{\text{int}})$ do termo de interação no modelo de regressão linear ajustado a cada réplica. As linhas tracejadas marcam os limiares clássicos de $\alpha = 0{,}05$ e $\alpha = 0{,}001$. Nas réplicas em que o par verdadeiro não foi recuperado no top-20 do ranking de EIM, a regressão correspondente não foi ajustada e a observação não comparece no gráfico.

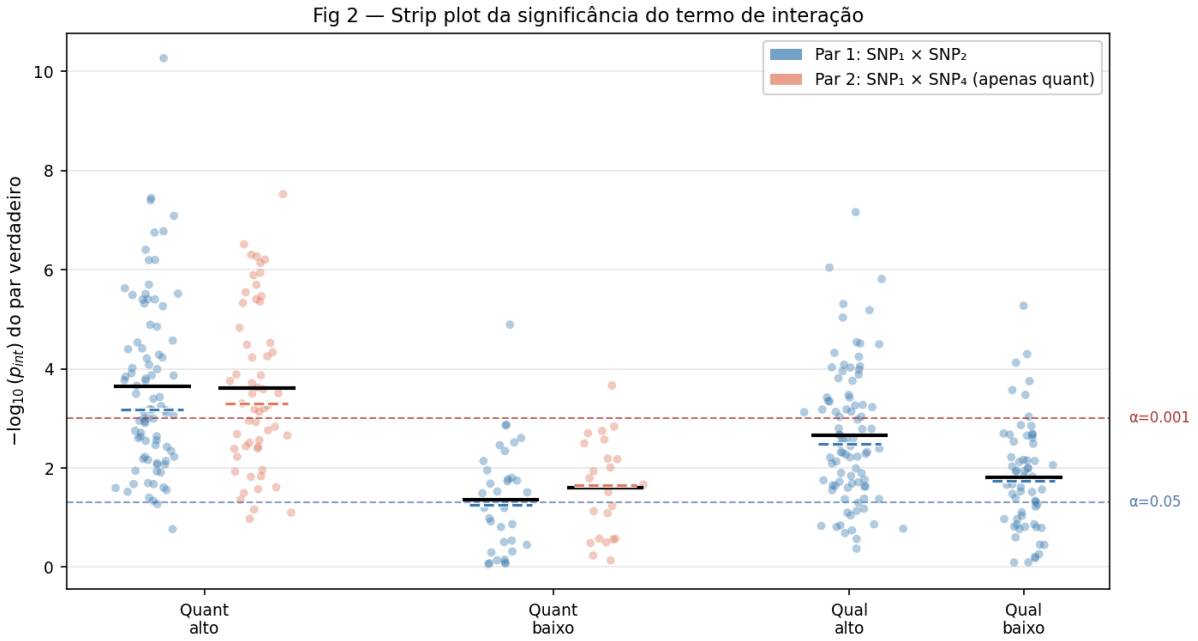

Figura 14 – Significância do termo de interação para os pares verdadeiros em 100 réplicas, por cenário. As linhas tracejadas marcam os limiares $\alpha = 0{,}05$ (azul) e $\alpha = 0{,}001$ (vermelho). Linha preta sólida = média, linha colorida tracejada = mediana.

A Figura 14 confirma o padrão da figura anterior. Nos cenários de *betas altos*, as medianas de $-\log_{10}(p_{\text{int}})$ ficam acima de 2,5, próximas ou superiores ao limiar de $\alpha = 0{,}001$. Nos cenários de *betas baixos*, as medianas mal ultrapassam o limiar de $\alpha = 0{,}05$, e uma fração apreciável das réplicas produz p-valores acima desse nível. A dispersão é também maior nos cenários de baixa magnitude, indicando que a evidência estatística para a interação verdadeira torna-se intermitente quando o sinal é fraco.

Tabela 3 – Taxa de recuperação do par verdadeiro no ranking de EIM em 100 réplicas por cenário. Cada célula indica a proporção de simulações em que o par verdadeiro apareceu até a posição K do ranqueamento.

Cenário	**Par verdadeiro**	**top-1**	**top-5**	**top-10**	**top-20**
Quantitativo (betas altos)	$\text{SNP}_1 \times \text{SNP}_2$	35%	70%	82%	87%
	$\text{SNP}_1 \times \text{SNP}_4$	4%	21%	32%	57%
Quantitativo (betas baixos)	$\text{SNP}_1 \times \text{SNP}_2$	4%	17%	22%	35%
	$\text{SNP}_1 \times \text{SNP}_4$	3%	11%	17%	23%
Qualitativo (betas altos)	$\text{SNP}_1 \times \text{SNP}_2$	54%	78%	84%	89%
Qualitativo (betas baixos)	$\text{SNP}_1 \times \text{SNP}_2$	24%	39%	53%	62%

A Tabela 3 consolida a frequência de recuperação. No cenário Qual alto, o par verdadeiro aparece em primeiro lugar do ranking em 54% das réplicas e dentro do top-20 em 89%. No outro extremo, no cenário Quant baixo, o Par 2 só é recuperado no top-20 em 23% das réplicas, indicando que sob efeitos fracos e com competição entre dois pares simultâneos a confiabilidade

do método cai substancialmente. Esses números reforçam a interpretação visual das Figura 13 e Figura 14: o desempenho do EIM como triagem é fortemente dependente da magnitude do efeito de interação e da existência de competição estrutural entre pares que compartilham SNPs.

A Figura 15 apresenta o número de pares não verdadeiros que aparecem no top-20 do ranking de EIM com $p_{\text{int}} < 0{,}05$, em cada uma das 100 réplicas, por cenário. Esse indicador quantifica a magnitude do ruído inferencial introduzido pelo procedimento de pré-seleção via Florestas de Interação seguido de validação paramétrica.

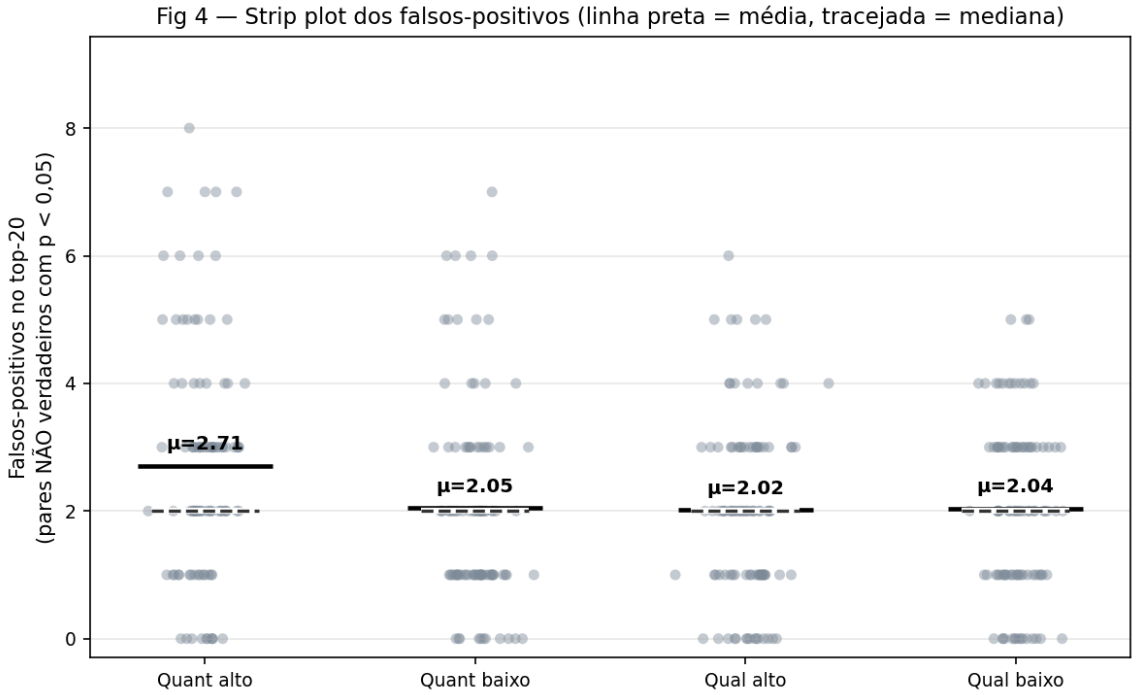

Figura 15 – Número de falsos-positivos no top-20 do ranking de EIM em 100 réplicas por cenário, definidos como pares não simulados com $p_{\text{int}} < 0{,}05$ no modelo de regressão. Linha preta sólida = média (anotada acima de cada coluna), linha tracejada = mediana.

Em todos os quatro cenários, a média de falsos-positivos é próxima de 2 por simulação ($\mu \approx 2{,}71$ em Quant alto, $\mu \approx 2{,}05$ em Quant baixo, $\mu \approx 2{,}02$ em Qual alto, $\mu \approx 2{,}04$ em Qual baixo), com mediana 2 e máximo de 8 em todos os casos. A estabilidade do número de falsos-positivos através dos cenários é informativa, ela indica que aproximadamente 10% das posições do top-20 são, em média, ocupadas por pares espuriamente significativos, independentemente de o cenário ter interações verdadeiras fortes ou fracas. Em outras palavras, os pares verdadeiros competem com um piso constante de $\sim$2 falsos-positivos no top-20 sob $\alpha = 0{,}05$.

4.3.3 Síntese

As 100 réplicas em cada um dos quatro cenários revelam um padrão consistente e mais matizado do que o sugerido por análises baseadas em uma única realização. Em cenários de *betas altos*, sejam interações quantitativas ou qualitativas, as Florestas de Interação recuperam

o par verdadeiro principal com alta confiabilidade (*Recall@20* entre 87% e 89%) e produzem evidência inferencial robusta (mediana de $-\log_{10}(p_{\text{int}})$ acima de 2,5). Em cenários de *betas baixos*, a recuperação cai substancialmente, com taxas *Recall@5* entre 11% e 39%, indicando que o método é sensível à magnitude da interação verdadeira.

Dois resultados adicionais merecem destaque. Primeiro, mesmo no cenário Quantitativo de *betas altos*, com $\gamma_{14} = 3$, o segundo par verdadeiro ($\text{SNP}_1 \times \text{SNP}_4$) só é recuperado no top-20 em 57% das réplicas, e nunca aparece em primeiro lugar com frequência semelhante ao primeiro par. Esse resultado expõe uma limitação estrutural do método quando dois pares verdadeiros compartilham um SNP em comum: a árvore tende a particionar os dados em torno desse SNP por meio de divisões com um único parceiro, em detrimento dos demais. Em estudos genômicos onde múltiplos pares de interação podem coexistir em torno de um SNP-âncora, esse comportamento implica que o método pode subestimar sistematicamente parte das interações relevantes.

Segundo, o número médio de falsos-positivos no top-20 sob $\alpha = 0{,}05$ é estável e próximo de 2 em todos os cenários, refletindo um ruído de fundo intrínseco ao procedimento de pré-seleção combinada com validação paramétrica. Esse achado responde diretamente à preocupação de que análises baseadas em rankings de importância possam produzir confiança falsa: mesmo na ausência de qualquer interação real fora dos pares simulados, o pipeline retorna em média dois pares espúrios significativos em cada simulação. Em aplicações reais, isso reforça a necessidade de critérios mais conservadores ($\alpha \leq 10^{-5}$, por exemplo) ou de análise de redes para destacar pares com alta consistência através de cromossomos e protocolos.

Em conjunto, os quatro cenários demonstram que as Florestas de Interação são um instrumento de triagem útil em alta dimensionalidade, com bom desempenho quando as interações verdadeiras têm magnitude apreciável, mas que devem ser combinadas com validação paramétrica e com julgamento crítico sobre a estrutura do espaço de pares para mitigar tanto subdetecção quanto falsos-positivos.

