Trabalho de Formatura

Modelagem Matemática de GWAS para Análise de Dados Genômicos

Beatriz Campanha Silva

Orientador: Diego Luiz Rovaris (ICB - USP)
Coorientadora: Cláudia Monteiro Peixoto (IME - USP)

Universidade de São Paulo
Instituto de Matemática e Estatística
Departamento de Matemática Aplicada

Fevereiro de 2025

Resumo

Estudos de associação por varredura genômica (GWAS) são capazes de identificar variações genéticas de polimorfismos de nucleotídeo único (SNPs) que possuem alguma associação significativa a uma doença em particular ou característica a ser estudada. Nesse método de associações do GWAS, utiliza-se o Modelo Biométrico, que fornece a base teórica sobre a distribuição estatística dos genótipos e dos efeitos na população. Assim, são formuladas hipóteses sobre a associação entre os genótipos (AA, Aa e aa) e as variáveis contínuas e binárias (e.g., nível de colesterol), avaliadas por regressão linear e logística, sob algumas suposições. Para isso, o genoma completo de voluntários é genotipado e são analisadas centenas, milhares ou milhões de variantes genéticas, utilizando-se também consórcios de genomas. A saída primária de uma análise de GWAS consiste de listas de tamanhos de efeitos estimados, os erros dessas estimativas e valores P, resultantes dos testes de associação, sobre as variantes identificadas. Um valor P indica se determinado SNP é significante para a expressão da característica alvo do teste, enquanto o efeito indica sua intensidade. Pela replicação do processo, são originados métodos pós-GWAS, capazes de estabelecer pontuações de risco para um indivíduo, conhecidas como escores de risco poligênico ou escores poligênicos, além de aplicar correlações locais e globais dentre diferentes características estudadas. Neste trabalho, o foco é o estudo da modelagem dos métodos GWAS e pós-GWAS (de Escore Poligênico, Correlação Global e Correlação Local), para pormenorizar seus respectivos usos e limitações. Além disso, é apresentada uma aplicação prática dos métodos pós-GWAS a partir de dados sumários de GWAS sobre TDAH (transtorno de déficit de atenção com hiperatividade), hipotireoidismo e níveis de TSH (hormônio tireoestimulante).

Palavras-chave: GWAS, estudos de associação por varredura genômica, modelagem, estatística, correlação, regressão, hipóteses.

Abstract

Genome-wide association studies (GWAS) are capable of identifying single nucleotide polymorphisms (SNPs), which are genetic variations that have some significant association with a particular disease or trait of interest. In the GWAS association method, the Biometric Model is used, which provides the theoretical basis for the statistical distribution of genotypes and their effects in the population. Thus, hypotheses are formulated about the association between genotypes (AA, Aa and aa) and continuous or binary variables (e.g., cholesterol levels), evaluated through linear and logistic regression, under some assumptions. For this purpose, the complete genome of volunteers is genotyped, and hundreds of thousands to millions of genetic variants are analyzed, also using genomic consortia. The primary output of a GWAS analysis is a list of estimated effect sizes, their errors, and P-values, derived from association tests, regarding the identified variants. The P-value indicates if a specific SNP is significant for the expression of the target trait, while the effect size indicates its intensity. Through the replication of this process, post-GWAS methods have emerged, enabling the calculation of risk scores for individuals, known as polygenic risk scores or polygenic scores, as well as the application of local and global correlations among different studied traits. In this final project, the focus is on the study of modeling GWAS and post-GWAS methods (Polygenic Score, Global Correlation, and Local Correlation) to detail their respective applications and limitations. Furthermore, a practical application of post-GWAS methods is presented from GWAS summary data on TSH (thyroid-stimulating hormone) levels, hypothyroidism, and ADHD (attention-deficit/hyperactivity disorder).

Keywords: GWAS, genome-wide association studies, model, statistics, correlation, regression, hypothesis.

Sumário

Glossário

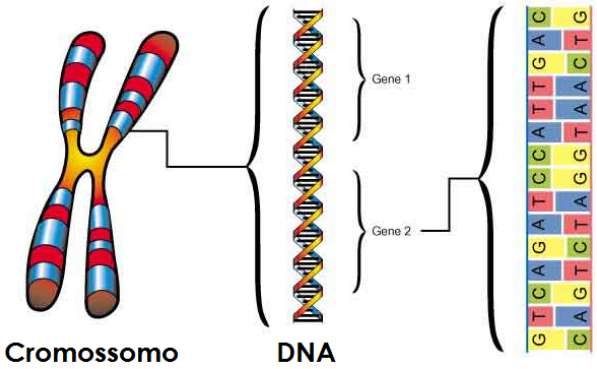

Figura 1: Imagem de suporte ao glossário. Fonte: Plant and Soil Sciences eLibrary.

