Universidade Federal de São Carlos
Centro de Ciências Exatas e de Tecnologia
Departamento de Estatística

Estudo de Associação Genômica Ampla para Obesidade por meio de Florestas Aleatórias

Milena Crnkovic Luzia

Trabalho de Conclusão de Curso

Universidade Federal de São Carlos
Centro de Ciências Exatas e de Tecnologia
Departamento de Estatística

Estudo de Associação Genômica Ampla para Obesidade por meio de Florestas Aleatórias

Milena Crnkovic Luzia
Orientadora: Profa. Dra. Andressa Cerqueira

Trabalho de Conclusão de Curso apresentado como parte dos requisitos para obtenção do título de Bacharel em Estatística.

São Carlos
Fevereiro de 2025

Federal University of São Carlos
Exact and Technology Sciences Center
Department of Statistics

Genome-Wide Association Study for Obesity through Random Forests

Milena Crnkovic Luzia
Advisor: Prof. Dr. Andressa Cerqueira

Bachelors dissertation submitted to the Department of Statistics, Federal University of São Carlos - DEs-UFSCar, in partial fulfillment of the requirements for the degree of Bachelor in Statistics.

São Carlos
February 2025

Milena Crnkovic Luzia

Estudo de Associação Genômica Ampla para Obesidade por meio de Florestas Aleatórias

Este exemplar corresponde à redação final do trabalho de conclusão de curso devidamente corrigido e defendido por nome do(a) aluno(a) e aprovado pela banca examinadora.

Aprovado em 13 de fevereiro de 2025

Banca Examinadora:

- Profa. Dra. Andressa Cerqueira
- Profa. Dra. Daiane Aparecida Zuanetti
- Prof. Dr. Thiago Rodrigo Ramos

Resumo

A obesidade é resultado de um desequilíbrio energético e vem crescendo drasticamente, porém não são todos os indivíduos que respondem da mesma forma ao mesmo ambiente, o que pode ser resultado de fatores genéticos. Segundo (29), a herdabilidade pode contribuir com cerca de 40% para a variação de uma doença. Dessa forma, noções de genética são de extrema importância para uma melhor compreensão de doenças crônicas como a obesidade. A maioria dos casos de obesidade é causada pela combinação de diversas variantes genéticas de pequeno efeito, assim existem vários desafios metodológicos e computacionais. Os métodos estatísticos de Estudo de Associação Genômica Ampla (GWAS) usualmente utilizados para análises são univariados, considerando apenas um marcador genético (SNP) por vez, logo tais métodos não capturam relações entre SNPs. O método de florestas aleatórias tem sido largamente aplicado em dados biológicos, sendo um método que suporta dados de alta dimensionalidade e pode capturar modelos não-lineares baseados em genótipos. Dessa forma, o presente trabalho pretende identificar variantes genéticas associadas à obesidade de forma multifatorial através de florestas aleatórias. Utilizando a medida de importância das florestas aleatórias e selecionando o quantil 99% de sua distribuição, os resultados identificaram diversos SNPs associados à obesidade e à circunferência da cintura, reforçando a complexidade genética dessas características. Um dos marcadores encontrados foi o SNP rs478582, presente em ambas as análises, sugerindo alta influência na regulação da adiposidade e na distribuição de gordura corporal. Muitos dos genes identificados no estudo estão envolvidos em processos metabólicos e inflamatórios, destacando a importância de fatores genéticos na predisposição à obesidade e reforçando a necessidade de estudos adicionais para aprofundar a compreensão dos mecanismos biológicos relacionados à obesidade.

Palavras-chave: *associação, obesidade, florestas aleatórias, GWAS, SNP.*

Abstract

The obesity is the result of an energy imbalance and has been growing significantly, but not all individuals respond in the same way to the same environment, which may be the result of genetic factors. According to (29), heritability can contribute around 40% to the variation of a disease. Therefore, notions of genetics are extremely important for a better understanding of chronic diseases such as obesity. Most cases of obesity are caused by the combination of several genetic variants of small effect, and there are several methodological and computational challenges. The statistical methods for Genomic-Wide Association Study (GWAS) normally used for the analyzes are univariate, considering only one genetic marker (SNP) at a time, therefore such methods do not capture relationships between SNPs. The random forests method has been widely applied to biological data, being a method that supports high-dimensional data and can capture non-linear models based on genotypes. Therefore, the present work aims to identify genetic variants associated with obesity in a multifactorial way through random forests. Using the importance measure of random forests and selecting the 99% quantile of its distribution, the results identified several SNPs associated with obesity and waist circumference, reinforcing the genetic complexity of these characteristics. One of the markers found was the SNP rs478582, present in both analyses, suggesting a high influence on the regulation of adiposity and body fat distribution. Many of the genes identified in the study are involved in metabolic and inflammatory processes, highlighting the importance of genetic factors in the predisposition to obesity and reinforcing the need for additional studies to deepen the understanding of the biological mechanisms related to obesity.

Keywords: *association, obesity, GWAS, random forests, SNP.*

Lista de Figuras

2.1 Estrutura da molécula de DNA (2). 20
2.2 Efeito de troca de bases nitrogenadas definida como SNP (23). 21
2.3 Representação da separação da amostra em casos e controle para condução do GWAS. Adaptado de (12). 22
2.4 Gráfico da distribuição dos marcadores SNPs de acordo com seu MAF e efeito genético, separados de variantes comuns a raras. Adaptado de (29). . 23

3.1 Gráfico de barras da obesidade por sexo. 33
3.2 Gráfico de barras da obesidade por faixa etária. 34
3.3 Correlação linear entre as variáveis quantitativas de interesse. 35
3.4 Gráfico da distribuição da circunferência de cintura por faixa etária. 36
3.5 Gráfico da distribuição da circunferência de cintura por sexo. 37
3.6 Gráfico de barras da quantidade de SNPs por cromossomo antes e após a aplicação dos filtros mencionados. 39

4.1 Exemplo de uma árvore de decisão. 42
4.2 Exemplo de uma árvore de decisão. 43
4.3 Exemplo de uma árvore de decisão. 44

5.1 Gráfico Manhattan para a variável obesidade. 53
5.2 Gráfico Manhattan para a variável circunferência de cintura abdominal. . . 56

Lista de Tabelas

2.1 Tabela dos valores observados na amostra. 27
2.2 Tabela do cálculo das observações esperadas sob H_0. 27

3.1 Tabela de frequência para a variável sexo. 32
3.2 Tabela de frequência para a variável faixa etária. 32
3.3 Tabela de frequência para a variável obesidade. 33
3.4 Tabela combinada dos testes qui-quadrado para obesidade. 34
3.5 Tabela de Medidas Resumo. 35

5.1 Medidas Resumo para os SNPs selecionados no quantil 99% para obesidade. 53
5.2 Quantidade de artigos encontrados para cada SNP, cromossomo, importância e descrição associada. 54
5.3 Medidas Resumo para os SNPs selecionados no quantil 99% para circunferência de cintura abdominal. 56
5.4 Quantidade de artigos encontrados para cada SNP, cromossomo, importância e descrição associada. 57

Sumário

1 Introdução **17**

1.1 Objetivos 18

2 Estudos de Associação Genômica Ampla - GWAS **19**

2.1 Funcionamento do Genoma Humano 19

2.2 Apresentação do GWAS 22

2.3 Filtros Estatísticos 23

2.3.1 Frequência do Menor Alelo - MAF 24

2.3.2 Desequilíbrio de Hardy-Weinberg 24

2.3.3 Desequilíbrio de Ligação 25

3 Banco de Dados **29**

3.1 Banco de Dados ISA 2015 29

3.2 Análise Exploratória 30

3.2.1 Resumo da Análise Exploratória 37

3.3 Aplicação ao Banco de Dados 38

3.3.1 Obtenção das Componentes de Ancestralidade Global 39

4 Florestas Aleatórias **41**

4.1 Árvores de Decisão 41

4.2 Florestas Aleatórias 45

4.3 Florestas Aletórias e GWAS 47

5 Resultados **49**

5.1 Imputação dos genótipos faltantes 49

5.2 Aplicação da floresta aleatória 51

5.3 Aplicação da floresta para obesidade 52

5.4 Aplicação da floresta para circunferência de cintura 56

6 Conclusão **59**

6.1 Discussão e conclusão 59

Referências Bibliográficas **61**

Capítulo 1

Introdução

A obesidade é resultado de desequilíbrio energético quando um indivíduo consome mais calorias do que seu corpo queima, se associando com doenças como diabetes, hipertensão, dislipidemia, entre outras. Atualmente a obesidade vem crescendo drasticamente, com a diminuição da prática de exercícios físicos e aumento de consumo de alimentos calóricos. Porém, não são todos os indivíduos que respondem da mesma forma ao mesmo ambiente, o que pode ser resultado de fatores genéticos. Segundo (29), estudos com gêmeos podem estimar a herdabilidade de uma doença comum em 40%, ou seja, 40% da variância total no risco de doença é devido a fatores genéticos. Dessa forma, noções de genética são de extrema importância para uma melhor compreensão de doenças crônicas como a obesidade.

De acordo com as informações descritas em (6), a maioria dos casos de obesidade é causada pela combinação de diversas variantes genéticas de pequeno efeito e fatores ambientais, conhecida como obesidade multifatorial. No entanto, em alguns pacientes, a doença está relacionada com mutações de grande efeito em um único gene, conhecida como obesidade monogênica. Raramente, porém, há um padrão claro de obesidade hereditária causado por uma variante específica de um único gene, sendo a maior parte proveniente de interações entre múltiplos genes e fatores ambientais, que permanecem pouco compreendidos.

Existem vários desafios metodológicos e computacionais para determinar qual variante influencia na susceptibilidade à doença. Os métodos estatísticos para o Estudo de Associação Genômica Ampla (GWAS) usualmente utilizados nas análises são univariados, isto é, são métodos que consideram apenas um marcador genético por vez. Logo, tais métodos não capturam relações entre marcadores. Portanto, estudos como GWAS

por meio de técnicas de regressão linear, por exemplo, apenas capturam relações lineares entre um único marcador e o fenótipo em estudo e não são capazes de captar as interações existentes entre os marcadores genéticos.

O método de florestas aleatórias (do inglês *random forest*, RF) tem sido largamente aplicado em dados biológicos, sendo um método que suporta dados de alta dimensionalidade e pode capturar relações não-lineares baseados em genótipos (7).

Com os resultados dos marcadores genéticos selecionados por métodos de GWAS podemos identificar os marcadores associados com a obesidade, a fim de direcionar melhores tratamentos e acompanhamento dos indivíduos. Dessa forma, o presente trabalho pretende identificar variantes genéticas associadas à obesidade de forma multifatorial através de florestas aleatórias.

1.1 Objetivos

O objetivo desse projeto é a identificação dos marcadores genéticos associados a obesidade. Além disso, esse projeto busca o estudo de florestas aleatórias, a fim de verificar sua aplicabilidade em estudos de GWAS.

Capítulo 2

Estudos de Associação Genômica Ampla - GWAS

O estudo de associação genômica ampla (GWAS) é um método de identificação de variantes genéticas associadas com fenótipos de interesse, com ampla aplicação no controle de doenças ou predição de fatores de risco para um indivíduo. Neste capítulo serão apresentados conceitos genéticos e características do GWAS.

2.1 Funcionamento do Genoma Humano

Todas as informações para funcionamento do organismo dos seres humanos, bem como de seu desenvolvimento se concentram no genoma. O genoma é formado pelo conjunto de todo DNA presente no organismo, que se divide em 23 pares de cromossomos, sendo 22 pares de cromossomos autossômicos e 1 par de cromossomos sexuais. DNA é a sigla em inglês para ácido desoxirribonucleico e é uma molécula formada por ligações de açúcar (desoxirribose) e fosfato. Em cada ligação de açúcar há uma das quatro bases nitrogenadas, sendo elas adenina (A), citosina (C), guanina (G) ou timina (T). As duas fitas de açúcar do DNA são conectadas pelas ligações entre as bases: a adenina se liga à timina e a citosina se liga à guanina. A Figura 2.1 demonstra essa estrutura.

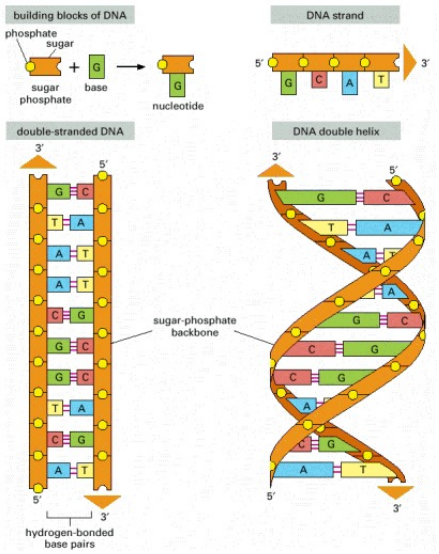

Figura 2.1: Estrutura da molécula de DNA (2).

A sequência das bases ao longo do DNA codifica informações biológicas, como produção de proteínas ou RNA. Cada particionamento dessa sequência dentro dos cromossomos é caracterizado como gene, que é responsável pelas características herdadas geneticamente. Dessa forma, um gene é uma subsequência do genoma composta, muitas vezes, por centenas e milhares de posições (locus) do DNA

As regiões podem sofrer mudanças em sua sequência, levando a diferentes variantes, conhecidas como alelos, cada lócus contém um alelo de um gene. Esses alelos codificam versões ligeiramente diferentes de uma proteína, o que causa diferentes características fenotípicas que observamos na população. Comumente cada gene terá dois alelos, ocupando o mesmo lócus em cada um dos pares do cromossomo.