CAPÍTULO

5

RESULTADOS

Neste capítulo, avaliamos o desempenho da Floresta de Interação nos dados genômicos reais do estudo ISA-Capital. Apresentamos inicialmente uma visão geral do conjunto de SNPs disponíveis após o controle de qualidade e filtragem por desequilíbrio de ligação, incluindo o gráfico de Manhattan global. Em seguida, descrevemos e comparamos os três cenários de aplicação do método (análise global, por cromossomo e com pré-seleção via regressão linear simples), detalhando os pares de SNPs prioritários e os resultados dos modelos de regressão com termos de interação. Por fim, exploramos as redes de interação construídas a partir dos pares selecionados, destacando os agrupamentos de SNPs que sugerem possíveis módulos funcionais associados aos níveis de glicose.

5.1 Avaliação do método em dados genômicos

Conforme descrito no Capítulo 4, utilizamos os dados do estudo ISA-Capital. Após o controle de qualidade e a filtragem por desequilíbrio de ligação, o conjunto integrado dos 22 cromossomos resultou em 324.471 SNPs disponíveis para análise, tendo como desfecho o nível de glicose. Para a análise de associação marginal, ajustou-se um modelo de regressão linear simples para cada SNP como o apresentado na Equação (4.1), considerado individualmente como covariável explicativa, a partir do qual foram obtidos os respectivos p-valores. A Figura 16 apresenta o gráfico de Manhattan para todos os cromossomos, destacando os principais picos de associação marginal.

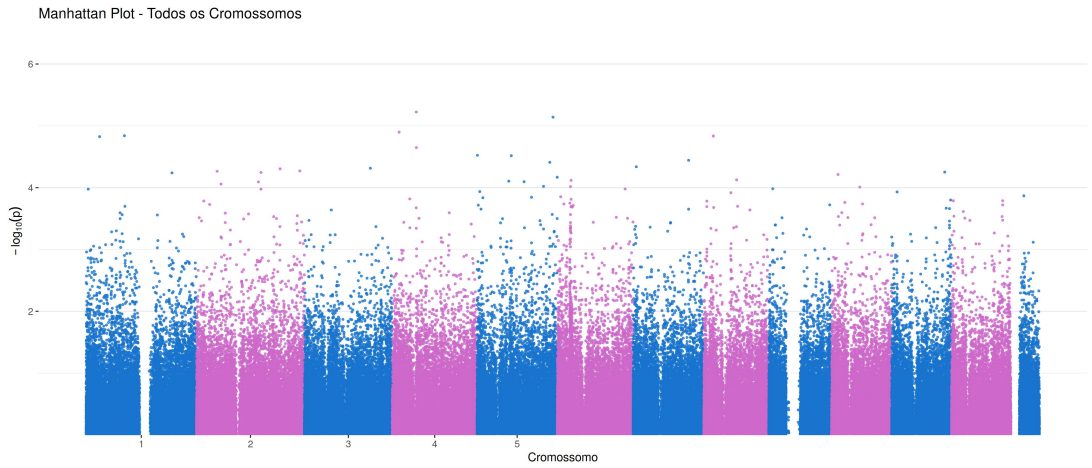

Figura 16 – Gráfico de Manhattan dos valores de $-\log_{10}(p)$ para todos os cromossomos.

5.1.1 Combinação entre EIM e p-valor: triagem preditiva e validação inferencial

A análise apresentada nas próximas seções emprega duas medidas distintas para a identificação de interações entre SNPs: a Effect Importance Measure (EIM), produzida pelas Florestas de Interação, e o p-valor do termo de interação em modelos de regressão linear. Embora ambas avaliem a relevância de pares de SNPs, elas pertencem a paradigmas diferentes e cumprem papéis complementares no fluxo metodológico.

A EIM é uma medida preditiva: quantifica o ganho de acurácia que o par $(\text{SNP}_i, \text{SNP}_j)$ traz à floresta quando usado em divisões bivariadas, em comparação com a desativação dessas divisões pelo procedimento de Proportional Randomization Hapfelmeier *et al.* 2014. Trata-se, portanto, de uma métrica de relevância para predição, baseada em validação out-of-bag, sem hipóteses paramétricas sobre a forma funcional do modelo. Sua principal vantagem é escalar bem em alta dimensionalidade, como os $\binom{324471}{2} \approx 5{,}3 \times 10^{10}$ pares possíveis no banco completo, e capturar efeitos não lineares ou de interação que escapariam de modelos puramente aditivos.

O p-valor do termo de interação, por sua vez, é uma medida inferencial: quantifica a evidência estatística contra a hipótese nula de ausência de interação no modelo paramétrico ajustado para o par específico, condicionalmente às covariáveis. Diferentemente da EIM, ele oferece uma escala calibrada de probabilidade que permite controle de erro tipo I, a custo de exigir a especificação do modelo (linearidade, codificação aditiva dos genótipos, normalidade dos resíduos) e de não ser viável aplicá-lo exaustivamente a todos os pares possíveis em um banco genômico de larga escala.

A combinação das duas ferramentas explora as forças de cada uma: a EIM atua como triagem preditiva, reduzindo o espaço de busca de milhões de pares para algumas centenas; e o p-valor atua como validação inferencial, fornecendo evidência estatística para os pares priorizados. Como mostrado no estudo de simulação do Capítulo 4, a EIM consegue priorizar

interações verdadeiras dentro do top-K, mas o procedimento gera, em média, cerca de dois falsos-positivos por simulação no top-20 sob $\alpha = 0{,}05$. Esse comportamento reforça que a EIM não deve ser interpretada como medida unica de significância: pares com EIM elevada e p-valor próximo de 1 não constituem evidência de interação, da mesma forma que pares com EIM modesta podem ainda assim apresentar evidência inferencial robusta. Em síntese, no fluxo proposto a EIM responde quais pares merecem investigação e o p-valor responde quão fortes são as evidências para cada um.

5.1.2 Os três cenários de aplicação

Com o objetivo de investigar diferentes estratégias para identificação de interações entre SNPs, consideramos três abordagens complementares, resumidas na Figura 17:

- Caso 1 (global): reunimos todos os SNPs dos 22 cromossomos em um único data frame e aplicamos diretamente a Floresta de Interação. Para conjuntos com mais de 30 000 variáveis, o método retorna apenas os 5 000 pares com maiores valores de EIM.
- Caso 2 (por cromossomo): ajustamos 22 Florestas de Interação, uma para cada cromossomo, obtendo aproximadamente 110 000 valores de EIM ao combinar os resultados cromossomo a cromossomo.
- Caso 3 (pré-seleção por regressão): inicialmente ajustamos uma regressão linear simples para cada SNP, incluindo como covariáveis o sexo, a idade, o uso de medicamento para diabetes e as duas primeiras componentes principais (PC1 e PC2). Em seguida, selecionamos os 400 SNPs com menores p-valores marginais, vistos anteriormente no gráfico Figura 16, e aplicamos a Floresta de Interação apenas a esse subconjunto de marcadores.

Após essas etapas específicas de cada cenário, adotamos um procedimento comum aos três casos:

(i) seleção dos pares de SNPs com maiores valores de EIM em cada abordagem (200 pares por cromossomo no Caso 2, 200 pares no total nos Casos 1 e 3);

(ii) ajuste, para cada par $(\text{SNP}_1, \text{SNP}_2)$ priorizado, de um modelo de regressão linear com termo de interação genética dado por

$$\begin{aligned} \log(Y_i) = {} & \beta_0 + \beta_1 \text{sexo}_i + \beta_2 \text{idade}_i + \beta_3 \text{medDM\_bioq}_i \\ & + \beta_4 \text{PC1}_i + \beta_5 \text{PC2}_i + \beta_6 \text{SNP}_{1i} + \beta_7 \text{SNP}_{2i} \\ & + \beta_8 (\text{SNP}_{1i} \times \text{SNP}_{2i}) + \varepsilon_i, \end{aligned} \tag{5.1}$$

em que as covariáveis fixas (sexo, idade, medDM_bioq, PC1 e PC2) são incluídas em todos os ajustes;

(iii) reporte dos p-valores marginais dos efeitos principais e do p-valor associado ao termo de interação (p_{int}), destacando-se os pares com $p_{\text{int}} < 10^{-5}$.

Os conjuntos de pares resultantes foram então avaliados em conjunto e são apresentados nas seções seguintes. O delineamento esquemático das três estratégias está ilustrado na Figura 17.

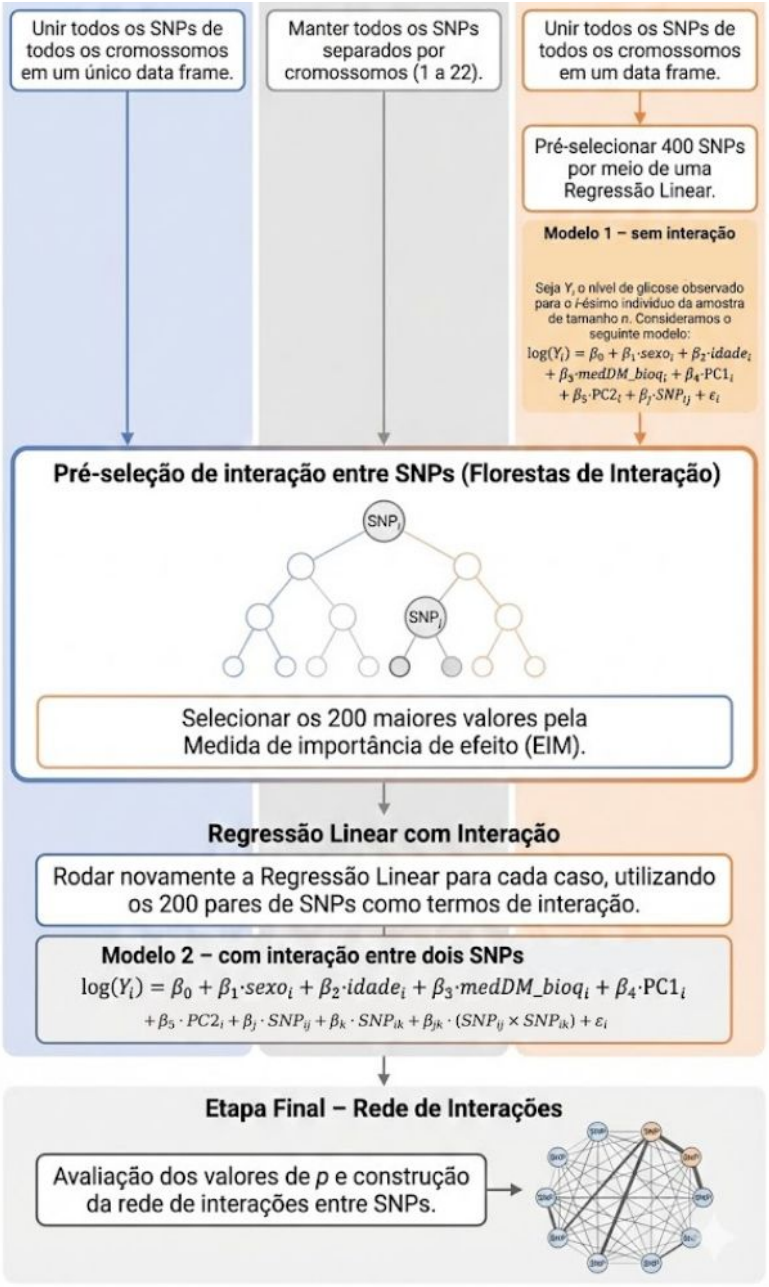

Figura 17 – Fluxo da análise de interação entre SNPs nos três cenários considerados.