Cerca de 99,9% de nosso DNA é idêntico, responsável por nosso desenvolvimento. Os diferentes 0,1% contêm as variações que combinadas a fatores ambientais propiciam nossa singularidade (24).

Quando nessa sequência de bases ocorre a alteração de uma única base nitrogenada por outra é definido biologicamente por SNP, sigla em inglês para polimorfismo de nucleotídeo único. Isto é, se por exemplo em uma posição de base específica o nucleotídeo C aparece na maioria dos indivíduos, mas em uma minoria a posição é ocupada por um A, isso indica um SNP nesta posição específica. As duas variações possíveis — C ou A — são chamadas de alelos para esta posição.

Os SNPs são distinguidos das variações raras por possuírem a frequência do menor alelo maior que 1% (3). Isto é, se mais de 1% da população analisada possui tal alteração elas são chamadas SNPs, caso seja menor que 1% é caracterizada como uma mutação. A

Figura 2.2 demonstra a ocorrência de um SNP na molécula de DNA.

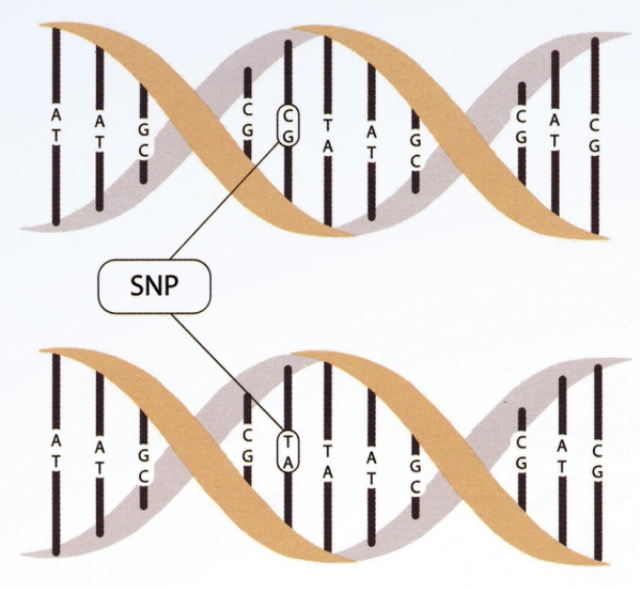

Figura 2.2: Efeito de troca de bases nitrogenadas definida como SNP (23).

Esses SNPs podem ser considerados marcadores genéticos e estima-se que ocorram em 1 a cada 1000 pares de bases nitrogenadas na sequência de DNA, porém como grande parte do código genético não é codificada muitas variações podem não causar efeitos no organismo (9).

Os SNPs são variações genéticas comuns e normalmente possuem dois alelos (A,a), o que significa que dentro de uma população existem duas possibilidades de pares de bases que ocorrem comumente para uma localização de SNP. A frequência de um SNP é dada em termos da frequência do alelo menor ou frequência do alelo menos comum.

Os SNPs podem ser codificados em 0, 1 e 2, representando diferentes genótipos de uma determinada posição genética. O valor 0 representa o genótipo homozigoto para o alelo de dominância, o valor 1 representa o genótipo heterozigoto (um alelo de dominância e um alelo recessivo) e o valor 2 representa genótipo homozigoto para o alelo recessivo. O conjunto de dados analisado seguirá essa codificação.

Ao representar os genótipos por 0, 1 e 2 atribuí-se maior peso aos indivíduos que possuem alelos recessivos, de forma a entender a relação entre o número de alelos recessivos e a severidade da doença.

2.2 Apresentação do GWAS

O Estudo de Associação Genômica Ampla (GWAS) é um método de identificação de variantes genéticas associadas com fenótipos de interesse. As variações genéticas são causadas por mutações na sua grande maioria durante a replicação do material genético e são responsáveis pelo aparecimento de doenças mas também são benéficas para evolução da espécie.

As variações com frequência maior que 1% na população são os denominados SNPs e para identificar contribuições genéticas para uma determinada doença podemos agrupar indivíduos portadores da doença e comparar seus padrões de SNP com indivíduos não portadores, como representado pela Figura 2.3. Dessa forma utilizamos os SNPs como marcadores genéticos para identificar as áreas do genoma influentes na doença.

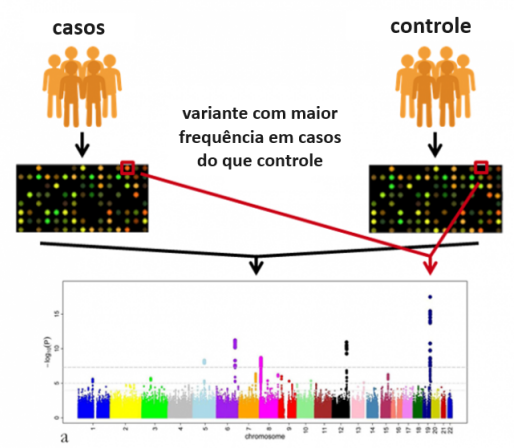

Figura 2.3: Representação da separação da amostra em casos e controle para condução do GWAS. Adaptado de (12).

O estudo do GWAS captura a variação genética existente em todo o genoma humano, avaliando de forma estatística esses marcadores a fim de verificar as associações com o fenótipo de interesse. Os resultados do GWAS podem ser usados em várias aplicações como no controle de doenças ou predição de fatores de risco para um indivíduo.

A identificação dos efeitos genéticos para uma doença pode não ser uma tarefa fácil. Estudos mostram que doenças comuns possuem múltiplos alelos comuns que influenciam a susceptibilidade à doença. Alelos comuns têm pequenos efeitos genéticos e portanto o efeito se deve a múltiplos fatores genéticos, o que dificulta a identificação por meio do GWAS (29).

A Figrura 2.4 demonstra como se distribuem as frequências e efeitos de determinados marcadores. Os efeitos genéticos no canto superior direito são mais receptivos a estu-

dos menores e portanto são utilizados poucos marcadores genéticos. Os efeitos no canto inferior direito são típicos dos resultados do GWAS, exigindo grandes tamanhos de amostra e uma considerável quantidade de marcadores. Os efeitos da parte central são mais difíceis de serem obtidos que os anteriores e normalmente combinam outras técnicas além do GWAS. Os efeitos no canto inferior esquerdo são os mais difíceis de obter, exigindo sequenciamento genômico de grandes amostras para associar variantes raras à doença.

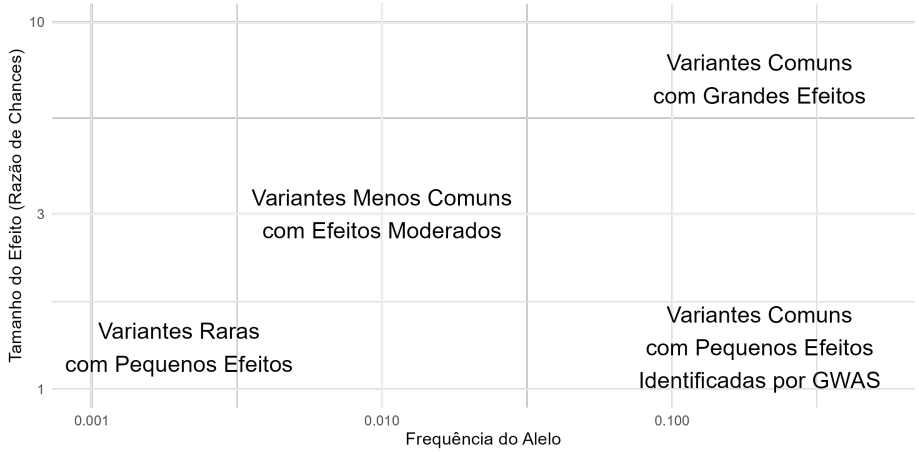

Figura 2.4: Gráfico da distribuição dos marcadores SNPs de acordo com seu MAF e efeito genético, separados de variantes comuns a raras. Adaptado de (29).

Dessa forma, a fim de avaliar da melhor forma possível os resultados do GWAS, filtros devem ser aplicados para controle desses SNPs. Os SNPs de ocorrência rara não serão captados pelo GWAS e portanto devem ser excluídos. As correlações entre os marcadores também devem ser determinadas para não incorporar informações redundantes. O método estatístico para condução do GWAS também deve ser escolhido estrategicamente para obtenção do maior número de SNPs possível.

2.3 Filtros Estatísticos

Antes da aplicação do GWAS, como mencionado, é necessária uma análise de controle de qualidade, onde são removidos das análises SNPs para os quais os genótipos não foram sequenciados, bem como SNPs que possuem correlação entre si ou baixa frequência observada na população, visando o mínimo de perda de informação.

Os métodos a seguir são aplicados na condução do GWAS como controle de qualidade para filtrar os SNPs.

2.3.1 Frequência do Menor Alelo - MAF

A frequência do menor alelo (do inglês *Minor Allele Frequency*, MAF) indica a frequência do alelo menos comum em uma população. Dado que o alelo de menor frequência em uma população seja o recessivo (a), podemos calcular sua frequência por

$$MAF = \frac{n_a}{2n},$$

onde n é o tamanho da amostra e n_a é a quantidade de alelos recessivos a na amostra. Como cada indivíduo de uma amostra de tamanho n possui dois alelos para formação do genótipo, dividimos a quantidade n_a por $2n$.

Essa frequência ajuda a medir a variabilidade da amostra, pois pela frequência consegue-se diferenciar SNPs comuns de raros. Assim, ao utilizar MAF $\geq 5\%$ são retirados os SNPs raros que não serão identificados pelo GWAS.

2.3.2 Desequilíbrio de Hardy-Weinberg

O Equilíbrio de Hardy-Weinberg (EHW) é um princípio que afirma que a variação genética em uma população permanecerá constante ao longo das gerações na ausência de fatores perturbadores. Portanto, sob certas condições, as frequências de alelo na população permanecerão constantes ao longo do tempo. O desvio desse equilíbrio pode ser indicativo de potenciais erros de genotipagem ou estratificação populacional.

As frequências genotípicas, considerando independência entre o alelo paterno e materno, podem ser expressas por $p^2 + 2pq + q^2 = 1$, onde

p = frequência do alelo dominante (A);

q = frequência do alelo recessivo (a);

p^2 = frequência do genótipo homozigoto dominante (AA);

$2pq$ = frequência do genótipo heterozigoto (Aa);

q^2 = frequência do genótipo homozigoto recessivo (aa).

Assim, na população um individuo pode apresentar dois alelos dominantes (AA), um dominante e um recessivo (Aa ou aA) ou dois recessivos (aa), por isso escrevemos $p^2 + 2pq + q^2 = 1$.

Para testar se o SNP está em Equilíbrio de Hardy-Weinberg, é feito um teste de hipóteses utilizando o teste qui-quadrado, com a hipótese nula de que o SNP está em

equilíbrio. Ou seja,

$$\begin{cases} H_0 : \text{os genótipos estão em EHW.} \\ H_1 : \text{os genótipos não estão em EHW.} \end{cases}$$

Denotamos as frequências observadas e esperadas dos genótipos sob H_0 por o_i e e_i, respectivamente, para $i \in \{AA, Aa, aa\}$. A estatística teste é

$$\tilde{\chi}^2 = \sum_{i \in \{AA,Aa,aa\}} \frac{(o_i - e_i)^2}{e_i},$$

onde $\sum_{i \in \{AA,Aa,aa\}} o_i = n$ e $e_i = nP_i$, em que $P_{AA} = p^2$, $P_{Aa} = 2pq$ e $P_{aa} = q^2$, sob H_0 e n é o tamanho da amostra. Sob H_0 $\tilde{\chi}^2 \sim \chi^2_{(1)}$. O teste é feito sob as suposições de que a amostra é aleatória e de que a unidade de medida da variável é no mínimo nominal, além disso é um teste aproximado, que possui melhor desempenho sob grandes amostras.

Os SNPs em desequilíbrio não estão em conformidade no genótipo e portanto, removeremos os SNPs cujo o valor-p do teste (valor-p = $P(\tilde{\chi}^2 \geq \tilde{\chi}^2_{obs})$) seja menor que o nível de significância desejado, pois para valores-p menores que o nível de significância devemos rejeitar a hipótese nula.

2.3.3 Desequilíbrio de Ligação

O Desequilíbrio de Ligação (do inglês *Linkage Desequilibrium*, LD) é a associação não aleatória de dois ou mais alelos em loci diferentes que ocorrem com probabilidade diferente da esperada sob aleatoriedade.

A presença de LD possibilita a identificação de dois possíveis SNPs estatisticamente associados ao fenótipo. Num primeiro caso, o SNP que influencia o fenótipo é considerado estatisticamente associado à característica e possui uma associação direta. Já para o segundo caso, esse SNP pode não ser captado diretamente, em vez disso um outro SNP, em LD com ele, é captado como estatisticamente associado ao fenótipo, caracterizando uma associação indireta. Assim, é difícil dizer qual SNP exatamente está influenciando a doença e todos os SNPs em LD são potenciais candidatos. Desse modo, o filtro de LD no GWAS como controle de qualidade otimiza o processo, evitando a análise de SNPs que fornecem informações redundantes.