5.2 Caso 1 (global)

No Caso 1, os 324 471 SNPs codificados como (0/1/2) são fornecidos em conjunto à Floresta de Interação. O algoritmo retorna listas separadas para interações quantitativas (mudança

de magnitude do efeito) e qualitativas (inversão de sinal). A partir dos 200 pares com maiores valores de EIM, ajustamos o modelo de regressão da Equação (5.1) e calculamos p_{int} para cada par.

A Figura 18 apresenta a relação entre EIM e $-\log_{10}(p_{\text{int}})$ para os 200 pares avaliados. O objetivo desse gráfico é verificar se há associação direta entre as duas métricas: caso a houvesse, valores altos de EIM corresponderiam a valores baixos de p. No entanto, os pontos distribuem-se de forma dispersa e sem tendência linear clara, o que é coerente com a discussão da Subseção 5.1.1: a EIM ranqueia pares pela contribuição preditiva na floresta, enquanto o p-valor mede evidência inferencial em um modelo paramétrico, dois critérios que não precisam estar fortemente correlacionados ponto a ponto.

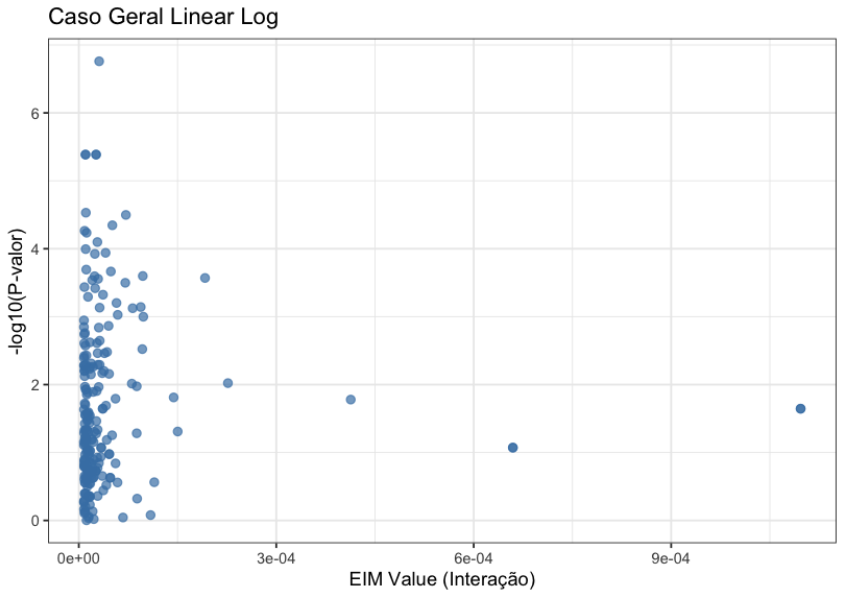

Figura 18 – Caso 1: EIM (Interaction Forest) versus $-\log_{10}(p_{\text{int}})$ para os 200 pares priorizados.

A Tabela 4 apresenta os cinco pares com menor p_{int} no Caso 1. O resultado mais notável é a presença de associações com $p_{\text{int}} < 10^{-7}$, com o par `rs6800849_A` (chr3) e `rs11007430_A` (chr10) apresentando $p_{\text{int}} = 1{,}74 \times 10^{-7}$. Uma versão ampliada desta tabela, contendo os 20 menores valores de p_{int}, é apresentada no Apêndice A (Tabela 10).

snp1	snp2	eim_value	coef_snp1	pval_snp1	coef_snp2	pval_snp2	coef_interacao	pval_interacao
rs6800849_A (chr3)	rs11007430_A (chr10)	3.09e-05	-0.0482216	2.46e-02	-0.0196231	1.22e-01	0.1043843	1.74e-07
rs411337_T (chr6)	rs8027766_C (chr15)	2.62e-05	0.0817586	3.00e-06	0.0470140	6.23e-03	-0.1214829	4.11e-06
rs411337_T (chr6)	rs8027766_C (chr15)	2.62e-05	0.0817586	3.00e-06	0.0470140	6.23e-03	-0.1214829	4.11e-06
rs1002174_G (chr15)	rs12917291_A (chr15)	1.07e-05	-0.0037859	7.93e-01	-0.0414024	1.39e-02	0.1005886	2.95e-05
rs7578812_T (chr2)	rs740420_A (chr12)	7.15e-05	0.0710932	1.85e-06	0.0249232	1.30e-01	-0.0725253	3.18e-05

Tabela 4 – Cinco menores p-valores do termo de interação (modelo linear ajustado) com EIM e cromossomos dos SNPs (Caso 1, global).

Para examinar visualmente a interação detectada no par com menor p_{int} deste caso, a Figura 19 apresenta o gráfico de perfil de médias para o par `rs6800849_A` × `rs11007430_A`.

No eixo horizontal estão dispostos os três genótipos do primeiro SNP (codificados como 0, 1 e 2, correspondendo, respectivamente, a homozigose para o alelo de referência, heterozigose e homozigose para o alelo alternativo). No eixo vertical aparece a média de log(glicose) com intervalos de confiança de 95%, e cada linha representa um dos três genótipos do segundo SNP.

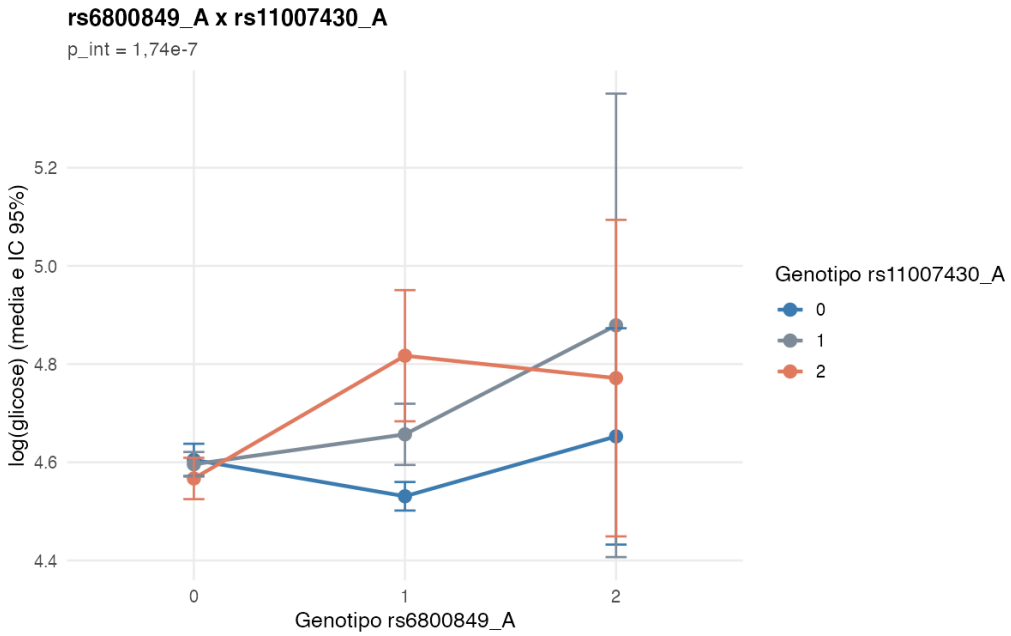

Figura 19 – Perfil de médias para o par com menor p_{int} do Caso 1: `rs6800849_A` (chr3) × `rs11007430_A` (chr10). Cada linha representa um genótipo de `rs11007430_A`; o eixo horizontal mostra o genótipo de `rs6800849_A`.

A Figura 19 mostra que as três linhas não são paralelas: o efeito do genótipo de `rs6800849_A` sobre a glicose depende do genótipo de `rs11007430_A`. Esse comportamento é a expressão visual da interação que o termo cruzado capta no modelo de regressão e cuja significância é avaliada por p_{int}.

5.3 Caso 2 (por cromossomo)

No Caso 2, as Florestas de Interação são ajustadas separadamente para cada um dos 22 cromossomos, preservando a estrutura cromossômica e produzindo 44 listas (22 cromossomos × componentes quantitativa e qualitativa). Para cada cromossomo, selecionamos os 200 pares com maiores valores de EIM e validamos por meio do modelo da Equação (5.1).

A Figura 20 apresenta a relação EIM versus $-\log_{10}(p_{\text{int}})$, com cores identificando o cromossomo de origem do par. Como no Caso 1, não emerge uma tendência linear clara entre as duas métricas; p-valores muito pequenos aparecem distribuídos numa faixa de EIM majoritariamente baixa. Esse padrão reforça o uso da figura como ferramenta de triagem visual, identificando pares que se destacam em relação aos demais por possuírem simultaneamente EIM elevada e forte evidência inferencial.

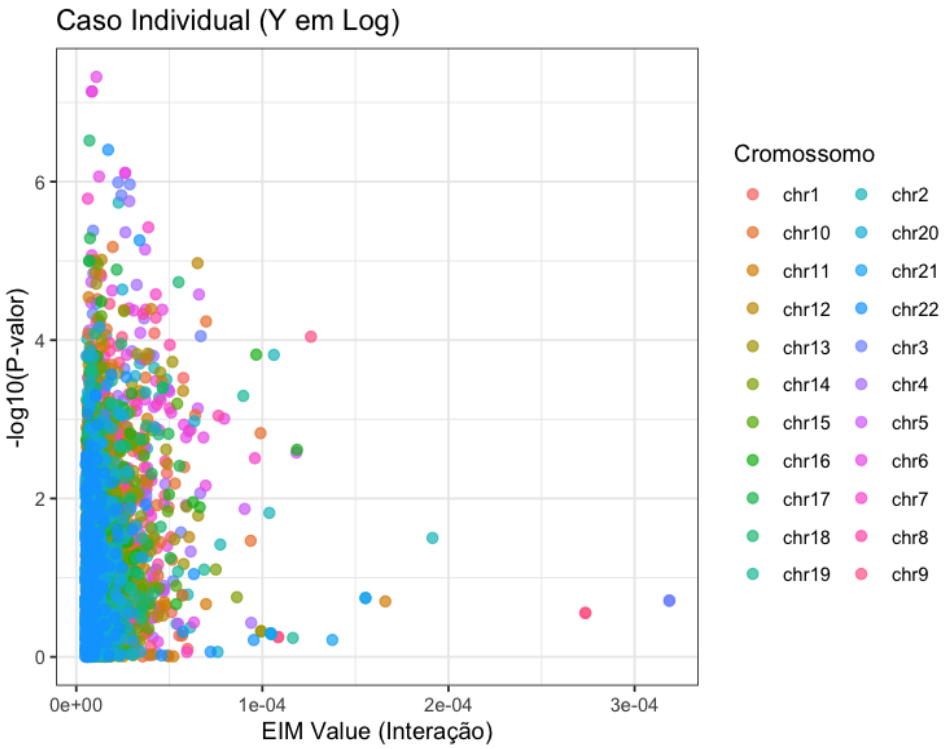

Figura 20 – Caso 2: EIM (Interaction Forest) versus $-\log_{10}(p_{\text{int}})$ para os pares priorizados em cada cromossomo. Cada ponto representa um par SNP×SNP; as cores identificam o cromossomo.