Definição 2.1 (Desequilíbrio de Ligação) *Considere dois loci ligados, o Locus 1 pos-*

sui um alelo A ocorrendo na frequência p_A e Locus 2 um alelo B ocorrendo na frequência p_B na população. O haplótipo pode ser denotado como AB com frequência $p_A p_B$. Diz-se que os dois loci ligados estão em equilíbrio de ligação (LE), se a ocorrência do alelo A e a ocorrência do alelo B no haplótipo forem eventos independentes. Por outro lado, os alelos estão em desequilíbrio de ligação (LD) quando não ocorrem aleatoriamente (4).

Sob desequilíbrio de ligação, a frequência do alelo A na primeira região não é independente da frequência do alelo B na segunda região. Matematicamente podemos definir o desequilíbrio de ligação D_{AB} para os alelos A e B nas duas regiões comparadas como

$$D_{AB} = p_{AB} - p_A p_B.$$

Quando $D_{AB} = 0$ temos que $p_{AB} = p_A p_B$, ou seja, as duas regiões estão em equilíbrio de ligação pois a frequência do alelo A na primeira região é independente da frequência do alelo B na segunda região.

O desequilíbrio de ligação positivo existe quando os alelos ocorrem com mais frequência do que o esperado e o negativo quando ocorrem com menos frequência do que o esperado, sob a independência.

Regiões genéticas próximas no DNA têm uma probabilidade maior de estarem em desequilíbrio de ligação porque são herdadas em conjunto e raramente são separadas por eventos de recombinação. Regiões genéticas de cromossomos diferentes são consideradas independentes.

Para avaliar o desequilíbrio um teste qui-quadrado pode ser realizado, sob as seguintes hipóteses:

$$\begin{cases} H_0: & \text{os loci estão em equilíbrio } (D_{AB} = 0). \\ H_1: & \text{os loci não estão em equilíbrio } (D_{AB} \neq 0). \end{cases}$$

Sob H_0, $p_{ij} = p_i p_j$, para $i = A, a$ e $j = B, b$.

A Tabela 2.1 de contingência pode ser construída utilizando as quantidades de cada genótipo observadas na amostra e utilizada para aplicação do teste.

	B	b	Total
A	n_{AB}	n_{Ab}	n_A
a	n_{aB}	n_{ab}	n_a
Total	n_B	n_b	$2n$

Tabela 2.1: Tabela dos valores observados na amostra.

Já sob H_0, o número de observações esperada de cada genótipo pode ser descrito pela Tabela 2.2.

	B	b
A	$2np_Ap_B$	$2np_Ap_b$
a	$2np_ap_B$	$2np_ap_b$

Tabela 2.2: Tabela do cálculo das observações esperadas sob H_0.

A estatística de teste Q, pode ser calculada como

$$Q = \sum_{i=A}^{a} \sum_{j=B}^{b} \frac{(o_{ij} - e_{ij})^2}{e_{ij}},$$

sendo e_{ij} o número de obervações esperadas sob a independência do haplótipo ij e o_{ij} o número observado, dados por $o_{ij} = n_{ij}$ e $e_{ij} = 2np_ip_j$. Sob H_0, $Q \sim \chi^2_1$.

Dessa forma será removido o SNP com menor valor-p dentre os dois analisados, cujo o valor-p do teste (valor-p $= P(Q \geq Q_{obs})$) seja menor que o nível de significância desejado.

Capítulo 3

Banco de Dados

Para este trabalho, estudaremos a prevalência de obesidade na população e a relação da obesidade com as características genotípicas dos indivíduos. Neste capítulo, apresentaremos as variáveis presentes no banco de dados e realizaremos uma análise descritiva das mesmas. Além disso, aplicaremos os filtros apresentados na Seção 2.3 ao banco de dados genotípicos.

3.1 Banco de Dados ISA 2015

Os dados utilizados nesse projeto fazem parte do projeto temático (17/05125-7) entitulado "Estilo de vida, marcadores bioquímicos e genéticos como fatores de risco cardiometabólico: Inquérito de Saúde na cidade de São Paulo". O Inquérito de Saúde de São Paulo (ISA-Capital 2015) - Estudo Foco em Nutrição (ISA-Nutrição 2015) é um inquérito transversal realizado de fevereiro de 2015 a fevereiro de 2016 que utilizou uma amostra multiestágio, estratificada por conglomerados, com seleção de setores censitários urbanos e domicílios, fornecendo estimativas representativas da população da cidade de São Paulo.

A estratificação ocorreu nas cinco Coordenadorias de Saúde de São Paulo: Norte, Centro-Oeste, Sudeste, Sul e Leste, que foram os domínios do estudo. No primeiro estágio amostral, foram selecionados aleatoriamente 30 setores censitários urbanos de cada área, totalizando 150 unidades. No segundo estágio, foram selecionados em média 18 domicílios particulares em cada setor censitário. Todos os indivíduos que pertenciam aos domicílios selecionados foram convidados a participar. Mais informações sobre o processo amostral podem ser consultadas em (21).

Nesse contexto, este projeto de delineamento transversal de base populacional pre-

tende avaliar, em residentes do município de São Paulo, fatores modificáveis relacionados ao estilo de vida, bem como sua associação com marcadores bioquímicos e genéticos relacionados a fatores de risco cardiometabólico.

Os dados foram coletados nos domicílios em forma de questionário - a fim de avaliar fatores sociodemográficos e de estilo de vida e IMC - e coleta de sangue, aplicados em indivíduos de ambos os sexos com 12 anos ou mais. Por meio do sangue foi aferido concentração de micronutrientes, glicemia, perfil lipídico, biomarcadores de inflamação e polimorfismos de nucleotídeo único. Variáveis sociais também foram coletadas para investigar associações, como sexo, idade, raça, estado civil, situação de renda, entre outras.

3.2 Análise Exploratória

Primeiramente, o banco de dados composto por 841 indivíduos foi lido, excluindo os indivíduos que possuem grau de parentesco, totalizando 707 indivíduos para análise. A exclusão foi feita calculando a matriz de parentesco genômico (GRM) e removendo pares de indivíduos com um coeficiente de parentesco maior ou igual a 0,125, o que corresponde aproximadamente a parentes de segundo grau. Essa análise foi conduzida por outros colaboradores do projeto ao qual o banco de dados pertence, e, por essa razão, uma descrição mais detalhada do método não está no escopo deste trabalho.

Dentre as variáveis presentes no banco, serão utilizadas as seguintes variáveis:

- Índice de Massa Corporal (IMC): variável contínua que mede o índice de massa corporal em kg/m^2. Apenas para adultos e idosos, visto que a medida contínua não é adequada para avaliar adolescentes;

- Circunferência de Cintura Abdominal: variável contínua que indica o valor da circunferência de cintura abdominal do indivíduo em centímetros;

- Pressão diastólica (PAD): variável contínua que indica o valor da pressão diastólica do indivíduo em $mmHg$, fazendo a média de três medidas e considerando a média do braço com maior valor;

- Pressão sistólica (PAS): variável contínua que indica o valor da pressão sistólica do indivíduo em $mmHg$, fazendo a média de três medidas e considerando a média do braço com maior valor;

- Nível de insulina: variável contínua que indica o nível de resistência a insulina do indivíduo, calculado por $glicose * insulina/405$;

- Colesterol Não HDL: variável contínua que indica o colesterol não HDL sérico dosado por meio de método enzimático colorimétrico;

- LDL: variável contínua que indica o colesterol LDL sérico dosado por meio de método enzimático colorimétrico;

- HDL: variável contínua que indica o colesterol HDL sérico dosado por meio de método enzimático colorimétrico;

- Triglicérides: variável contínua que indica os triacilgliceróis sérico dosado por meio de método enzimático colorimétrico;

- Nível de Inflamação: variável contínua que indica o valor da Proteína C Reativa (PCR), em mg/dL. Valores muito altos indicam inflamação aguda;

- CP1: primeira componente de ancestralidade global, cuja obtenção é detalhada na Seção 3.3;

- CP2: segunda componente de ancestralidade global, cuja obtenção é detalhada na Seção 3.3;

- Sexo: variável binária representando o sexo do indivíduo, sendo 1 para indivíduos masculinos e 2 para indivíduos femininos;

- Faixa etária: variável categórica já previamente criada indicadora da faixa etária do indivíduo, sendo 0 para adolescente (12-19 anos), 1 para adulto (20-59 anos) e 2 para idoso (60-9 anos);

- Obesidade: variável binária que indica a presença ou ausência de obesidade na população. É calculada para todas as faixas etárias, segundo pontos de corte de IMC específicos utilizados para cada uma delas;

- Nível de Inflamação categórica: variável categórica já previamente criada que indica o nível de inflamação em um indivíduo. Seu valor é 1 para inflamações baixas, 2 para intermediárias, 3 para altas e 4 para agudas;

- Diabetes: variável binária que indica a presença ou ausência de diabetes na população. Seu valor é calculado por glicemia de jejum > 126 ou uso de medicação (hipoglicemiantes orais/insulina injetável);
- Hipertensão: variável binária que indica a presença ou ausência de hipertensão na população. Seu valor é calculado da seguinte forma:
  - Para adolescentes de 12 e 13 anos: PAS ou PAD $>$ Percentil 95 de sexo, idade e altura;
  - Para adolescentes de 14 a 19 anos: PAS ≥ 130 ou PAD ≥ 80;
  - Para adultos e idosos: PAS ≥ 140 ou PAD ≥ 90. Ou uso de medicação (hipotensores e diuréticos);
- Dislipidemia: variável binária que indica a presença ou ausência de dislipidemia na população. Seu cálculo é feito com base em condições relacionadas a triglicérides LDL e HDL ou pelo uso de medicação hipolipemiante.

Tabelas de frequência foram feitas para as variáveis sexo e faixa etária, a fim de visualizar como estão distribuídos os indivíduos.

Tabela 3.1: Tabela de frequência para a variável sexo.

	Frequência	%
Masculino	365	51,6
Feminino	342	48,4
Total	707	100

Tabela 3.2: Tabela de frequência para a variável faixa etária.

	Frequência	%
Jovem	166	23,5
Adulto	249	35,2
Idoso	292	41,3
Total	707	100

De acordo com as Tabelas 3.1 e 3.2 observamos que os indivíduos estão bem distribuídos quanto ao gênero, com 48,4% de indivíduos do sexo feminino e 51,6% do sexo

masculino. Já para a faixa etária há uma ligeira prevalência de adultos e idosos com relação aos jovens.

Também foi feita a tabela de frequência para a variável obesidade. A coluna %(NA+) considera as observações sem categoria definida para o cálculo das porcentagens, já a coluna %(NA-) são desconsideradas essas observações.

Tabela 3.3: Tabela de frequência para a variável obesidade.

	Frequência	%(NA+)	%(NA-)
Não Possuem obesidade	530	75,0	75,3
Possuem obesidade	174	24,6	24,7
NA	3	0,4	0,0
Total	707	100	100

Observando a Tabela 3.3, este conjunto de dados apresenta 24,7% dos indivíduos com obesidade.

O gráfico da Figura 3.1 foi feito para verificar a influência do sexo na variável obesidade.

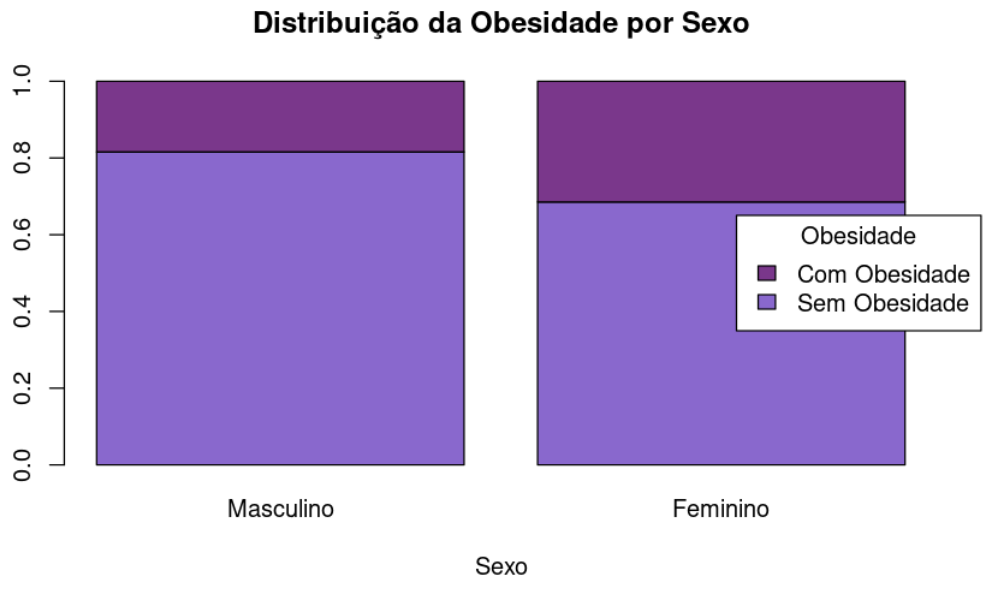

Figura 3.1: Gráfico de barras da obesidade por sexo.

A partir do gráfico observa-se uma possível influência do sexo, visto que há uma quantidade maior de mulheres com obesidade do que homens.