A Tabela 5 apresenta os cinco pares com menor p_{int} no Caso 2. Comparativamente ao Caso 1, observa-se uma melhora na magnitude da evidência estatística para os pares de topo, com associações chegando a $p_{\text{int}} < 10^{-7}$ e até $< 10^{-8}$. Notavelmente, o cromossomo 6 fornece dois dos cinco pares de maior destaque, ambos envolvendo `rs6457131`: a interação com `rs3131787_C` ($p_{\text{int}} = 4{,}76 \times 10^{-8}$) e com `rs4678_A` ($p_{\text{int}} = 7{,}25 \times 10^{-8}$). Uma versão ampliada desta tabela, contendo os 20 menores valores de p_{int}, é apresentada no Apêndice A (Tabela 9).

	snp1	snp2	eim_value	coef_snp1	pval_snp1	coef_snp2	pval_snp2	coef_interacao	pval_interacao
chr6	rs6457131_T	rs3131787_C	1.08e-05	0.0371699	8.08e-03	0.1122360	2.91e-07	-0.1587490	4.76e-08
chr6	rs6457131_T	rs4678_A	8.40e-06	0.0369294	8.64e-03	0.1101974	4.43e-07	-0.1561574	7.25e-08
chr18	rs9947371_C	rs56374046_A	7.09e-06	-0.0572183	7.34e-04	-0.0444386	2.08e-03	0.1078512	3.03e-07
chr22	rs78380435_A	rs4821513_C	1.71e-05	0.1780199	4.33e-08	0.0133702	2.54e-01	-0.1740723	3.96e-07
chr6	rs2575174_T	rs13195572_C	2.62e-05	0.0366364	2.51e-03	9.95e-02	4.19e-05	-0.1031389	7.74e-07

Tabela 5 – Cinco menores p-valores do termo de interação (modelo linear ajustado) com EIM e cromossomos dos SNPs (Caso 2, por cromossomo).

A Figura 21 apresenta o perfil de médias para o par com menor p_{int} deste caso, `rs6457131_T` × `rs3131787_C`, ambos no cromossomo 6.

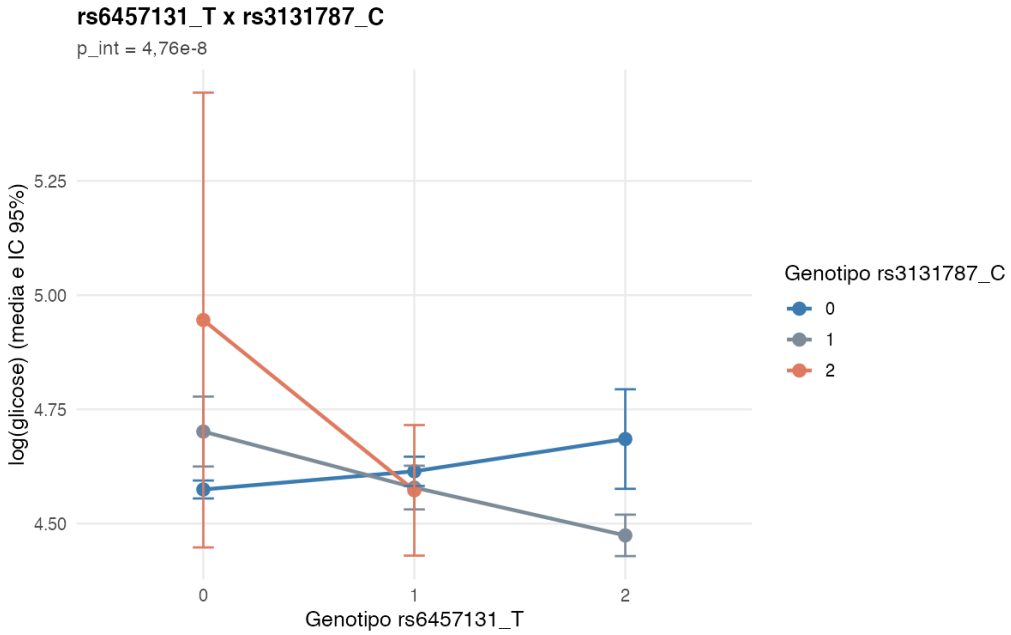

Figura 21 – Perfil de médias para o par com menor p_{int} do Caso 2: `rs6457131_T` (chr6) × `rs3131787_C` (chr6). Cada linha representa um genótipo de `rs3131787_C`; o eixo horizontal mostra o genótipo de `rs6457131_T`.

O comportamento das linhas neste caso é particularmente ilustrativo da interação capturada pelo termo cruzado: o coeficiente de interação negativo ($\beta_{\text{int}} = -0{,}159$) implica que o efeito do alelo menor de `rs6457131_T` sobre a glicose se inverte conforme o genótipo de `rs3131787_C`, padrão consistente com o tipo de interação qualitativa discutida no Capítulo 3.

5.4 Caso 3 (pré-seleção por regressão)

No Caso 3, partimos de uma etapa de pré-seleção baseada na evidência marginal: ajustamos regressões lineares simples de log(glicose) em cada SNP individualmente, conforme o modelo da Equação (4.1), e selecionamos os 400 SNPs com menores p-valores marginais. Em seguida, aplicamos a Floresta de Interação apenas a esse subconjunto de marcadores e validamos os 200 pares com maiores EIM por meio do modelo da Equação (5.1).

A Figura 22 apresenta a relação EIM versus $-\log_{10}(p_{\text{int}})$ para os pares priorizados. O comportamento qualitativo é análogo ao dos casos anteriores, com ausência de relação linear estrita entre EIM e $-\log_{10}(p_{\text{int}})$, mas a magnitude dos p-valores no topo do painel é substancialmente maior. Vários pares atingem $-\log_{10}(p_{\text{int}}) > 12$, correspondendo a $p_{\text{int}} < 10^{-12}$, valores extremamente baixos que excedem com folga os limiares clássicos de genome-wide significance (5×10^{-8}).

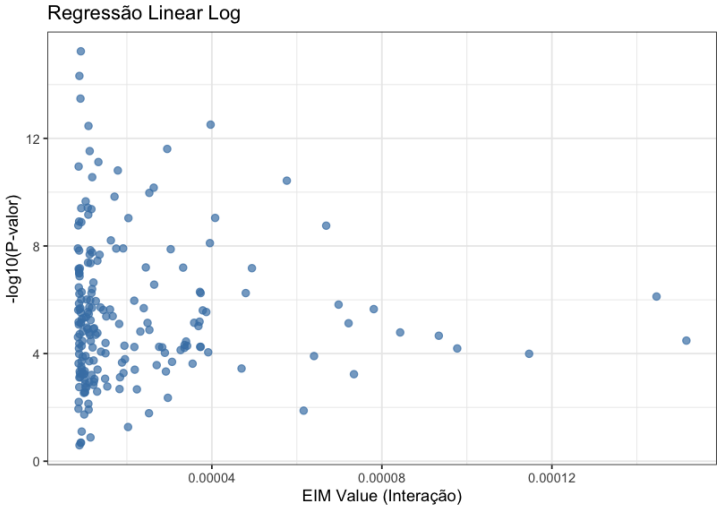

Figura 22 – Caso 3 (pré-seleção por regressão): EIM (Interaction Forest) versus $-\log_{10}(p_{\text{int}})$. A linha vermelha é um ajuste linear simples.

A Tabela 6 apresenta os cinco pares com menor p_{int} no Caso 3, todos com $p_{\text{int}} < 10^{-12}$. O par `rs55768887_T` (chr5) e `rs73256166_A` (chr6) apresenta o menor valor ($p_{\text{int}} = 5{,}81 \times 10^{-16}$), com coeficiente de interação positivo e elevado ($\beta_{\text{int}} = 0{,}425$) e efeitos marginais de ambos os SNPs próximos de zero ($p > 0{,}6$), exemplo claro de interação detectável apenas pela inclusão do termo cruzado, que não emergiria de uma análise puramente marginal.

snp1	snp2	eim_value	coef_snp1	pval_snp1	coef_snp2	pval_snp2	coef_interacao	pval_interacao
rs55768887_T (chr5)	rs73256166_A (chr6)	9.17e-06	0.0050677	8.33e-01	0.0080426	6.75e-01	0.4253349	5.81e-16
rs73286055_A (chr17)	rs75689771_G (chr10)	8.84e-06	0.0171030	3.94e-01	0.0124439	4.18e-01	0.2606503	4.79e-15
rs29623_A (chr5)	rs4432172_A (chr14)	9.08e-06	0.0117901	5.88e-01	0.0047521	8.09e-01	0.3273849	3.32e-14
medDM_bioq (chrNA)	rs13127462_A (chr4)	3.97e-05	0.5267797	5.35e-38	-0.0089331	4.15e-01	-0.2476684	3.09e-13
rs4704396_C (chr5)	rs17179053_C (chr9)	1.10e-05	0.0096938	6.24e-01	-0.0102908	6.25e-01	0.3104587	3.46e-13

Tabela 6 – Cinco menores p-valores do termo de interação (modelo linear ajustado) com EIM e cromossomos dos SNPs (Caso 3, pré-seleção por regressão).

Uma versão ampliada desta tabela, contendo os 20 menores valores de p_{int}, é apresentada no Apêndice A (Tabela 11).

A Figura 23 apresenta o perfil de médias para o par com menor p_{int} do Caso 3 e da dissertação como um todo, `rs55768887_T` $\times$ `rs73256166_A`.

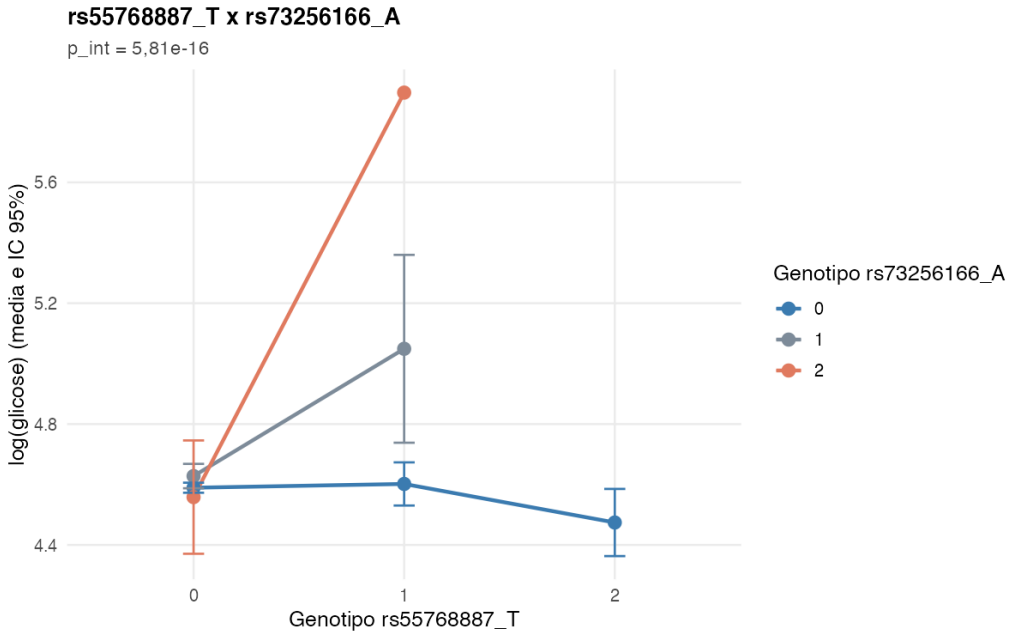

Figura 23 – Perfil de médias para o par com menor p_{int} do Caso 3: `rs55768887_T` (chr5) $\times$ `rs73256166_A` (chr6). Cada linha representa um genótipo de `rs73256166_A`; o eixo horizontal mostra o genótipo de `rs55768887_T`.

A divergência das linhas é especialmente acentuada neste caso, refletindo a magnitude do coeficiente de interação $\beta_{\text{int}} = 0{,}425$, o maior em valor absoluto entre os pares-top dos três casos. Como ambos os efeitos marginais individuais são essencialmente nulos ($p > 0{,}6$), o padrão observado configura uma interação pura: a glicose só é afetada quando os dois SNPs operam em conjunto, sem que nenhum deles, isoladamente, contribua de forma detectável. Para examinar se o padrão de interação é homogêneo entre subgrupos populacionais, as Figura 24 e Figura 25 apresentam estratificações do mesmo par por sexo e por faixa etária, respectivamente.

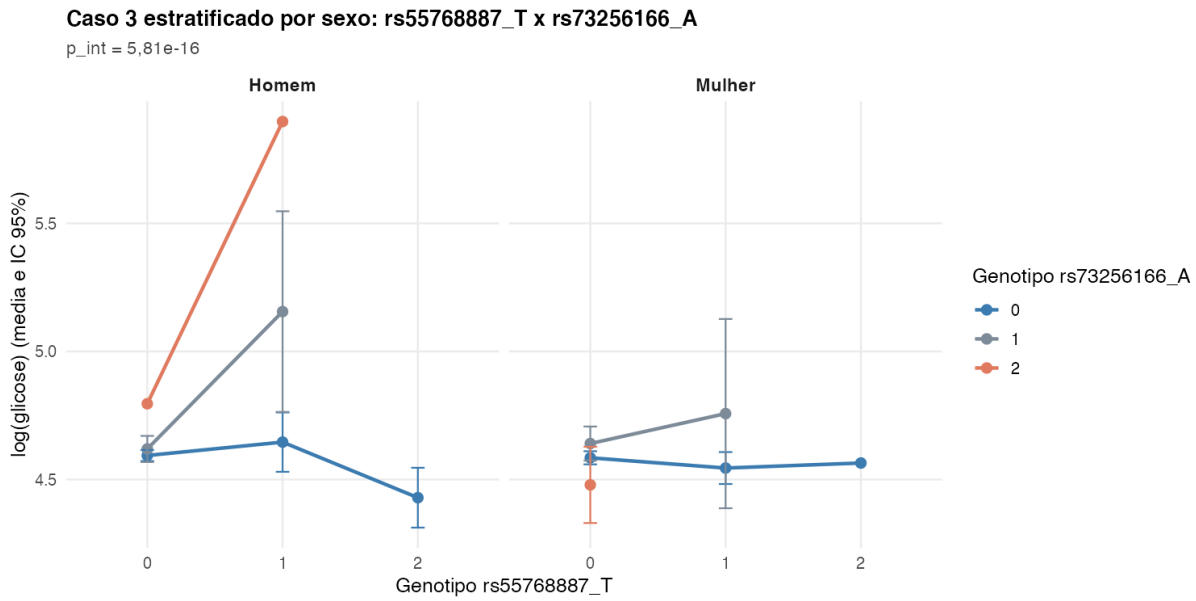

Figura 24 – Perfil de médias do par `rs55768887_T` × `rs73256166_A` estratificado por sexo. Painéis: homens (esquerda) e mulheres (direita).

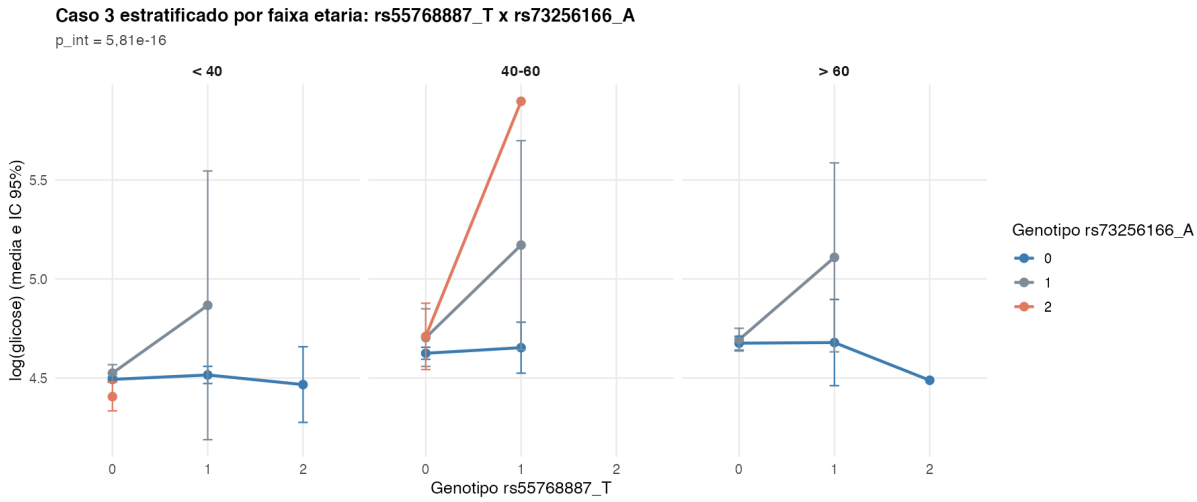

Figura 25 – Perfil de médias do par `rs55768887_T` × `rs73256166_A` estratificado por faixa etária. Painéis: < 40 anos, 40 a 60 anos e > 60 anos.

A análise estratificada permite avaliar se o padrão de interação observado no painel agregado (Figura 23) é consistente entre subgrupos. A inspeção visual da Figura 24 indica que o padrão de divergência das linhas é qualitativamente preservado tanto em homens quanto em mulheres, embora as magnitudes médias da glicose difiram entre os dois grupos. Já na Figura 25, observa-se que a interação se manifesta com clareza nas faixas etárias intermediária e superior, enquanto na faixa de menor idade os intervalos de confiança são consideravelmente mais amplos, reflexo do menor número de indivíduos jovens com diagnóstico de alterações glicêmicas relevantes na amostra. A consistência qualitativa entre os subgrupos sugere que a interação detectada não é artefato de uma subpopulação específica, embora uma avaliação

formal exigiria a inclusão de termos triplos do tipo SNP×SNP×covariável, fora do escopo deste trabalho.

5.4.1 Comparação entre os três casos

A comparação entre as tabelas 4, 5 e 6 mostra que não há SNPs em comum entre os três casos, nem repetição de pares específicos entre as análises. A única recorrência de variantes observada ocorre dentro da Tabela 5, onde o SNP `rs6457131` (chr6) participa de dois pares distintos, interagindo com `rs3131787` e `rs4678`. Ainda assim, observa-se uma consistência em nível cromossômico: o cromossomo 6 está presente em todas as três listas, e o cromossomo 10 aparece nos Casos 1 e 3.

A inexistência de pares específicos comuns entre os três casos, a despeito da consistência cromossômica, ilustra um achado já antecipado pelo estudo de simulação do Capítulo 4: o ranqueamento por EIM apresenta variabilidade não negligenciável, sensível ao espaço de pares considerado e à estrutura local dos dados. A pré-seleção por regressão (Caso 3) restringe o espaço de busca aos SNPs com efeito marginal já estabelecido e produz, sistematicamente, os menores p-valores de interação. Em contraste, a análise global (Caso 1) e a análise por cromossomo (Caso 2) operam sobre conjuntos de pares mais amplos, nos quais a EIM disputa atenção entre milhares de candidatos sem filtro prévio de evidência marginal. Como consequência, embora os três casos identifiquem pares com forte evidência inferencial, eles iluminam regiões diferentes do espaço de interações, e seu uso conjunto fornece uma cobertura mais ampla do que cada estratégia isoladamente.

5.5 Redes de interação

Os resultados obtidos nas três abordagens (global, por cromossomo e com pré-seleção por regressão) geraram listas de pares SNP×SNP com evidência de interação, quantificada tanto pela EIM quanto pelos p-valores dos termos de interação (p_{int}). Para sintetizar essas informações e destacar padrões estruturais, construímos redes de interação entre SNPs.

Do ponto de vista computacional, as redes foram construídas a partir das listas de pares selecionados em cada caso, considerando diferentes valores de corte para o termo de interação ($p_{\text{int}} < \tau$), com $\tau \in \{10^{-8}, 10^{-7}, 10^{-6}, 10^{-5}\}$. Para cada conjunto de pares (definido por um caso e por um valor de τ), definimos uma matriz de adjacência binária $\mathbf{A} = (a_{ij})$, em que as linhas e colunas correspondem aos SNPs presentes na lista e

$$a_{ij} = \begin{cases} 1, & \text{se o par } (\text{SNP}_i,\ \text{SNP}_j) \text{ foi selecionado,} \\ 0, & \text{caso contrário.} \end{cases} \tag{5.2}$$

A matriz é construída de forma simétrica (isto é, $a_{ij} = a_{ji}$), representando uma rede não direcionada. A partir de $\mathbf{A}$, calculamos o grau de cada nó (número de interações em que o SNP

participa) e restringimos a visualização aos SNPs com grau maior que 1, de modo a destacar apenas os marcadores que compõem componentes conectados da rede. Em seguida, geramos coordenadas bidimensionais para os nós e traçamos as arestas correspondentes, resultando nas redes apresentadas adiante.

Nos Casos 1 e 2, mesmo variando o corte τ em $\{10^{-8}, 10^{-7}, 10^{-6}, 10^{-5}\}$, não houve formação de componentes conectados após a filtragem por grau, de modo que não emergiram subestruturas (por exemplo, ciclos) passíveis de análise em rede. Esse resultado é coerente com a observação de que os pares priorizados nesses dois casos tendem a envolver SNPs distintos sem reaparição em múltiplos pares; assim, não reportamos redes nesses dois cenários.

No Caso 3 (pré-seleção de SNPs por regressão linear), em contraste, a construção das redes de interação evidencia a emergência de estruturas conectadas com fechamento topológico à medida que o critério de significância do termo de interação é relaxado. Conforme ilustrado na Figura 26, para $p_{\text{int}} < 10^{-5}$ observa-se uma rede mais densa, com maior número de nós e arestas e a presença de múltiplos ciclos (cinco ciclos detectados), sugerindo a formação de subestruturas (módulos) nas quais os SNPs se conectam de forma recorrente. Ao adotar limiares mais conservadores ($p_{\text{int}} < 10^{-6}$ e $p_{\text{int}} < 10^{-7}$), a rede torna-se progressivamente mais esparsa e a quantidade de ciclos diminui (dois e um ciclo, respectivamente), indicando que apenas as interações mais robustas permanecem. Em particular, no limiar $p_{\text{int}} < 10^{-7}$ subsiste um ciclo mínimo (triângulo) composto por `rs55768887_T`, `rs55660132_G` e `rs73256166_A`, o que caracteriza um núcleo de interações que persiste sob maior rigor estatístico.