A seguir foi feito o gráfico da Figura 3.2 para verificar agora a influência da faixa etária na variável obesidade.

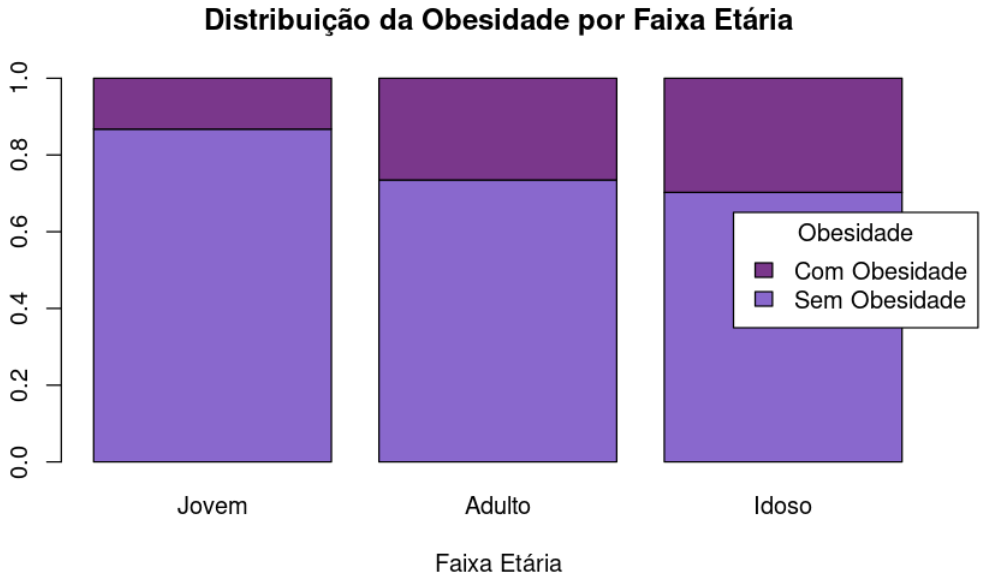

Figura 3.2: Gráfico de barras da obesidade por faixa etária.

A partir do gráfico observa-se uma possível influência da faixa etária, visto que há uma quantidade menor de jovens com obesidade em relação aos adultos e idosos.

Testes qui-quadrado também foram aplicados para as variáveis categóricas, com o objetivo de verificar independência da variável obesidade.

Tabela 3.4: Tabela combinada dos testes qui-quadrado para obesidade.

	P-valor para Obesidade
Hipertensão	0,0000002
Dislipidemia	0,0004364
Diabetes	0,0000114
Sexo	0,0000856
Faixa etária	0,0003192
Nível de Inflamação categórica	0,0000000

A partir dos resultados dos testes apresentados na Tabela 3.4, podemos afirmar ao nível de 5% de significância que as variáveis não são independentes de obesidade.

Já para as variáveis quantitativas do banco, primeiramente foi construída a Tabela 3.5 de medidas resumo.

Tabela 3.5: Tabela de Medidas Resumo.

	Mín.	1º Quartil	Mediana	Média	3º Quartil	Máx.	NA's
Circunferência de cintura	53,50	81,01	93,00	92,34	102,88	171,30	9
Pressão diastólica	46,33	69,00	77,00	77,16	84,00	132,67	4
Pressão sistólica	93,00	115,67	127,50	130,99	143,00	220,50	4
Nível de insulina	0,23	1,73	2,63	3,74	4,21	54,54	7
Colesterol não HDL	24,00	98,00	126,00	130,34	156,00	392,00	15
LDL	6,00	79,00	104,00	106,13	129,00	373,00	15
HDL	10,00	34,00	43,00	44,16	52,25	119,00	15
Triglicérides	10,00	75,00	102,00	122,95	144,00	1.402,00	14
Nível de Inflamação	0,02	0,11	0,32	0,56	0,81	2,24	22

A partir das medidas de referência em (8) podemos verificar a presença de indivíduos com altos valores de circunferência de cintura (ccm), representando cerca de 25% da amostra, sendo condizente com a análise categórica apresentada para a obesidade. Para as demais variáveis, podemos observar de forma geral grande discrepância entre os valores mínimo e máximo, com médias mais próximas do valor mínimo.

A seguir foi calculada a correlação linear (de Pearson) para cada par de variáveis contínuas, apresentado na Figura 3.3.

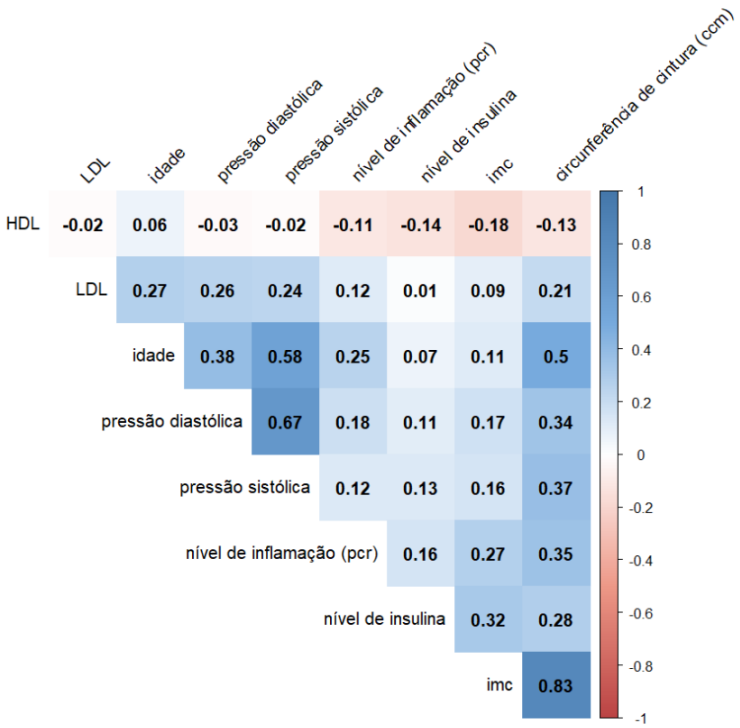

Figura 3.3: Correlação linear entre as variáveis quantitativas de interesse.

Pode-se observar a partir da Figura 3.3 uma alta correlação já esperada entre as variáveis imc e circunferência de cintura abdominal (ccm), além de uma alta correlação da idade com ambas as pressões e da idade com a circunferência de cintura.

A partir das relações vistas, o gráfico da Figura 3.4 foi construído para uma investigação mais detalhada. A variável circunferência de cintura foi escolhida como versão contínua da variável obesidade pois a variável IMC não possui bom desempenho para jovens.

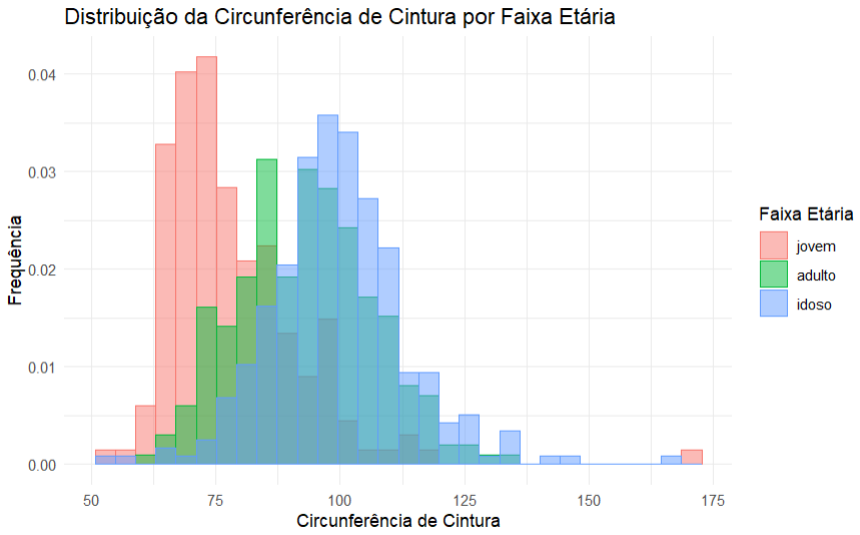

Figura 3.4: Gráfico da distribuição da circunferência de cintura por faixa etária.

A estrutura etária mostrada pela Figura 3.4 apresenta diferentes distribuições para a circunferência de cintura, demonstrando a influência da faixa etária na circunferência de cintura observada anteriormente. Observa-se que o pico para os jovens é em torno de 70 cm de circunferência, enquanto que para os adultos esse valor é de 80 cm e pra os idosos 100 cm.

O gráfico da Figura 3.5 foi construído para uma investigação mais detalhada de como se comporta a circunferência de cintura por sexo.

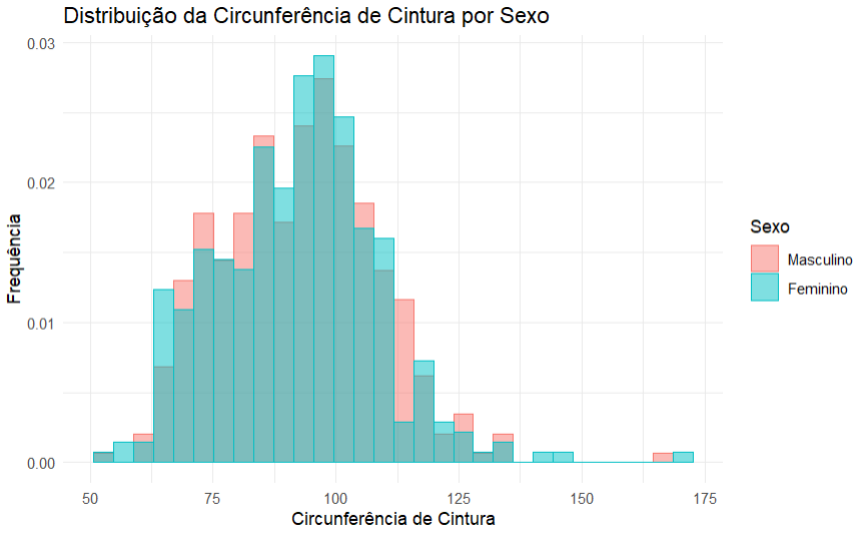

Figura 3.5: Gráfico da distribuição da circunferência de cintura por sexo.

A estrutura mostrada pela Figura 3.5 parece ser semelhante para ambos os sexos, com um pico em torno de 100 cm de circunferência. Em alguns valores mais altos pode-se observar uma sutil diferença para os homens, como por exemplo em torno de 115 cm.

3.2.1 Resumo da Análise Exploratória

A análise exploratória destacou diversos aspectos importantes. Primeiramente, com relação ao indivíduos analisados, temos que o banco inicial contava com 841 indivíduos, mas após a exclusão daqueles com grau de parentesco significativo (coeficiente de parentesco $\geq 0,125$), a amostra final utilizada é de 707 indivíduos.

Com relação a características demográficas, a distribuição por sexo foi equilibrada, com 51,6% de homens e 48,4% de mulheres. Já em relação à faixa etária, observou-se uma maior proporção de idosos (41,3%), seguidos por adultos (35,2%) e jovens (23,5%).

A obesidade esteve presente em 24,7% da amostra, com indícios de diferenças na prevalência entre os sexos, sendo mais comum entre mulheres. Além disso, foi observada maior frequência de obesidade entre adultos e idosos, enquanto os jovens apresentaram taxas menores.

As variáveis quantitativas apresentaram ampla variabilidade, especialmente em medidas como circunferência de cintura, pressão arterial, colesterol e níveis de inflamação.

Foi observada uma forte relação entre obesidade e condições como hipertensão, dislipidemia e diabetes, sugerindo potenciais inter-relações. Testes estatísticos de associação indicaram que a obesidade está significativamente relacionada a variáveis categóricas como

hipertensão, dislipidemia, diabetes, sexo, faixa etária e níveis de inflamação categórica, com p-valores inferiores a 0,05. Em linhas gerais, a análise exploratória evidenciou diferenças demográficas e de saúde que poderão ser aproveitadas para condução do GWAS.

3.3 Aplicação ao Banco de Dados

Neste estudo, identificamos as variantes genéticas associadas a variável obesidade. Para este trabalho, gostaríamos de estudar a existência de obesidade na população e a relação da obesidade com as características genotípicas dos indivíduos.

As características observadas na amostra evidenciam que a população de forma geral possui poucos indivíduos portadores de características como diabetes, hipertensão, dislipidemia e até mesmo obesidade. Quando muitos genes estão envolvidos no controle de uma característica, em geral, tais genes são de efeito menor, consequentemente observamos poucos indivíduos portadores na população. Isso dificulta a obtenção de genes associados, reafirmando a importância de um bom modelo para obtenção desses marcadores. Além disso, as potenciais inter-relações entre as doenças dificultam a obtenção de marcadores exclusivos a obesidade, visto que estes podem aparecer devido a presença de outra doença no modelo, como diabetes por exemplo.

Os métodos de controle de qualidade apresentados na Seção 2.3 foram aplicados no banco de dados antes da execução do modelo, removendo os SNPs que não foram genotipados, bem como aplicação do MAF $\geq 5\%$, LD e o critério de desequilíbrio de Hardy-Weinberg ao nível de significância de 5%. O controle de qualidade dos SNPs foi realizado no PLINK 2.0.

Para a aplicação do LD, foi considerada uma janela de 50 SNPs e calculado o LD entre cada par de SNPs na janela. Foram removidos um de um par de SNPs se o LD fosse maior que 0,5, após foi deslocada a janela em 5 SNPs para frente e repetido o procedimento.

A Figura 3.6 a seguir mostra a configuração dos SNPs em todos os cromossomos antes e após a aplicação dos filtros do controle de qualidade.

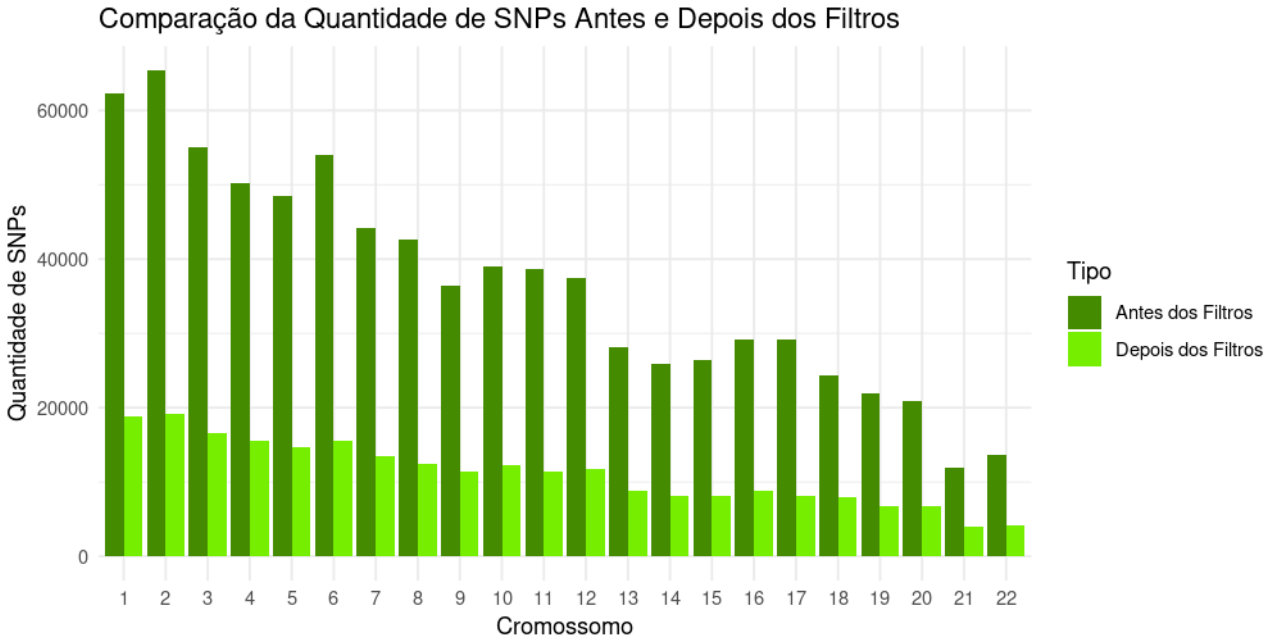

Figura 3.6: Gráfico de barras da quantidade de SNPs por cromossomo antes e após a aplicação dos filtros mencionados.

Após a aplicação dos filtros, a quantidade total de SNPs foi reduzida de 873.177 para 244.338. Os SNPs mostrados em verde claro na Figura 3.6 serão utilizados na construção do modelo, conjuntamente com variáveis físicas como idade e sexo.

3.3.1 Obtenção das Componentes de Ancestralidade Global

A análise de ancestralidade global foi realizada por outros colaboradores do projeto temático ao qual o banco de dados pertence, por essa razão o método de forma detalhada não será apresentado no escopo deste trabalho.

Para conduzir a análise de ancestralidade global, o controle de qualidade do SNP foi realizado no PLINK 2.0 usando 873.177 SNPs. Após aplicação dos filtros MAF, Desequilibrio de HW e LD resultaram 417.192 SNPs de alto desempenho.

Nas análises a amostra ISA-Nutrição (ISA) de 2015 foi comparada com a fase 3 do Projeto 1000 Genomas (1KGP), abrangendo indivíduos de populações ancestrais conhecidas, mantendo indivíduos com $> 95\%$ de ancestralidade homogênea, selecionando assim 1585 indivíduos. As populações ancestrais foram selecionadas de várias regiões, como da África Subsaariana (AFR), nativos americanos (AMR), asiáticos orientais (EAS) e europeus (EUR).

Os dados 1KGP foram mesclados com o conjunto de dados ISA, removendo 178.376 SNPs do ISA que não correspondiam ao 1KGP, resultando em 226.346 SNPs para ava-

liação.

Ferramentas como análise de componentes principais (PCA) foram usadas para analisar a estrutura populacional. A melhor configuração de clusters foi definida com k=4, refletindo as quatro superpopulações principais. O PCA foi realizado com os 226.346 SNPs em uma matriz de genótipo de 2305 indivíduos (ISA + 1KGP). Os componentes principais (PCs) foram usados para separar as populações com base em variações genômicas.

Os autovetores ou componentes principais (PCs) foram classificados em ordem decrescente dos autovalores correspondentes, ou seja, o primeiro autovetor (PC1) é responsável pela maior variação nos dados, o segundo autovetor (PC2) é responsável pela segunda maior, e assim por diante.

Dessa forma, os componentes CP1 e CP2 agregam a maior variação ancestral e podem ser aplicados no modelo a fim de homogeneizar os indivíduos utilizados no GWAS.

Dessa forma, os resultados obtidos por meio desta análise serão utilizados neste trabalho. Mais detalhes podem ser consultados em (25).

Capítulo 4

Florestas Aleatórias

Os algoritmos de aprendizado de máquina se tornam cada vez mais frequentes na classificação e seleção de características com fatores de proteção ou risco, principalmente algoritmos baseados em florestas aleatórias (do inglês *random forest*, RF). O benefício da aplicação do método RF está na habilidade de suportar bases de dados com um grande número de preditores (como por exemplo SNPs), além de poder capturar relações não-lineares e interações entre elas.

4.1 Árvores de Decisão

Árvore de decisão é uma metodologia não paramétrica, construída para propósitos de predição e utilizada para tarefas de classificação e regressão. Cada particionamento da árvore recebe o nome de nó e cada resultado final recebe o nome de folha. As árvores de regressão são utilizadas para variáveis resposta contínuas, enquanto que as árvores de classificação são utilizadas por variáveis resposta categóricas.

A utilização da árvore para prever uma nova observação é feita começando pela raiz da árvore, verificando se a condição do primeiro nó é satisfeita, caso seja, seguimos a esquerda. Caso contrário, seguimos a direita e assim prosseguimos até atingir uma folha (13).

No exemplo da Figura 4.1, caso as condições 1 e 2 sejam satisfeitas a predição será dada por F1.

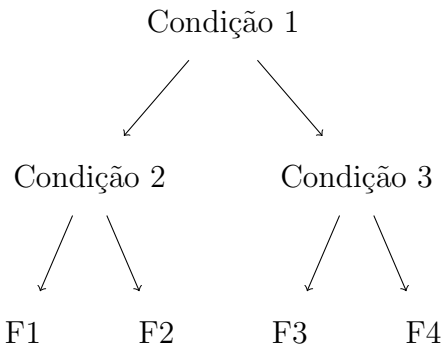

Figura 4.1: Exemplo de uma árvore de decisão.

Para prever o valor da variável resposta Y com base na observação de uma variável X, observamos a região a qual a observação de x pertence e calculamos uma estatística dos valores da variável resposta Y das observações do conjunto de treinamento pertencentes àquela mesma região. Para o caso de árvores de regressão é calculada a média, já para as árvores de classificação a moda.

Mais especificamente, uma árvore cria uma partição do espaço das covariáveis em regiões distintas e disjuntas: $R_1, ..., R_j$. A predição para a resposta Y de uma observação pertencente a R_k, com covariáveis x, tal que $x = (x_1, ..., x_p)$ onde p é a quantidade de covariáveis, para um problema de regressão é dada por:

$$g(x) = \frac{1}{|i : x_i \in R_k|} \sum_{i:x_i \in R_k} y_i,$$

onde y_i é o valor da variável resposta da i-ésima observação e x_i os valores das covariáveis da i-ésima observação, tal que $i = 1, ..., n$ onde n é a quantidade de observações em R_k, e $x_i = (x_{i1}, ..., x_{ip})$.

Já para um problema de classificação adaptamos a expressão acima para

$$g(x) = \text{moda}\{y_i : x_i \in R_k\}.$$

A criação da estrutura de uma árvore de regressão é feita através de duas grandes etapas: (i) a criação de uma árvore completa e complexa e (ii) a poda dessa árvore, com a finalidade de evitar o super ajuste.

Para criação da árvore no passo (i) buscam-se os valores de Y homogêneos nas observações em cada uma das folhas, desse modo utiliza-se o erro quadrático médio (EQM), dado pela Equação (4.1), para encontrar uma árvore T com baixo EQM.

$$EQM(T) = \sum_{R_k} \sum_{i:x_i \in R_k} \frac{(y_i - \hat{y}_{r_k})^2}{n}, \tag{4.1}$$

onde $\hat{y}_{r_k}$ é o valor predito para a resposta de uma observação i pertencente à região R_k. Essa avaliação consiste na criação de divisões binárias recursivas, dado que encontrar T na equação (4.1) que minimize $EQM(T)$ é computacionalmente inviável.

Assim, para escolha da primeira partição busca-se, dentre todas as covariáveis x_j, para $j = 1, ..., p$, e cortes t_1 a combinação que leva a uma partição (R_1, R_2) com menor EQM, que é dada por:

$$EQM(T) = \sum_{i:x_i \in R_1} (y_i - \hat{y}_{R_1})^2 + \sum_{i:x_i \in R_2} (y_i - \hat{y}_{R_2})^2. \tag{4.2}$$

Assim podemos estabelecer $R_1 = \{x : x_j < t_1\}$ e $R_2 = \{x : x_j \geq t_1\}$. A árvore da Figura 4.2 a seguir demonstra essa partição.

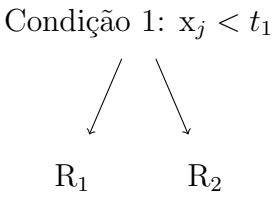

Figura 4.2: Exemplo de uma árvore de decisão.

Uma vez estabelecidas as regiões, a raiz da árvore é fixada. No próximo passo busca-se particionar R_1 ou R_2 em regiões menores com a mesma estratégia. Ou seja, busca-se, dentre todas as covariáveis x_j e cortes t_2 a combinação que leva ao menor EQM. Esse processo é repetido de forma recursiva como mostrado na Figura 4.3. É estabelecido como critério de parada quando as folhas tem menos de cinco observações.

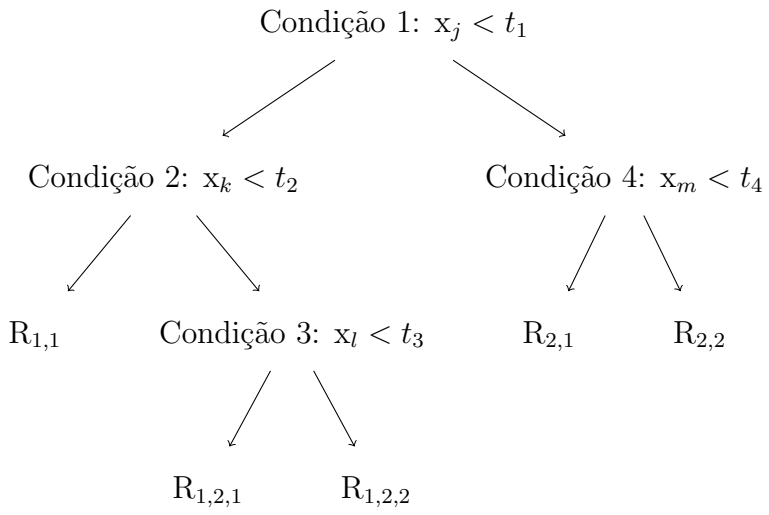

Figura 4.3: Exemplo de uma árvore de decisão.

Após a construção da árvore é iniciado o processo (ii) de poda, no qual cada nó é retirado, um por vez, e observa-se como o erro de predição varia no conjunto de validação, decidindo quais nós permanecerão na árvore. Essa etapa tem como finalidade evitar o super-ajuste do conjunto de treinamento utilizado para construção.

Para as árvores de classificação o processo (i) de criação é adaptado pelo critério do índice de Gini:

$$G(t) = 1 - \sum_{k=1}^{K} p_k^2(t),$$

onde:

- $p_k(t)$ é a proporção de observações da classe k no nó t.

- K é o número total de classes.

Para esse processo avaliamos a homogeneidade das classes dentro de um nó de decisão, com o índice variando de 0 a 0.5 para problemas binários. $G(t) = 0$ indica que o nó possui todas as observações de uma única classe. Já o valor $G(t) = 0,5$ indica que as observações estão distribuídas de forma uniforme pelas classes. Busca-se, portanto, a redução desse índice.

Já para a etapa (ii) da poda, em geral, utiliza-se a proporção de erros no conjunto de validação como estimativa do risco.

4.2 Florestas Aleatórias

Árvores de regressão são extremamente interpretáveis, porém costumam apresentar baixo poder preditivo. As florestas aleatórias combinam diversas árvores para fazer uma melhor predição.