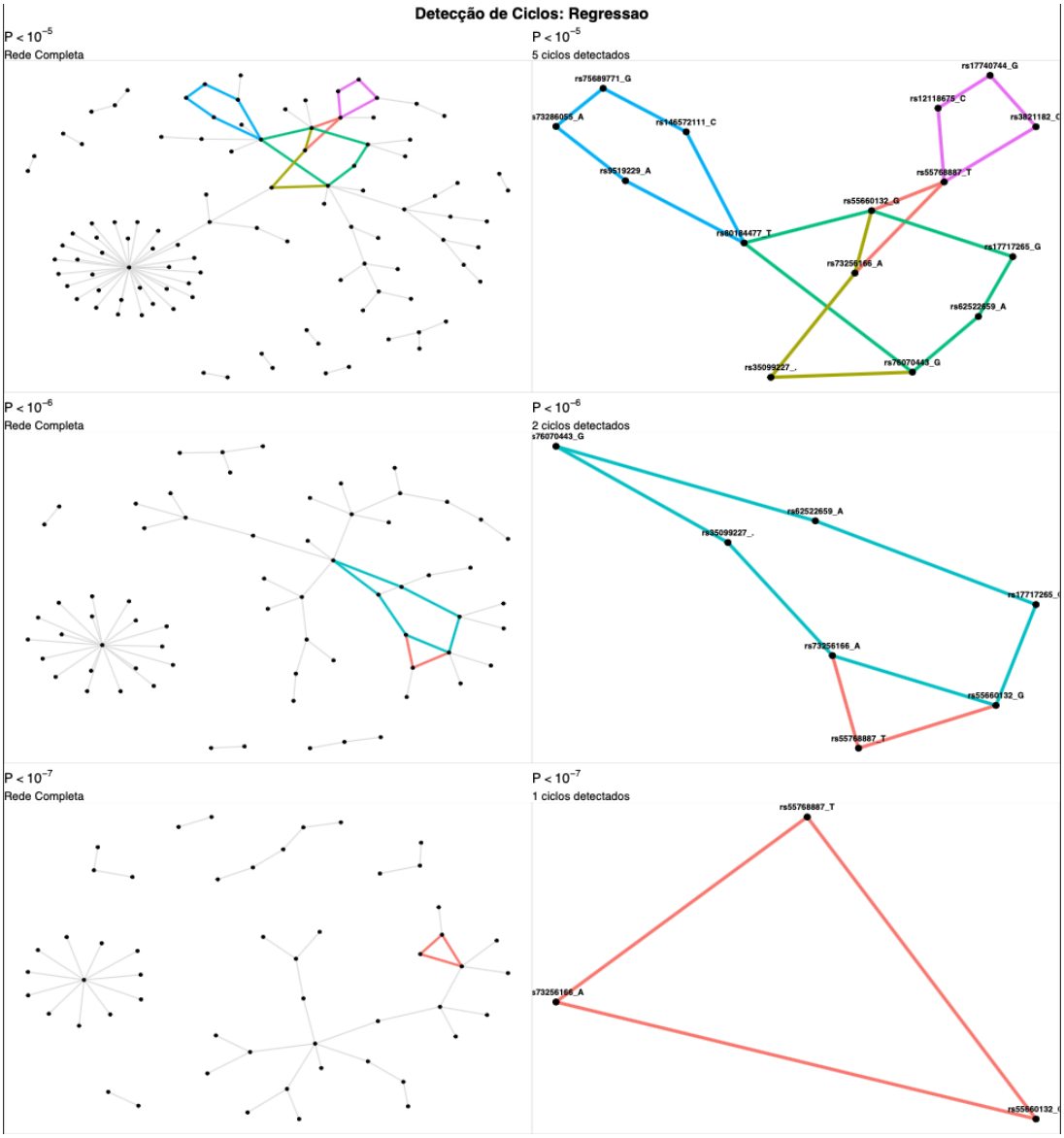

Figura 26 – Rede de interações no Caso 3 (pré-seleção por regressão). As cores das arestas indicam ciclos identificados no grafo; nós com maior grau correspondem a SNPs que participam de múltiplas interações significativas.

A Tabela 7 complementa a interpretação da Figura 26 ao sumarizar, para cada valor de corte, a complexidade estrutural do grafo e a composição dos ciclos identificados. Em termos gerais, a redução de nós e arestas acompanhada do decréscimo no número de ciclos (de $5 \to 2 \to 1$ ao passar de 10^{-5} para 10^{-7}) é consistente com a ideia de que limiares menos restritivos revelam módulos de interação mais amplos, enquanto limiares mais conservadores preservam apenas subestruturas recorrentes e mais estáveis. Assim, a análise topológica fornece uma síntese quantitativa e qualitativa das interações SNP×SNP no cenário com pré-seleção, destacando simultaneamente módulos mais extensos (em 10^{-5}) e um subconjunto central de interações (em 10^{-7}) potencialmente prioritário para investigação subsequente.

A rede da Figura 27 destaca o SNP de maior grau, `rs76070443_G`, no Caso 3, considerando pares com $p_{\text{int}} < 10^{-5}$. O nó central representa `rs76070443_G` e os nós periféricos

Tabela 7 – Análise topológica e identificação de ciclos epistáticos no Caso 3 (pré-seleção por regressão), para diferentes limiares de significância do termo de interação ($p_{\text{int}} < \tau$).

P-Valor	Nós	Arestas	Ciclos	SNPs nos Ciclos
10^{-5}	103	98	5	1. rs55768887_T, rs55660132_G, rs17717265_G, rs62522659_A, rs76070443_G, rs35099227_., rs73256166_A 2. rs55660132_G, rs17717265_G, rs62522659_A, rs76070443_G, rs35099227_., rs73256166_A 3. rs55660132_G, rs17717265_G, rs62522659_A, rs76070443_G, rs80184477_T 4. rs80184477_T, rs146572111_C, rs75689771_G, rs73286055_A, rs9519229_A 5. rs55768887_T, rs12118675_C, rs17740744_G, rs3821182_C
10^{-6}	66	62	2	1. rs55768887_T, rs55660132_G, rs17717265_G, rs62522659_A, rs76070443_G, rs35099227_., rs73256166_A 2. rs55660132_G, rs17717265_G, rs62522659_A, rs76070443_G, rs35099227_., rs73256166_A
10^{-7}	52	46	1	1. rs55768887_T, rs55660132_G, rs73256166_A

correspondem aos SNPs que interagem significativamente com ele. A relevância dessas conexões deve ser confirmada pela sua persistência sob limiares mais conservadores e pela consistência com módulos/ciclos e evidências biológicas.

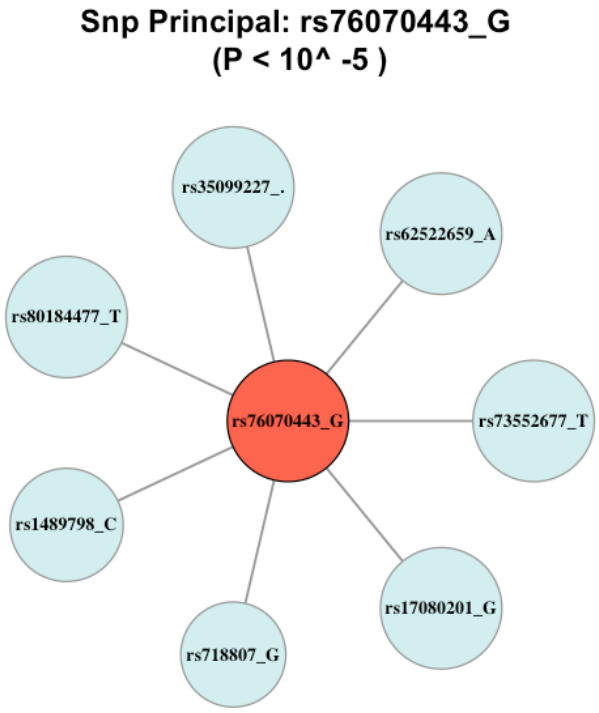

Figura 27 – Rede de vizinhança do SNP principal `rs76070443_G` no Caso 3.

Com base nos pares SNP×SNP selecionados pelas Interaction Forests e posteriormente avaliados em modelos de regressão com termo de interação, os SNPs retidos foram utilizados para a construção da rede de interações, da qual se obteve um conjunto final de 15 SNPs para uma pesquisa mais aprofundada (Tabela 8). Para cada rsID, buscamos caracterizar a anotação funcional mais severa (most severe consequence) e, quando disponível, o(s) gene(s) mapeado(s), integrando evidências de bases de variação (Ensembl/NCBI).

SNP	Most severe consequence	Mapped gene(s)	Referência
rs55768887	intergenic variant		Ensembl
rs55660132			
rs17717265	intron variant	USP15	NCBI
rs62522659	intron variant	CFAP418	NCBI
rs76070443	intron variant		Ensembl
rs35099227	intron variant	LNX1; LNX1-AS1	NCBI
rs73256166	intron variant		Ensembl
rs80184477	intron variant	THSD4	NCBI
rs146572111	intergenic variant		Ensembl
rs75689771	intron variant	CACNB2	NCBI
	genic upstream transcript variant		
rs73286055	intergenic variant		Ensembl
rs9519229	intergenic variant		Ensembl
rs12118675	intergenic variant		Ensembl
rs17740744	intron variant	LINC01412	NCBI
rs3821182	genic upstream transcript variant	CALCRL; CALCRL-AS1	NCBI
	intron variant		

Tabela 8 – SNPs apresentados na rede de interações e anotações externas quando disponíveis. Células em branco indicam ausência de informação recuperada nas consultas.

Em termos interpretativos, observa-se predominância de SNPs em regiões não codificantes, com anotações do tipo intron variant (isto é, localizados em íntrons) ou intergenic variant (isto é, localizados entre genes). Na prática, isso significa que, para a maior parte dos SNPs anotados, não há indicação direta de alteração em sequência de proteína; por isso, a consequência reportada por ferramentas como o Ensembl/VEP costuma aparecer com impacto do tipo modifier, isto é, uma classificação que aponta efeito previsto como pouco específico ou indeterminado apenas com base na localização genômica. Dessa forma, a tabela deve ser entendida como uma etapa de contextualização dos achados estatísticos, descrevendo onde os SNPs tendem a estar localizados (dentro de íntrons ou entre genes), e não como evidência de mecanismo funcional, e a interpretação biológica das interações detectadas exigiria estudos complementares de expressão gênica, regulação cis/trans e validação experimental.

CAPÍTULO

6

CONCLUSÃO

6.1 Conclusão

Neste trabalho investigamos interações entre SNPs associadas aos níveis de glicose, combinando modelos de regressão linear com termos de interação e o método Interaction Forest. A contribuição central desta dissertação reside na proposta e validação de uma abordagem em duas etapas para detecção de epistasia em dados genômicos de alta dimensionalidade: uma triagem não paramétrica baseada nas Florestas de Interação, seguida de validação inferencial por regressão linear com termo de interação. Diferentemente de abordagens puramente exaustivas ou puramente paramétricas, essa estratégia combina a capacidade das florestas de capturar padrões de interação sem pressupostos de linearidade com o rigor interpretativo dos modelos de regressão, resultando em um procedimento computacionalmente viável e estatisticamente fundamentado para a identificação de pares de SNPs epistáticos. A aplicação ao estudo ISA-Capital, com 324.471 SNPs e foco nos níveis de glicose sanguínea, representa, ao nosso conhecimento, uma das primeiras utilizações das Interaction Forests em dados genômicos populacionais brasileiros, reforçando a relevância da metodologia proposta para o avanço da genética estatística aplicada.

No estudo de simulação, construímos cenários com interações quantitativas e qualitativas sob diferentes intensidades de efeito. Quando os betas de interação foram mais elevados, os pares simulados foram consistentemente priorizados pelos maiores valores de EIM e apresentaram p-valores extremamente baixos para o termo de interação na regressão, evidenciando boa capacidade de recuperação de epistasia quando o sinal estatístico é suficientemente forte. Com a redução da intensidade de efeito, a detecção tornou-se mais desafiadora; ainda assim, os pares verdadeiros tenderam a permanecer entre os candidatos priorizados, indicando que as Interaction Forests operam como um mecanismo efetivo de triagem para reduzir o espaço de busca antes da validação inferencial. Adicionalmente, o tamanho amostral utilizado nas simulações foi fixado em um único valor, não sendo possível avaliar como variações no número de indivíduos impactam o desempenho do método. Ambos os pontos são reconhecidos como limitações do

estudo de simulação e representam extensões naturais para trabalhos futuros.

Na aplicação aos dados reais do estudo ISA-Capital, após controle de qualidade e filtragem por LD, analisamos 324.471 SNPs em três cenários: (i) análise global com todos os cromossomos, (ii) análise cromossomo a cromossomo e (iii) pré-seleção por regressão seguida das Florestas de Interação, em um subconjunto de 400 SNPs. Em todos os casos, utilizamos a EIM para ranquear pares SNP×SNP e, em seguida, ajustamos modelos lineares com termo de interação, priorizando os p-valores p_{int}. Embora não tenha sido observada relação linear evidente entre EIM e $-\log_{10}(p_{\text{int}})$, o Caso 3 se destacou ao produzir os menores p-valores de interação, inclusive abaixo de limiares tipicamente empregados em GWAS, sugerindo que a combinação entre pré-seleção marginal e florestas de interação pode ser particularmente eficiente em contextos de alta dimensionalidade.