Entre as vantagens da floresta aleatória podemos citar baixo risco de super-ajuste, flexibilidade para tratar de diferentes variáveis resposta e interações complexas e possibilidade de aplicação em situações com poucas observações mas muitas covariáveis, que muitas vezes são altamente correlacionadas. Além disso, é possível aplicar uma medida de importância para as covariáveis, possibilitando medir o impacto de cada covariável na variável resposta.

Para construção da floresta criam-se B árvores distintas utilizando B amostras bootstrap da amostra original. Todavia, para a criação de cada nó em cada uma das B árvores são disponibilizadas apenas m covariáveis, selecionadas de forma aleatória, tal que m seja menor que o número total de covariáveis (p), a fim de diversificar as árvores criadas para redução da correlação entre as mesmas. As árvores criadas utilizam as técnicas descritas na Seção 4.1. Porém, a fim de não viesar os estimadores não podamos as árvores criadas e portanto utilizamos apenas o passo (i) descrito. Usualmente são escolhidos valores de $m = \frac{p}{3}$ e valores $B > 100$ (13).

Seja $g_b(x)$ a função de predição obtida segundo a b-ésima árvore. A função de predição da floresta é dada por

$$g(x) = \frac{1}{B}\sum_{b=1}^{B} g_b(x).$$

Já para o problema de classificação essa função é dada por

$$g(x) = \text{moda}\{g_b(x), b = 1, ..., B\}.$$

Assim, os resultados das diferentes árvores são utilizados em termos de média ou moda para potencializar os resultados de uma única árvore individual.

Dada a função $g_b(x)$, podemos avaliar a importância de cada covariável no modelo. Existem diversas maneiras para o cálculo de importância, sendo o índice de Gini e a redução da soma de quadrados dos resíduos as mais usuais.

Para o primeiro caso, cada árvore é construída dividindo os nós com base no índice de Gini, que é utilizado nos problemas de classificação. A redução do índice de Gini para

um nó t é calculada como:

$$\Delta G(t) = G(t) - \left(\frac{N_{\text{esq}}}{N_t} G(\text{esq}) + \frac{N_{\text{dir}}}{N_t} G(\text{dir}) \right),$$

onde:

- $G(t)$ é o índice de Gini no nó t.
- $G(\text{esq})$ e $G(\text{dir})$ são os índices de Gini dos filhos esquerdo e direito, respectivamente.
- N_t é o número de observações no nó t.
- N_{esq} e N_{dir} são o número de observações nos filhos esquerdo e direito, respectivamente.

A importância da covariável X_j em uma árvore b é a soma das reduções do índice de Gini para todos os nós t onde X_j é usada:

$$\text{M}_{\text{Gini b}}(X_j) = \sum_{t \in T} \Delta G(t) \cdot \mathbb{1}(h(t) = X_j),$$

tal que $\mathbb{1}(h(t) = X_j)$ é uma função indicadora que vale 1 se a variável X_j é usada no nó t e 0 caso contrário, e T é o conjunto de todos os nós na árvore.

Para o cálculo na floresta aleatória, a importância de X_j é a média das importâncias calculadas em todas as árvores:

$$\text{M}_{\text{Gini floresta}}(X_j) = \frac{1}{B} \sum_{b=1}^{B} \text{M}_{\text{Gini b}}(X_j),$$

onde B é o número total de árvores na floresta e $\text{M}_{\text{Gini b}}(X_j)$ é a importância da variável X_j na árvore b.

Ao calcular essa média pode-se avaliar a importância de uma covariável. Se o índice relativo a covariável é próximo de um tem-se que ela é importante para o modelo, já para valores próximos de zero indicam que a covária vel não é importante. Ou seja, quanto maior o valor, maior a contribuição da covariável para o modelo.

Já para o segundo caso, a importância de uma covariável em uma floresta aleatória pode ser medida pela variação na soma dos quadrados dos resíduos quando os valores da covariável são permutados.

A soma dos quadrados dos resíduos (do inglês *residual sum of squares*, RSS) é definida como:

$$\text{RSS} = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2,$$

onde N é o número de observações no conjunto de teste, y_i é o valor da resposta para a i-ésima observação e $\hat{y}_i$ é o valor predito da resposta para a i-ésima observação.

Para cálculo da importância de uma covariável X_j primeiramente são construídas estimativas avaliando os dados de treinamento que não são incluídos no conjunto de amostras *bootstrap*, chamados de dados *out-of-bag* (OOB). Cada árvore construída com os dados que foram escolhidos pela amostra *bootstrap* é utilizada para predizer os valores de Y dos dados OOB. Essas predições são comparadas com os valores verdadeiros para obter uma estimativa do RSS, denominado $\text{RSS}_{\text{original}}$. Após, é feita para cada covariável X_j uma permutação aleatória dos valores de X_j no conjunto OOB, mantendo todas as outras variáveis inalteradas e calculada a RSS para essa situação, denominada $\text{RSS}_{\text{permutada}}$.

Dado um número B de árvores, a importância da covariável X_j é dada pela diferença média entre a RSS permutada e a RSS original.

$$\text{M}_{\text{RSS}}(X_j) = \frac{1}{B} \sum_{b=1}^{B} \left(\text{RSS}_{\text{permutada},b}(X_j) - \text{RSS}_{\text{original},b}(X_j) \right).$$

Ao calcular essa diferença pode-se avaliar a importância de uma covariável. Se a permutação da covariável aumenta significativamente a RSS tem-se que ela é importante para o modelo. Ou seja, a importância da covariável é medida pelo impacto que a retirada da informação que ela possui causa no erro de predição (1).

As florestas podem ser implementadas por meio do pacote `randomForest` do *R* (28) com a utilização da função `randomForest` (10). O pacote *ranger* (19) também pode ser utilizado para construção das florestas, comumente conhecido por sua rápida aplicação em grandes bancos de dados.

4.3 Florestas Aletórias e GWAS

Estudos de GWAS são caracterizados pela coleta de informações genéticas dos indivíduos por meio de polimorfismos de nucleotídeo único (SNPs) existentes no genoma. Para esse estudo supõe-se que a variação em SNPs específicos leva a mudanças nas características fenotípicas, como doenças. Para isso temos que o SNP associado pode ser causador da doença ou estar em desequilíbrio de ligação (LD) com outras variantes cau-

sadoras da doença.

Observando o genótipo dos indivíduos com e sem a doença é possível identificar tal variação. Há diversas maneiras de fazer essa identificação, porém muitas supõe independência entre os genes como a regressão linear ou logística, enquanto outras não são computacionalmente viáveis por sua vasta extensão, como por exemplo testar todos os modelos genéticos possíveis, incluindo aqueles para interação. Desse modo as florestas aleatórias parecem ser uma ótima abordagem para a integração de vários de SNPs, além das interações entre os mesmos, visto que lidam naturalmente com interações entre variáveis além capacidade de identificar as variáveis de interesse em conjuntos de dados extensos.

Outra característica que torna as florestas aleatórias como um modelo atrativo para o GWAS é sua capacidade de identificar os SNPs importantes, seguindo os métodos descritos na Seção 4.2.

Após o cálculo da importância para todos os SNPs podemos visualizar os resultados obtidos por meio de um Gráfico Manhattan. Esse gráfico possui em seu eixo x os cromossomos com suas distâncias genéticas e em seu eixo y sua respectiva pontuação de importância. Para definição de um limite de importância para seleção dos SNPs várias métricas podem ser utilizadas, como por exemplo a seleção de 1% dos SNPs com maior importância (1).

Capítulo 5

Resultados

Este capítulo apresenta os resultados obtidos com a aplicação das florestas aleatórias ao conjunto de dados analisado, com a imputação dos genótipos faltantes. Em seguida, são identificados os SNPs mais relevantes para os fenótipos investigados, permitindo a construção de um ranking de importância das variantes genéticas, discutindo as implicações e relevâncias dos genes encontrados.

5.1 Imputação dos genótipos faltantes

As florestas aleatórias desconsideram os dados faltantes nos SNPs, o que pode levar à perda de informações importantes. Para evitar essa perda, antes de prosseguir com a análise, foi realizada a imputação dos genótipos faltantes por meio da fração de recombinação.

A fração de recombinação (também chamada de frequência de recombinação) entre dois loci é definida como a razão entre o número de haplótipos recombinados e o número total de haplótipos produzidos. Haplótipo é um conjunto de alelos próximos que tendem a ser herdados juntos, como um bloco. Dado dois loci A e B, com alelos A e a, B e b, os haplótipos possíveis são AB e ab. Quando há recombinação presenciamos a existência de haplótipos Ab ou aB.

Quanto menor a frequência de recombinação mais próximos estão os loci. Geralmente ela é denotada por r, com seu valor mínimo sendo zero, representando a ligação perfeita, e seu valor máximo sendo 0.5, que representa independência completa entre os dois loci.

Ela pode ser calculada via funções biológicas como a função de Haldane, calculada a partir da distância genética entre dois loci. A função de Haldane é definida como:

$$r = \frac{1}{2}(1 - \exp(-2d)),$$

em que d é a distância em Morgan entre os dois loci.

Por ser uma medida de associação genética entre os dois loci, a fração de recombinação é geralmente utilizada para imputação de genótipos faltantes. O genótipo faltante de um loci pode ser imputado através do genótipo dos loci vizinhos, considerando a associação que existe entre essas regiões caso estejam em desequilíbrio de ligação ou muito próximas dentro do DNA.

As probabilidades de cada genótipo no loci faltante são calculadas através de probabilidades condicionais em uma cadeia de Markov e usando a fração de recombinação entre os loci.

Seja X_f o genótipo faltante e X_e e X_d o genótipo do loci à esquerda do faltante e do loci à direita do faltante, respectivamente. Então, usando a propriedade Markoviana:

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

Essas probabilidades são calculadas para $X_f \in \{0, 1, 2\}$, com X_e e X_d observados, e usando a fração de recombinação entre os loci. Para mais detalhes, ver (32).

A fração de recombinação (r) entre dois loci é usada para calcular as probabilidades de transição entre os genótipos de forma que:

- Se r é pequeno ($r \approx 0$): Existe uma ligação forte entre os loci, indicando que os genótipos são altamente correlacionados. O genótipo faltante provavelmente será igual ao loci vizinho.

- Se r é grande ($r \approx 0.5$): Os loci são independentes, e o genótipo faltante será distribuído aleatoriamente de acordo com as frequências alélicas observadas.

A matriz de transição entre dois loci (X_e e X_f) é construída com base na fração de recombinação (r), que reflete a probabilidade de que uma recombinação tenha ocorrido entre esses loci. Para loci com três possíveis genótipos cada (0 = homozigoto dominante, 1 = heterozigoto, 2 = homozigoto recessivo), a matriz será 3×3, representando as probabilidades de transição de um genótipo em X_e para outro genótipo em X_f.

Assim, a matriz de transição, T, indica as probabilidades $P(X_f = j \mid X_e = i)$, onde:

- i é o genótipo do loci X_e (linha da matriz).

- j é o genótipo do loci X_f (coluna da matriz).

As probabilidades de transição são dadas por:

- **Sem recombinação** $(1-r)^2$**:** O genótipo do loci X_f será o mesmo que o do loci X_e. Para não ter mudança de genótipo, não pode haver recombinação (r) no gene materno e paterno.
- **Com recombinação** $(r)^2$**:** O genótipo do loci X_f será distribuído aleatoriamente entre os outros genótipos.

Assim, a matriz T será:

$$T = \begin{bmatrix} (1-r)^2 & 2r(1-r) & r^2 \\ r(1-r) & 1-2r(1-r) & r(1-r) \\ r^2 & 2r(1-r) & (1-r)^2 \end{bmatrix}$$

Quando r é pequeno ($r \approx 0$), a matriz se aproxima de uma matriz identidade, indicando forte ligação entre os loci, já quando $r \approx 0.5$, os loci tornam-se independentes, e a matriz apresenta distribuições mais aleatórias.

5.2 Aplicação da floresta aleatória

Após a aplicação dos filtros de controle de qualidade discutidos no Capítulo 3, foram selecionados 244.338 SNPs para análise. Os SNPs com genótipos faltantes foram imputados utilizando o método descrito na Seção 5.1.

O algoritmo de florestas aleatórias foi executado por meio o pacote `randomForest` do *R*, com a função `randomForest`. A referência para o pacote pode ser encontrada no PDF oficial do R, disponível em (15). Para a aplicação da função foram estabelecidos os seguintes parâmetros principais:

- **Número total de árvores** (B)**:** 500 (default);
- **Número de variáveis consideradas por nó** (m)**:** $m = \sqrt{p}$ para a variável categórica e $m = \frac{p}{3}$ para a variável contínua, onde p é o número total de SNPs. Para obesidade, $m = \sqrt{244.338} \approx 494$, e para CCM, $m = \frac{244.338}{3} = 81.446$;
- **Critério de divisão:** índice de Gini, para a variável categórica e Soma de Quadrados dos Resíduos para a variável contínua.

Além dos marcadores genéticos, variáveis sociodemográficas foram utilizadas como sexo, idade, CP1 e CP2. É importante ressaltar que não foram utilizadas como covariáveis as variáveis clínicas como diabetes, dislipidemia e hipertensão, visto que estas possuem alta relação com a obesidade e podem coagir a obtenção de SNPs que não são diretamente associados a doença.

Os valores ausentes existentes nas variáveis resposta (cerca de 3 valores para obesidade e 9 para circunferência de cintura) foram imputados via média para a circunferência de cintura abdominal e de modo proporcional as categorias existentes para a obesidade.

A predição do modelo para uma nova observação é realizada pela média para a circunferência de cintura e moda para obesidade.

A importância dos SNPs foi calculada utilizando a *redução da soma de quadrados dos resíduos (RSS)* e a redução do índice de Gini (G). A importância total de um SNP foi obtida somando $\Delta RSS(t)$ ou $\Delta G(t)$ ao longo de todos os nós onde esse SNP foi utilizado.

É importante ressaltar que o desempenho da floresta, bem como o desbalanceamento de classe existente, não foram levados em consideração, visto que o estudo não possui objetivo de predição.

5.3 Aplicação da floresta para obesidade

Após execução da floresta para a obesidade, os SNPs foram rankeados de acordo com sua pontuação de importância.

O gráfico Manhattan da Figura 5.1 demonstra a distribuição dos SNPs nos cromossomo por sua respectiva pontuação de importância.

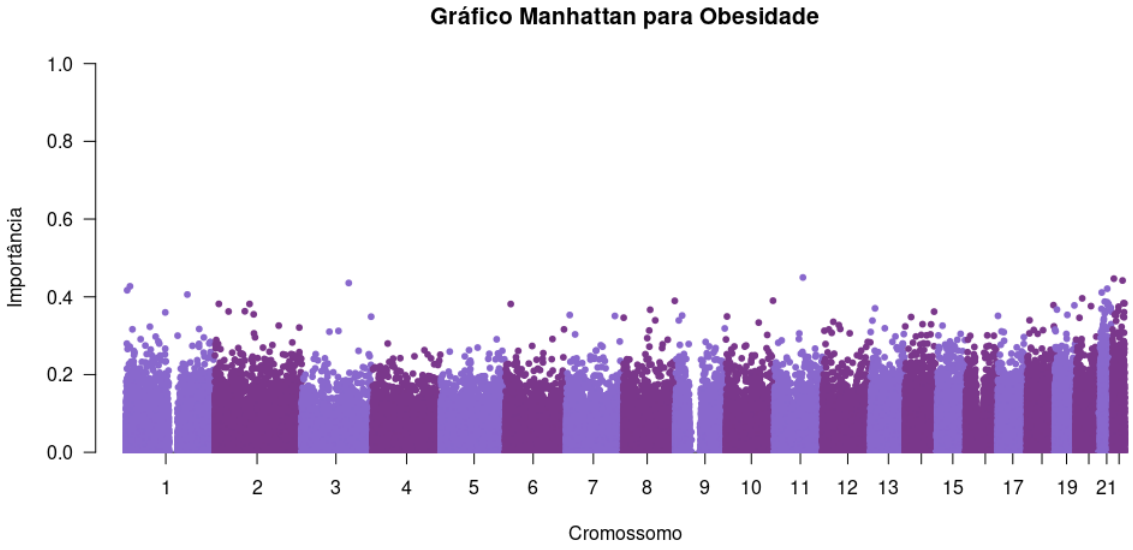

Figura 5.1: Gráfico Manhattan para a variável obesidade.

Por meio do Gráfico Manhattan apresentado na Figura 5.1, podemos observar que a maioria dos SNPs se concentra abaixo de 40%, com a média sendo 2,5%, mediana 0% e quantil 75% de 3,8%.

Estabelecendo um corte nas 1% maiores importâncias encontradas (quantil 99%), obtemos 2445 SNPs. A Tabela 5.1 apresenta a distribuição das pontuações de importância obtidas para o quantil 99% selecionado.

Tabela 5.1: Medidas Resumo para os SNPs selecionados no quantil 99% para obesidade.

Mínimo	1º Quartil (Q1)	Mediana	Média	3º Quartil (Q3)	Máximo
0,18	0,19	0,21	0,22	0,24	0,45

A partir deste quantil selecionado, podemos destacar 10 SNPs encontrados em mais de 5 estudos prévios pelo PubMed (35), descritos na Tabela 5.2. Também foram selecionados o SNP com maior valor de importância associado e o SNP com importância maior que 0,4 que possuía artigos.

Tabela 5.2: Quantidade de artigos encontrados para cada SNP, cromossomo, importância e descrição associada.

SNP	Cromossomo	Importância	Artigos	Descrição
rs1005753	1	0,23	5	Gene PADI2
rs7521902	1	0,23	15	Gene WNT4
rs2201841	1	0,18	47	Gene IL-23R
rs2229094	6	0,18	30	Gene LTA
rs1801275	16	0,21	55	Gene IL-4Ralpha
rs478582	18	0,20	9	Gene PTPN2
rs28362459	19	0,27	13	Gene FUT3
rs2740210	20	0,24	13	Gene OXT
rs44707	20	0,26	12	Gene ADAM33
rs9616915	22	0,20	9	Gene SHANK3
rs7945201	11	0,45	-	-
rs2284553	21	0,42	3	Gene IFNGR2

A partir da Tabela 5.2 podemos verificar diversos genes associados à obesidade, cada um com um papel específico em processos metabólicos e inflamatórios.

O gene *PADI2* (peptidilarginina deiminase tipo 2) codifica a enzima PAD2, envolvida na citrulinação de proteínas, convertendo resíduos de arginina em citrulina. A citrulinação desempenha um papel em processos autoimunes e inflamatórios, como a artrite reumatoide (34).

Outro gene é o *WNT4*, que tem um papel crucial no desenvolvimento embrionário e na homeostase dos tecidos, regulando a adipogênese, o processo pelo qual células precursoras se diferenciam em adipócitos. Alterações desse gene podem influenciar a formação de tecido adiposo, contribuindo para a obesidade (16).

O gene *IL-23R*, que codifica um receptor para a interleucina-23, que participa da regulação da resposta inflamatória e pode influenciar a suscetibilidade a doenças inflamatórias, como artrite psoriática (27). Do mesmo modo, o gene *LTA* (linfotoxina alfa) é uma citocina pró-inflamatória que pode estar envolvida no desenvolvimento de um estado inflamatório crônico (5), bem como o gene *IL-4Ralpha*, que codifica um receptor envolvido na regulação da resposta imune, com algumas variantes desse gene associadas a doenças inflamatórias como hipercolesterolemia (33).

O gene *ADAM33* está envolvido na remodelação tecidual e já foi associado a doenças respiratórias, como a asma, sugerindo indícios de que possam estar ligados a processos inflamatórios (30).

O gene *PTPN2* codifica a proteína tirosina fosfatase não receptora tipo 2, e está envolvido em vias de sinalização que regulam a função imunológica e a inflamação. Alterações genéticas nesse gene têm sido associadas a diversas condições autoimunes, com variantes do gene comprometendo a função da barreira epitelial intestinal e aumentando o risco de distúrbios inflamatórios (31).

Além da inflamação, a composição da microbiota intestinal também pode influenciar o risco de obesidade. O gene *FUT3*, responsável pela síntese de antígenos do grupo sanguíneo Lewis, pode afetar a expressão de moléculas e modificar o metabolismo energético. Algumas pesquisas sugerem que variantes desse gene podem alterar a microbiota intestinal, impactando a digestão e a absorção de nutrientes, influenciando assim o peso corporal (11).

Outro gene de grande interesse é o *OXT*, que codifica a oxitocina, um neuropeptídeo com funções importantes no comportamento social. Estudos indicam que a oxitocina pode atuar na regulação do apetite e do metabolismo energético, e há evidências de que a administração desse hormônio pode reduzir a ingestão alimentar e auxiliar no controle do peso (14).

Por fim, o gene *SHANK3* codifica uma proteína essencial para a formação e função das sinapses no cérebro e está fortemente associado a transtornos do neurodesenvolvimento, como o autismo, podendo impactar circuitos neurais ligados ao comportamento alimentar (20).

Já analisando os SNPs com maior valor de importância agregado, temos que o SNP rs7945201 não possui gene ou estudos associados. Porém, o SNP rs2284553 está no gene *IFNGR2* (Interferon Gamma Receptor 2), que codifica a subunidade beta do receptor de interferon-gama ($IFN-\gamma$). Esse receptor é essencial para a resposta imunológica, pois permite que as células respondam ao interferon-gama, uma citocina crucial na defesa contra infecções e no controle da inflamação (37).

5.4 Aplicação da floresta para circunferência de cintura

Após execução da floresta para a circunferência de cintura, os SNPs foram rankeados de acordo com sua pontuação de importância.

O gráfico Manhattan da Figura 5.2 demonstra a distribuição dos SNPs nos cromossomo por sua respectiva pontuação de importância.

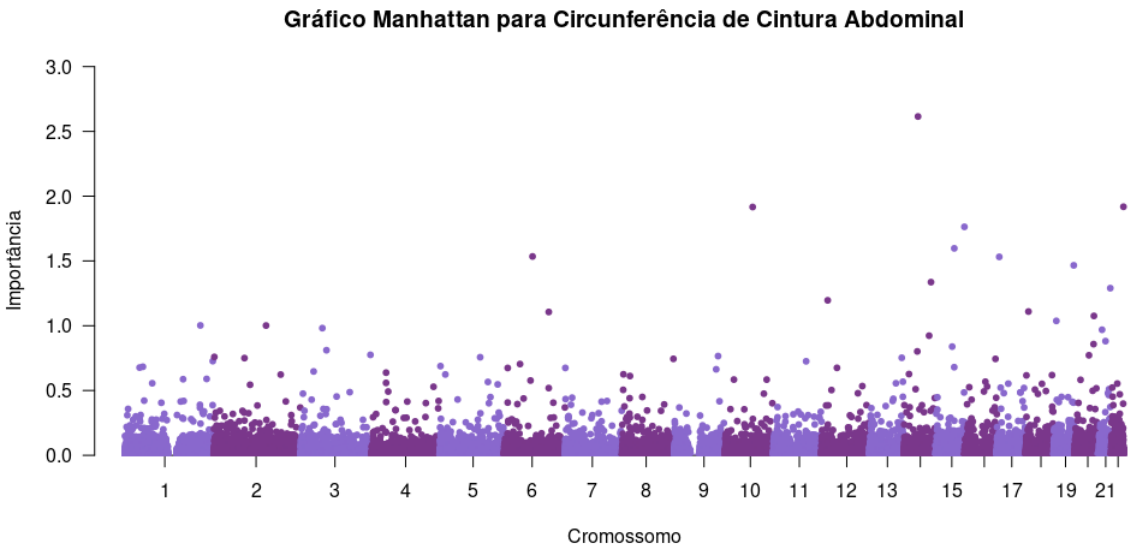

Figura 5.2: Gráfico Manhattan para a variável circunferência de cintura abdominal.

Por meio do Gráfico Manhattan apresentado na Figura 5.2, podemos observar que a maioria dos SNPs se concentra abaixo de 1, com a média sendo 0,0078, mediana 0 e quantil 75% de 0,01.

Estabelecendo um corte nas 1% maiores importâncias encontradas (quantil 99%), obtemos 2445 SNPs.A Tabela 5.3 apresenta a distribuição das pontuações de importância obtidas para o quantil 99% selecionado.

Tabela 5.3: Medidas Resumo para os SNPs selecionados no quantil 99% para circunferência de cintura abdominal.

Mínimo	**1º Quartil (Q1)**	**Mediana**	**Média**	**3º Quartil (Q3)**	**Máximo**
0,09	0,11	0,14	0,18	0,19	3,04

A partir deste quantil selecionado, podemos destacar 5 SNPs encontrados em estudos prévios pelo PubMed (35), e o SNP com a maior importância observada, descritos na Tabela 5.4.

Tabela 5.4: Quantidade de artigos encontrados para cada SNP, cromossomo, importância e descrição associada.

SNP	Cromossomo	Importância	Artigos	Descrição
rs2251746	1	0,1	18	Gene FCER1A
rs4988235	2	0,14	78	Gene MCM6
rs187116	11	0,01	10	Gene CD44
rs12785878	11	0,17	84	Gene DHCR7
rs478582	18	0,15	9	Gene PTPN2
rs543659	16	3,04	-	Gene ZNF423

A partir da Tabela 5.4 podemos verificar alguns genes associados à circunferência de cintura abdominal, cada um com um papel específico em processos metabólicos e inflamatórios.

O gene *FCER1A* codifica a cadeia alfa do receptor Fc de IgE. Ele é fundamental nas reações alérgicas e na ativação de células imunes, como os mastócitos. A alteração nesse gene pode estar relacionada a distúrbios alérgicos. Estudos avaliam a correlação entre obesidade e altos níveis de IgE (26).

O gene *MCM6* está relacionado à lactase, que é a enzima responsável pela quebra da lactose no intestino, sendo uma característica importante para a digestão de leite e produtos lácteos. Alguns estudos mostraram associações específicas com obesidade e suas comorbidades relacionadas (18).

O gene *CD44* codifica uma glicoproteína envolvida na adesão celular e na migração celular, sendo importante para processos inflamatórios e o sistema imunológico. Evidências recentes sugerem um papel desse gene no metabolismo, especialmente na resistência à insulina na obesidade e diabetes (36).

O gene *DHCR7* é responsável pela conversão de 7-deidrocolesterol em colesterol e também desempenha um papel crucial na síntese de vitamina D. Deficiências nesse gene estão associadas à síndrome de Smith-Lemli-Opitz e a problemas no metabolismo de colesterol e vitamina D (22).