Por fim, a construção de redes de interação a partir dos pares selecionados por significância do termo de interação permitiu avaliar, de forma exploratória, a existência de estruturas nos conjuntos priorizados. No entanto, tanto no cenário global quanto na análise cromossomo a cromossomo, não foi possível obter uma rede informativa após a aplicação de filtros por p_{int}: mesmo ao variar o limiar de seleção ($10^{-8}, 10^{-7}, 10^{-6}, 10^{-5}$), os pares significativos permaneceram essencialmente como arestas isoladas, sem formar componentes conectados com grau maior que 1. Em contraste, no Caso 3 (pré-seleção por regressão), a rede tornou-se visível e progressivamente mais conectada para limiares menos restritivos, com estruturas observáveis para $p_{\text{int}} < 10^{-7}$, $p_{\text{int}} < 10^{-6}$ e $p_{\text{int}} < 10^{-5}$, incluindo nós de maior grau e pequenos ciclos.

Do ponto de vista das limitações metodológicas, é importante reconhecer que a abordagem adotada neste trabalho utiliza exclusivamente a codificação aditiva (dosagem 0/1/2) para os SNPs, o que implica testar apenas a componente *linear* × *linear* da epistasia. Componentes de ordem superior, como as interações *aditivo* × *dominante* e *dominante* × *dominante*, não são capturadas por esse modelo, uma vez que o efeito de dominância, entendido como o desvio do heterozigoto em relação ao ponto médio esperado entre os homozigotos, é inteiramente absorvido pelo erro residual na codificação aditiva. A herdabilidade de fenótipos complexos relacionados ao metabolismo da glicose, como os níveis de glicemia de jejum, tem sido estimada em estudos populacionais na faixa de 40% a 60% [Dupuis *et al.* 2010, Manning *et al.* 2012], e parte substancial dessa variância permanece inexplicada por modelos puramente aditivos, sugerindo que componentes epistáticas e de dominância podem desempenhar papel relevante. A incorporação de contrastes de dominância explícitos nos modelos de regressão e nas divisões bivariadas das Florestas de Interação representa, portanto, uma extensão metodológica promissora para trabalhos futuros.

Como perspectivas de continuidade a este trabalho, pensamos em: (i) avaliar estratégias alternativas de pré-seleção e redução de dimensionalidade, incluindo métodos baseados em Random Forest e variantes para seleção de variáveis em dados genômicos, que têm sido amplamente utilizados para priorização e detecção de padrões não lineares, inclusive interações; (ii)

estender a aplicação para outras variáveis-resposta disponíveis no ISA-Capital, como desfechos relacionados a excesso de peso e obesidade; (iii) realizar, em um mesmo conjunto de dados, uma comparação sistemática entre diferentes abordagens de identificação de interações, incluindo métodos de redução de dimensionalidade multilocus como o Multifactor Dimensionality Reduction (MDR) e extensões como OR-MDR e MB-MDR; e (iv) incorporar codificação de dominância nas divisões bivariadas das Interaction Forests, possibilitando a detecção de componentes epistáticas do tipo *aditivo* $\times$ *dominante* e *dominante* $\times$ *dominante*, que permanecem invisíveis à abordagem atual.

REFERÊNCIAS

ASCHARD, H.; VILHJÁLMSSON, B. J.; GRELICHE, N.; MORANGE, P.-E.; TRÉGOUËT, D.-A.; KRAFT, P. A perspective on interaction effects in genetic association studies. **Genetic Epidemiology**, v. 40, n. 4, p. 273–281, 2016. Citado nas páginas 44 e 46.

BLUMENTHAL, D. B.; BAUMBACH, J.; HOFFMANN, M.; KACPROWSKI, T.; LIST, M. A framework for modeling epistatic interaction. **Bioinformatics**, Oxford University Press, v. 37, n. 12, p. 1708–1716, 2021. Citado nas páginas 46 e 48.

BREIMAN, L. Random forests. **Machine learning**, Springer, v. 45, p. 5–32, 2001. Citado nas páginas 27 e 52.

CALLE, M. L.; URREA, V.; MALATS, N.; STEEN, K. van. Model-based multifactor dimensionality reduction (MB-MDR). **Bioinformatics**, v. 24, n. 14, p. 1881–1887, 2008. Citado na página 26.

CHANG, C. C.; PURCELL, S. M. **Epistasis tests — PLINK 1.9 documentation**. 2025. <https://www.cog-genomics.org/plink/1.9/epistasis>. Acesso em 10 out 2025. Citado na página 26.

CHARGAFF, E. Chemical specificity of nucleic acids and mechanism of their enzymatic degradation. **Experientia**, v. 6, n. 6, p. 201–209, 1950. Citado na página 31.

CHUNG, Y.-A.; PARK, T. Odds ratio-based multifactor-dimensionality reduction method for detecting gene–gene interactions. **Bioinformatics**, v. 23, n. 1, p. 71–76, 2007. Citado na página 26.

COCKERHAM, C. C. An extension of the concept of partitioning hereditary variance for analysis of covariances among relatives when epistasis is present. **Genetics**, v. 39, n. 6, p. 859–882, 1954. Citado nas páginas 44, 45 e 46.

CORDELL, H. J. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. **Human Molecular Genetics**, v. 11, n. 20, p. 2463–2468, 2002. Citado nas páginas 27 e 45.

____. Detecting gene–gene interactions that underlie human diseases. **Nature Reviews Genetics**, v. 10, n. 6, p. 392–404, 2009. Citado na página 26.

CORDER, E. H.; SAUNDERS, A. M.; STRITTMATTER, W. J.; SCHMECHEL, D. E.; GASKELL, P. C.; SMALL, G. W.; ROSES, A. D.; HAINES, J. L.; PERICAK-VANCE, M. A. Gene dose of apolipoprotein e type 4 allele and the risk of alzheimer's disease in late onset families. **Science**, v. 261, n. 5123, p. 921–923, 1993. Citado na página 38.

DINU, I.; MAHASIRIMONGKOL, S.; LIU, Q.; YANAI, H.; ELDIN, N. S.; KREITER, E.; WU, X.; JABBARI, S.; TOKUNAGA, K.; YASUI, Y. Snp-snp interactions discovered by logic regression explain crohn's disease genetics. Public Library of Science San Francisco, USA, 2012. Citado na página 38.

DUPUIS, J.; LANGENBERG, C.; PROKOPENKO, I.; SAXENA, R. *et al.* New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. **Nature Genetics**, v. 42, n. 2, p. 105–116, 2010. Citado na página 84.

EVANS, D. M.; CORTES, A.; HU, X.; GUO, J.; AL. et. Interaction between erap1 and hla-b27 in ankylosing spondylitis implicates peptide handling by hla-b. **Nature Genetics**, v. 43, n. 8, p. 761–767, 2011. Disponível em: <https://www.nature.com/articles/ng.873>. Citado na página 47.

GRIFFITHS, A. J. F.; WESSLER, S. R.; CARROLL, S. B.; DOEBLEY, J. **Introdução à Genética**. 11. ed. Rio de Janeiro: Guanabara Koogan, 2015. Citado na página 25.

HAHN, L. W.; RITCHIE, M. D.; MOORE, J. H. Multifactor dimensionality reduction software for detecting gene–gene and gene–environment interactions. **Bioinformatics**, v. 19, n. 3, p. 376–382, 2003. Citado na página 26.

HAPFELMEIER, A.; HOTHORN, T.; ULM, K.; STROBL, C. A new variable importance measure for random forests with missing data. **Statistics and Computing**, Springer, v. 24, p. 21–34, 2014. Citado nas páginas 52 e 68.

HORNUNG, R.; BOULESTEIX, A.-L. Interaction forests: Identifying and exploiting interpretable quantitative and qualitative interaction effects. **Computational Statistics & Data Analysis**, Elsevier, v. 171, p. 107460, 2022. Includes Supplementary Material 1. Citado nas páginas 27, 48, 50, 51, 53 e 54.

JARA, L.; AMPUERO, S.; SANTIBÁÑEZ, E. *et al.* Mutations in brca1, brca2 and other breast and ovarian cancer susceptibility genes in central and south american populations. **Biological Research**, v. 50, n. 1, p. 35, 2017. Citado na página 38.

KIM, H. I.; CHO, S.; PARK, T. Ordinal multifactor dimensionality reduction method for detecting gene–gene interactions. **Bioinformatics**, v. 29, n. 18, p. 2358–2365, 2013. Citado na página 26.

LIU, Y.; ZHOU, J.; LIU, Z.; CHEN, L.; NG, M. K. Construction and analysis of genome-wide snp networks. In: **2012 IEEE 6th International Conference on Systems Biology (ISB)**. [S.l.: s.n.], 2012. p. 327–332. Citado na página 43.

MANNING, A. K.; HIVERT, M.-F.; SCOTT, R. A. *et al.* A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits. **Nature Genetics**, v. 44, n. 6, p. 659–669, 2012. Citado na página 84.

MANOLIO, T. A.; COLLINS, F. S.; COX, N. J.; GOLDSTEIN, D. B.; HINDORFF, L. A.; HUNTER, D. J.; MCCARTHY, M. I.; RAMOS, E. M.; CARDON, L. R.; CHAKRAVARTI, A. *et al.* Finding the missing heritability of complex diseases. **Nature**, v. 461, n. 7265, p. 747–753, 2009. Citado na página 45.

MAREES, A. T.; KLUIVER, H. D.; STRINGER, S.; VORSPAN, F.; CURIS, E.; MARIE-CLAIRE, C.; DERKS, E. M. A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. **International journal of methods in psychiatric research**, Wiley Online Library, v. 27, n. 2, p. e1608, 2018. Citado nas páginas 26 e 34.

MEGIORNI, F.; PIZZUTI, D. Hla-dqa1 and hla-dqb1 in celiac disease predisposition: Practical implications of the hla molecular typing. **Autoimmunity Reviews**, v. 12, n. 2, p. 197–202, 2012. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1568997212002037>. Citado na página 47.

MOORE, J. H.; WILLIAMS, S. M. Epistasis and its implications for personal genetics. **The American Journal of Human Genetics**, v. 85, n. 3, p. 309–320, 2009. Citado na página 45.

NIEL, C.; SINOQUET, C.; DINA, C.; ROCHELEAU, G. A survey about methods dedicated to epistasis detection. **Frontiers in Genetics**, v. 6, p. 285, 2015. Citado nas páginas 27 e 48.

NUSSBAUM, R. L.; MCINNES, R. R.; WILLARD, H. F. **Thompson & Thompson Genetics in Medicine**. 7. ed. Philadelphia: Saunders/Elsevier, 2007. ISBN 978-1416030805. Citado na página 25.

____. **Thompson & Thompson Genetics in Medicine**. 8. ed. Philadelphia: Elsevier, 2016. Citado na página 25.

PEREIRA, J. L.; SOUZA, C. A. de; NEYRA, J. E.; LEITE, J. M.; CERQUEIRA, A.; MINGRONI-NETTO, R. C.; SOLER, J. M.; ROGERO, M. M.; SARTI, F. M.; FISBERG, R. M. Genetic ancestry and self-reported "skin color/race" in the urban admixed population of são paulo city, brazil. **Genes**, MDPI, v. 15, n. 7, p. 917, 2024. Citado na página 56.

PHILLIPS, P. C. Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems. **Nature Reviews Genetics**, v. 9, n. 11, p. 855–867, 2008. Citado nas páginas 27 e 45.

RITCHIE, M. D.; HAHN, L. W.; ROODI, N.; BAILEY, L. R.; DUPONT, W. D.; PARL, F. F.; MOORE, J. H. Multifactor-dimensionality reduction reveals high-order interactions among genetic variants that influence disease risk. **The American Journal of Human Genetics**, v. 69, n. 1, p. 138–147, 2001. Citado na página 26.

STEPHENS, D.; FISCH, R. Bayesian analysis of quantitative trait locus data using reversible jump markov chain monte carlo. **Biometrics**, JSTOR, p. 1334–1347, 1998. Citado nas páginas 36, 40 e 56.

STRANGE, A.; CAPON, F.; SPENCER, C. C.; AL. et. A genome-wide association study identifies new psoriasis susceptibility loci and an interaction between hla-c and erap1. **Nature Genetics**, v. 42, n. 11, p. 985–990, 2010. Disponível em: <https://www.nature.com/articles/ng.694>. Citado na página 47.

STURM, R. A.; DUFFY, D. L.; MARTIN, N. G.; MONTGOMERY, G. W.; LLAMAS, L. A.; NYHOLT, L. J. R.; LI, Z.; ABECASIS, G. R.; JAMES, M. A.; HAYWARD, N. K.; KAYSER, M. A single snp in an evolutionary conserved region within intron 86 of the herc2 gene determines human blue-brown eye color. **American Journal of Human Genetics**, v. 82, n. 2, p. 424–431, 2008. Citado na página 47.

UFFELMANN, E.; HUANG, Q. Q.; MUNUNG, N. S.; VRIES, J. de; OKADA, Y.; MARTIN, A. R.; MARTIN, H. C.; LAPPALAINEN, T.; POSTHUMA, D. Genome-wide association studies. **Nature Reviews Methods Primers**, v. 1, p. 59, 2021. Citado na página 39.

VITEZICA, Z. G.; LEGARRA, A.; TORO, M. A.; VARONA, L. Orthogonal estimates of variances for additive, dominance, and epistatic effects in populations. **Genetics**, v. 206, n. 3, p. 1297–1307, 2017. Citado nas páginas 44, 45 e 46.

WAN, X.; YANG, C.; YANG, Q.; XUE, H.; FAN, X.; TANG, N.; YU, W. BOOST: A fast approach to detecting gene–gene interactions in genome-wide case–control studies. **Bioinformatics**, v. 26, n. 10, p. 1295–1301, 2010. Citado na página 26.

ZUK, O.; HECHTER, E.; SUNYAEV, S. R.; LANDER, E. S. The mystery of missing heritability: Genetic interactions create phantom heritability. **Proceedings of the National Academy of Sciences of the United States of America**, v. 109, n. 4, p. 1193–1198, 2012. Citado na página 26.

APÊNDICE

A

TABELAS DOS 20 MENORES P-VALORES

snp1	snp2	eim_value	coef_snp1	pval_snp1	coef_snp2	pval_snp2	coef_interacao	pval_interacao
rs6457131_T	rs3131787_C	1.08e-05	0.0371699	8.08e-03	0.1122360	2.91e-07	-0.1587490	4.76e-08
rs6457131_T	rs4678_A	8.40e-06	0.0369294	8.64e-03	0.1101974	4.43e-07	-0.1561574	7.25e-08
rs6457131_T	rs4678_A	6.41e-06	0.0369294	8.64e-03	0.1101974	4.43e-07	-0.1561574	7.25e-08
rs9947371_C	rs56374046_A	7.09e-06	-0.0572183	7.34e-04	-0.0444386	2.08e-03	0.1078512	3.03e-07
rs78380435_A	rs4821513_C	1.71e-05	0.1780199	4.33e-08	0.0133702	2.54e-01	-0.1740723	3.96e-07
rs2575174_T	rs13195572_C	2.63e-05	0.0366364	2.51e-03	0.0995365	4.20e-05	-0.1031389	7.75e-07
rs2575174_T	rs13195572_C	8.15e-06	0.0366364	2.51e-03	0.0995365	4.20e-05	-0.1031389	7.75e-07
rs6457131_T	rs3095151_T	1.23e-05	0.0375797	9.27e-03	0.0794036	3.75e-05	-0.1258986	8.60e-07
rs4678544_C	rs6444305_G	2.25e-05	-0.0550421	1.48e-03	-0.0312808	3.86e-02	0.0842932	1.02e-06
rs4678544_C	rs6444304_C	2.88e-05	-0.0551569	1.47e-03	-0.0317699	3.85e-02	0.0846518	1.08e-06
rs1507825_A	rs9834871_T	2.43e-05	-0.0434048	1.51e-03	-0.0747642	1.41e-04	0.0888796	1.48e-06
rs17207196_T	rs879003_T	6.21e-06	-0.0245267	4.90e-02	-0.0585755	8.62e-03	0.0983598	1.64e-06
rs62294120_G	rs6832816_G	2.85e-05	-0.0143038	4.69e-01	-0.0357005	6.51e-02	0.2030528	1.76e-06
rs388367_C	rs6708834_A	2.27e-05	-0.0215904	9.07e-02	-0.0620065	1.68e-03	0.0802958	1.84e-06
rs9648343_C	rs7804456_C	3.87e-05	-0.0229436	1.13e-01	-0.0518995	8.78e-04	0.0830852	3.77e-06
rs7374258_A	rs11720703_T	9.03e-06	-0.0647026	2.75e-05	-0.0706220	1.65e-04	0.0702154	4.15e-06
rs17290107_C	rs34363522_AA	2.64e-05	0.0833120	4.51e-05	0.0168993	2.15e-01	-0.1005620	4.36e-06
rs3760423_G	rs72847801_A	7.39e-06	-0.0286570	4.70e-02	-0.0234341	2.08e-01	0.1219388	5.18e-06
rs13054962_G	rs9606709_C	3.40e-05	-0.0145094	4.13e-01	-0.0329023	3.16e-02	0.1112418	5.49e-06
rs41325747_C	rs943206_T	1.96e-05	-0.0560960	1.01e-03	-0.0401358	9.21e-03	0.0683579	6.67e-06

Tabela 9 – Top 20 menores p-valores do termo de interação (Análise Individual).

snp1	snp2	coef_snp1	pval_snp1	coef_snp2	pval_snp2	coef_interacao	pval_interacao
rs6800849_A	rs11007430_A	-0.0482216	2.46e-02	-0.0196231	1.22e-01	0.1043843	1.74e-07
rs411337_T	rs8027766_C	0.0817586	3.00e-06	0.0470140	6.23e-03	-0.1214829	4.11e-06
rs411337_T	rs8027766_C	0.0817586	3.00e-06	0.0470140	6.23e-03	-0.1214829	4.11e-06
rs1002174_G	rs12917291_A	-0.0037859	7.93e-01	-0.0414024	1.39e-02	0.1005886	2.95e-05
rs7578812_T	rs740420_A	0.0710932	1.85e-06	0.0249232	1.30e-01	-0.0725253	3.18e-05
rs11896845_T	rs6048546_A	-0.0181228	1.94e-01	-0.0139615	4.62e-01	0.0844992	4.51e-05
rs10973634_C	rs60628344_G	-0.0368851	2.57e-02	-0.0408281	1.28e-02	0.0568542	5.44e-05
rs2148174_T	rs9387601_G	0.0723615	1.35e-04	0.0234473	9.14e-02	-0.0729410	5.82e-05
rs6444652_T	rs214711_C	-0.0351882	4.93e-02	0.0087193	5.39e-01	0.0889513	7.97e-05
rs17804312_G	rs581534_C	-0.0047158	6.62e-01	0.1694824	2.74e-06	-0.1207597	1.01e-04
rs61410289_G	rs10402887_G	-0.0162705	3.25e-01	-0.0038124	8.03e-01	0.0826664	1.15e-04
rs4745380_C	rs12901901_C	0.0852020	1.33e-05	0.0110204	4.11e-01	-0.0634601	1.19e-04
rs2704112_A	rs12326011_C	0.0162241	1.92e-01	-0.0340756	1.52e-01	0.0915405	2.03e-04
rs2109517_G	rs444804_T	0.0335771	2.55e-02	0.0704955	9.02e-05	-0.0575573	2.16e-04
rs1935743_T	rs7081208_A	-0.0076524	5.89e-01	-0.0204078	3.11e-01	0.0897589	2.52e-04
rs17623453_G	rs62468927_G	-0.0292300	3.59e-01	-0.0190334	1.86e-01	0.1412837	2.54e-04
rs5894958_CAGA	rs12608173_G	-0.0042034	7.50e-01	-0.0312711	1.44e-01	0.0865041	2.69e-04
rs11894928_G	rs1060570_A	0.0302829	6.74e-02	0.0746176	4.00e-06	-0.0560310	2.78e-04
rs489077_C	rs5759635_A	-0.0631726	2.25e-03	-0.0357654	8.76e-03	0.0887380	2.89e-04
rs9638734_T	rs34567980_A	-0.0130648	4.10e-01	0.0020601	9.24e-01	0.1109887	3.17e-04

Tabela 10 – Top 20 menores p-valores do termo de interação (Análise Geral).

snp1	snp2	coef_snp1	pval_snp1	coef_snp2	pval_snp2	coef_interacao	pval_interacao
rs55768887_T	rs73256166_A	0.0050677	8.33e-01	0.0080426	6.75e-01	0.4253349	5.81e-16
rs73286055_A	rs75689771_G	0.0171030	3.94e-01	0.0124439	4.18e-01	0.2606503	4.79e-15
rs29623_A	rs4432172_A	0.0117901	5.88e-01	0.0047521	8.09e-01	0.3273849	3.32e-14
medDM_bioq	rs13127462_A	0.5267797	5.35e-38	-0.0089331	4.15e-01	-0.2476684	3.09e-13
rs4704396_C	rs17179053_C	0.0096938	6.24e-01	-0.0102908	6.25e-01	0.3104587	3.46e-13
medDM_bioq	rs2875973_T	0.0800241	4.14e-02	0.0101627	3.36e-01	0.2295346	2.47e-12
rs73552677_T	rs12891831_C	0.0239788	1.78e-01	-0.0004883	9.84e-01	0.2636929	2.97e-12
rs76070443_G	rs1489798_C	0.0214880	1.38e-01	0.0098625	6.45e-01	0.2634664	7.58e-12
rs73669153_A	rs62279925_T	0.0306359	1.78e-01	0.0253097	2.91e-01	0.4958091	1.11e-11
rs76070443_G	rs62522659_A	0.0037007	8.12e-01	-0.0029246	8.58e-01	0.1722648	1.56e-11
rs55660132_G	rs10866415_C	0.1477071	2.60e-14	0.0245477	8.85e-02	-0.1076966	2.77e-11
medDM_bioq	rs13025591_C	0.4678898	3.14e-36	-0.0177346	1.11e-01	-0.2515731	3.73e-11
medDM_bioq	rs632957_G	0.1254970	4.67e-04	0.0126039	2.65e-01	0.2218603	6.82e-11
rs55768887_T	rs55660132_G	-0.0146407	5.98e-01	0.0196469	8.67e-02	0.2188330	1.06e-10
rs55660132_G	rs7554322_T	0.0082140	4.96e-01	-0.0467726	5.57e-02	0.1676807	1.47e-10
rs55660132_G	rs17717265_G	0.0079414	5.11e-01	-0.0408328	6.10e-02	0.1548404	2.19e-10
rs17717265_G	rs75795339_G	0.0196636	2.47e-01	0.0089827	6.80e-01	0.2886319	3.79e-10
rs146572111_C	rs75689771_G	0.0203782	3.97e-01	0.0196833	2.10e-01	0.2760568	3.93e-10
rs62522659_A	rs17717265_G	-0.0038481	8.17e-01	-0.0135714	4.76e-01	0.1786968	4.27e-10
rs4432172_A	rs6432182_T	0.0221205	2.67e-01	0.0194896	3.09e-01	0.2842523	6.86e-10

Tabela 11 – Top 20 menores p-valores do termo de interação (Análise de Regressão).


ICMC
USP
SÃO CARLOS