O gene *PTPN2* codifica a proteína tirosina fosfatase não receptora tipo 2, e está envolvido em vias de sinalização que regulam a função imunológica e a inflamação. Alterações genéticas nesse gene têm sido associadas a diversas condições autoimunes, com variantes do gene comprometendo a função da barreira epitelial intestinal e aumentando

o risco de distúrbios inflamatórios (31).

O gene *ZNF423* codifica uma proteína nuclear que pertence à família de proteínas de dedo de zinco, codificando um fator de transcrição envolvido em diversos processos celulares, incluindo desenvolvimento neural e adipogênese. Pesquisas sugerem que modificações epigenéticas nesse gene podem afetar a capacidade das células precursoras de se diferenciarem em adipócitos, contribuindo para o desenvolvimento da obesidade hipertrófica (17).

Capítulo 6

Conclusão

Este capítulo discute os resultados encontrados, com a conclusão do trabalho e sugestões para estudos futuros que possam existir.

6.1 Discussão e conclusão

Os resultados deste estudo reforçam a complexidade da genética da obesidade, evidenciando que múltiplos marcadores genéticos desempenham importância na predisposição à doença. A utilização de florestas aleatórias mostrou-se uma abordagem promissora para a identificação de marcadores genéticos, permitindo captar interações não lineares e identificar SNPs com impacto significativo à obesidade e circunferência de cintura abdominal.

Muitos dos genes identificados estão relacionados a processos metabólicos e inflamatórios, destacando a influência da genética na regulação do balanço energético, resposta inflamatória e metabolismo. Isso sugere que a obesidade não é apenas influenciada pela ingestão calórica e gasto energético, mas também predisposições genéticas que podem afetar a forma como o organismo responde a fatores externos, como dieta e estilo de vida.

Dentre os achados, destaca-se o SNP rs478582, identificado tanto na análise da obesidade quanto na circunferência da cintura. Esse marcador genético está localizado no gene *PTPN2*, que desempenha um papel fundamental na regulação da inflamação e do metabolismo energético. A recorrência desse SNP em ambas as análises sugere sua relevância como um possível marcador genético, evidenciando sua potencial influência em vias metabólicas e processos inflamatórios relacionados à obesidade.

Destacam-se também dois SNPs com grandes valores de importância associados. O gene IFNGR2, que demostrou grande importância com obesidade, está envolvido na mo-

dulação da resposta imune e inflamatória, indicando que a variante desse gene pode influenciar o desenvolvimento de condições inflamatórias crônicas como a obesidade. Já o gene ZNF423, que demostrou grande importância com circunferência de cintura, regula processos biológicos relacionados ao metabolismo e à adipogênese, com variante nesse gene alterando a forma como os adipócitos armazenam gordura e respondem a estímulos inflamatórios.

Além disso, foram identificados marcadores genéticos associados a características psicológicas, como comportamento alimentar. Esses marcadores genéticos sugerem que a predisposição genética para obesidade pode não estar apenas relacionada a mecanismos fisiológicos, mas também fatores neurocomportamentais. Assim, uma abordagem integrada envolvendo genética, neurociência e psicologia no estudo da obesidade poderia trazer ganhos substanciais.

É importante ressaltar que este trabalho é de foco inferencial e não preditivo, com objetivo de identificar associações genéticas, e não na construção de modelos para previsão. Dessa forma, características comuns em abordagens preditivas, como a divisão do banco de dados em conjuntos de treino e teste, balanceamento das classes e otimização de métricas de desempenho, não foram de interesse neste trabalho. O objetivo do trabalho foi a análise das relações estatísticas entre marcadores genéticos e obesidade, contribuindo assim para o conhecimento da influência do genoma na predisposição à essa condição.

Também é importante ressaltar que critérios mais rígidos ou mais flexíveis do que o adotado neste trabalho para seleção dos marcadores genéticos trarão diferentes resultados. O critério utilizado neste trabalho foi executado de forma a ranquear os marcadores por sua respectiva pontuação de importância, com foco em marcadores que já possuíam estudo prévio.

Por fim, os resultados obtidos neste trabalho reforçam a importância de estudos adicionais para aprofundar a compreensão dos genes, em especial do PTPN2, IFNGR2 e ZNF423, analisando suas interações com fatores ambientais e outros componentes genéticos. Além disso, futuras pesquisas podem explorar como as variantes desses genes interagem com fatores ambientais e outros componentes genéticos.

Referências Bibliográficas

[1] Alves AAC, da Costa RM, Fonseca LFS, Carvalheiro R, Ventura RV, Rosa GJM, and Albuquerque LG. A random forest-based genome-wide scan reveals fertility-related candidate genes and potential inter-chromosomal epistatic regions associated with age at first calving in nellore cattle. https://pubmed.ncbi.nlm.nih.gov/35692843/, 2022.

[2] Alberts B, Johnson A, Lewis J, and et al. Molecular biology of the cell. 4th edition. https://www.ncbi.nlm.nih.gov/books/NBK26821/, 2002.

[3] Anthony J. Brookes. The essence of snps. https://doi.org/10.1016/S0378-1119(99)00219-X, 1999.

[4] Barbara Calabrese. Linkage disequilibrium. https://www.sciencedirect.com/science/article/pii/B9780128096338202343, 2019.

[5] M.C. Campbell, B. Ashong, S. Teng, et al. Multiple selective sweeps of ancient polymorphisms in and around lt located in the mhc class iii region on chromosome 6. *BMC Evolutionary Biology*, 19(218), 2019.

[6] Instituto Oswaldo Cruz. Fatores genéticos da obesidade. https://www.ioc.fiocruz.br/noticias/fatores-geneticos-da-obesidade, 2021.

[7] Gilderlanio Santana de Araújo. Uso de random forests e redes biológicas na associação de polimorfismos à doença de alzheimer. https://repositorio.ufpe.br/bitstream/123456789/18012/1/Dissertacao%20-Gilderlanio%20Santana%20de%20Araujo.pdf, 2013.

[8] Sociedade Brasileira de Cardiologia. I diretriz brasileira de diagnóstico e tratamento de síndrome metabólica. http://departamentos.cardiol.br/DECAGE/esquina/superligado/Diretriz%20Bras%20SM%2005.pdf.

[9] Rita de Cássia Borges de Castro. O que são polimorfismos e qual a sua relação com a nutrigenética? https://nutritotal.com.br/pro/o-que-sa-o-polimorfismos-e-qual-a-sua-relaa-a-o-com-a-nutrigena-tica/, 2013.

[10] Fortran original by Leo Breiman and Adele Cutler, R port by Andy Liaw and Matthew Wiener. *Breiman and Cutler's Random Forests for Classification and Regression*, 2022.

[11] D.Y. Hu, X.X. Shao, C.L. Xu, S.L. Xia, L.Q. Yu, L.J. Jiang, J. Jin, X.Q. Lin, and Y. Jiang. Associations of fut2 and fut3 gene polymorphisms with crohn's disease in chinese patients. *Journal of Gastroenterology and Hepatology*, 29(10):1778–1785, 2014.

[12] EMBL's European Bioinformatics Institute. What are genome wide association studies (gwas)? https://www.ebi.ac.uk/training/online/courses/gwas-catalogue-exploring-snp-trait-associations/what-is-gwas-catalog/what-are-genome-wide-association-studies-gwas/.

[13] Rafael Izbicki and Tiago Mendonça dos Santos. *Aprendizado de máquina: uma abordagem estatística*. 2020.

[14] E.A. Lawson. The effects of oxytocin on eating behavior and metabolism in humans. *Nature Reviews Endocrinology*, 13(12):700–709, 2017.

[15] Andy Liaw and Matthew Wiener. randomforest: Breiman and cutler's random forests for classification and regression. https://cran.r-project.org/web/packages/randomForest/randomForest.pdf, 2002. R package version 4.7-1.

[16] K.A. Longo, W.S. Wright, S. Kang, I. Gerin, S.H. Chiang, P.C. Lucas, M.R. Opp, and O.A. MacDougald. Wnt10b inhibits development of white and brown adipose tissues. *Journal of Biological Chemistry*, 279(34):35503–35509, 2004.

[17] M. Longo, G. A. Raciti, F. Zatterale, et al. Modificações epigenéticas do gene zfp / znf423 controlam o comprometimento adipogênico murino e são desreguladas na obesidade hipertrófica humana. *Diabetologia*, 61:369–380, 2018.

[18] D. A. Luis, O. Izaola, and D. Primo. The lactase rs4988235 is associated with obesity related variables and diabetes mellitus in menopausal obese females. *Eur Rev Med Pharmacol Sci*, 25(2):932–940, 2021.

[19] Marvin N. Wright [aut, cre] and Stefan Wager [ctb] and Philipp Probst [ctb]. *A Fast Implementation of Random Forests*, 2023.

[20] F. Mashayekhi, N. Mizban, E. Bidabadi, and Z. Salehi. The association of shank3 gene polymorphism and autism. *Minerva Pediatrica (Torino)*, 73(3):251–255, 2021.

[21] Alves MCGP, Escuder MML, Goldbaum M, Barros MBA, Fisberg RM, and Cesar CLG. Sampling plan in health surveys, city of são paulo, brazil, 2015. https://www.revistas.usp.br/rsp/article/view/149639/146701, 2018.

[22] S. Miyazaki, N. Shimizu, H. Miyahara, H. Teranishi, R. Umeda, S. Yano, T. Shimada, H. Shiraishi, K. Komiya, A. Katoh, A. Yoshimura, R. Hanada, and T. Hanada. Dhcr7 links cholesterol synthesis with neuronal development and axonal integrity. *Biochemical and Biophysical Research Communications*, 712-713:149932, Jun 18 2024.

[23] Revista Saúde News. Farmacogenética: o uso do dna nos tratamentos personalizados em psiquiatria e neurologia. https://www.revistasaudenews.com.br/post/416/farmacogenetica:-o-uso-do-dna-nos-tratamentos-personalizados-em-psiquiatria-e-neurologia.

[24] NHI. National human genome research institute. https://www.genome.gov/genetics-glossary/Deoxyribonucleic-Acid/, 2024.

[25] Jaqueline L. Pereira, Camila A. de Souza, Jennyfer E. M. Neyra, Jean M. R. S. Leite, Andressa Cerqueira, Regina C. Mingroni-Netto, Julia M. P. Soler, Marcelo M. Rogero, Flavia M. Sarti, and Regina M. Fisberg. Genetic ancestry and self-reported "skin color/race" in the urban admixed population of são paulo city, brazil. *Genes*, 15(7), 2024.

[26] F. C. M. Pinheiro, M. A. C. N. Silva, H. C. M. Pinheiro, and J. D. M. Pinheiro. Correlação entre o índice de massa corporal (imc) e os níveis de ige total em indivíduos asmáticos de um programa estruturado de asma em são luís - ma. *Research, Society and Development*, 11(8):e17711830464, 2022.

[27] S. Popadic, Z. Ramic, Lj. Medenica, V. Pravica, and D. Popadic. Il-23r gene polymorphism rs2201841 is associated with psoriatic arthritis. *International Journal of Immunogenetics*, 41(4):335–337, 2014.

[28] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2019.

[29] Bush W. S. and Moore J. H. Chapter 11: Genome-wide association studies. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3531285/, 2012.

[30] B. Shen, R. Lin, C.C. Wang, J. Rei, Y. Sun, Y.L. Yang, and Y.Y. Lin. Adam33 gene polymorphisms identified to be associated with asthma in a chinese li population. *Biomedicine Pharmacotherapy*, 6(3):323–328, 2017.

[31] Markus R. Spalinger, Angie Sayoc-Becerra, Andrea N. Santos, Ali Shawki, Valentina Canale, Mahesh Krishnan, Anna Niechcial, Nnenna Obialo, Michael Scharl, Jun Li, Meher G. Nair, and Declan F. McCole. Ptpn2 regulates interactions between macrophages and intestinal epithelial cells to promote intestinal barrier function. *Gastroenterology*, 159(5):1763–1777.e14, Nov 2020.

[32] M. Stephens and S. Fish. Imputation of missing genotypes in genetic association studies. *American Journal of Human Genetics*, 62(2):545–552, 1998.

[33] J.F. Sánchez Muñoz-Torrero, M.D. Rivas, J. Zamorano, R. Alonso, P. Joya-Vázquez, T. Padró, and P. Mata. rs1801275 interleukin-4 receptor alpha polymorphism in familial hypercholesterolemia. *Journal of Clinical Lipidology*, 8(4):418–422, 2014.

[34] C.L. Too, S. Murad, J.S. Dhaliwal, et al. Polymorphisms in peptidylarginine deiminase associate with rheumatoid arthritis in diverse asian populations: evidence from myeira study and meta-analysis. *Arthritis Research & Therapy*, 14(R250), 2012.

[35] U.S. National Library of Medicine. Pubmed, 2024.

[36] X. Weng, S. Maxwell-Warburton, A. Hasib, L. Ma, and L. Kang. The membrane receptor cd44: novel insights into metabolism. *Trends in Endocrinology and Metabolism*, 33(5):318–332, May 2022.

[37] Y. Xia, Q. Zhang, Y. Ye, X. Wu, F. He, Y. Peng, Y. Yin, and W. Ren. Melatonergic signalling instructs transcriptional inhibition of ifngr2 to lessen interleukin-1-dependent inflammation. *Clinical and Translational Medicine*, 12(2):e716, Feb 2022.