Ácido desoxirribonucleico (DNA): é um polímero composto por duas cadeias polinucleotídicas, enroladas uma sobre a outra em forma de dupla hélice. O DNA carrega instruções genéticas para o desenvolvimento, funcionamento e reprodução de todos os organismos. A sua principal função é armazenar as informações necessárias para a transcrição em RNA e a construção de proteínas.

Sequência de DNA: é uma série de nucleotídeos que formam a estrutura primária de uma molécula ou cadeia de DNA, representados pelas letras A, C, G e T que correspondem aos quatro nucleotídeos de uma cadeia de DNA: as bases Adenina, Citosina, Guanina e Timina, respectivamente.

Gene: é a unidade fundamental da hereditariedade. Cada gene é formado por uma sequência de DNA que codifica uma proteína ou molécula de RNA.

Cromossomo: é a molécula de DNA condensada, onde cada região corresponde a um gene, e cada posição mapeia uma base do DNA. É a estrutura responsável por transmitir os genes ao longo das gerações. Em humanos, através de gametas que contém 23 cromossomos cada, é formado um indivíduo com 23 pares de cromossomos.

Genoma: é a sequência completa de DNA de um organismo, ou seja, o conjunto de todos os cromossomos de um ser vivo.

Locus: é uma palavra de origem latina que significa uma região, no caso, genômica. Nesse trabalho, considera-se como *locus* a menor região possível, que é exatamente uma base de DNA. Seu plural é *loci*.

Polimorfismo de nucleotídeo único (SNP): é uma variação na sequência de DNA que afeta somente uma base (nucleotídeo).

Alelos: são as formas alternativas de um determinado gene, e ocupam uma mesma posição (locus) em pares de cromossomos homólogos.

Desequilíbrio de ligação (LD): é medido pela associação não aleatória de alelos em diferentes loci. Descreve uma situação em que algumas combinações de alelos ocorrem mais ou menos frequentemente numa população do que era esperado pela segregação aleatória, sendo influenciado principalmente pela distância física entre os genes no cromossomo, que permanecem fisicamente juntos quando acontece a meiose.

Fenótipo: é um traço observável ou que pode ser mensurado de um indivíduo, como características físicas, bioquímicas, fisiológicas e comportamentais.

Genótipo: é a composição genética formada pelo conjunto de alelos presentes em um indivíduo.

Efeito genotípico: quantifica o efeito que uma variante genética tem em um determinado fenótipo.

Pleiotropia: é o conjunto de múltiplos efeitos de um gene.

Lista de símbolos

$\in$: Pertence à.

$\mathbb{N}$: Conjunto dos números Naturais ($\{0, 1, 2, \dots\}$).

$\mathbb{R}$: Conjunto dos números Reais.

$\mathbb{R}_+$: Conjunto dos números Reais positivos.

$\sim$: Segue uma distribuição.

Bin: Distribuição Binomial.

Bernoulli: Distribuição Bernoulli.

N: Distribuição Normal.

χ^2: Distribuição Qui-Quadrado.

t: Distribuição t de Student

E: Função de esperança.

V: Função de variância.

α: Letra grega *alpha*.

β: Letra grega *beta*.

θ: Letra grega *theta*.

μ: Letra grega *mu*.

σ: Letra grega *sigma*.

ϵ: Letra grega *epsilon*.

ψ: Letra grega *psi*.

ρ: Letra grega *rho*.

$\sum_{i=1}^{n}$: Somatório para $i = 1, 2, \dots, n$.

$\approx$: Aproximadamente.

$\neq$: Diferente.

1 Introdução

A metodologia de Estudos de Associação por Varredura Genômica, ou *Genome-Wide Association Studies* (GWAS) consiste no estudo sobre o efeito de variações genéticas de polimorfismos de nucleotídeo único (SNPs) em diversos fenótipos de interesse, como patologias, distúrbios ou níveis hormonais, por meio de testes estatísticos de associação. Assim, o GWAS se destaca como uma importante ferramenta para a identificação de variantes genéticas associadas à expressão de uma doença, condição ou característica individual, que seja mensurável qualitativa ou quantitativamente (Pereira Ciochetti et al. 2023).

Fundamentada pela Inferência Estatística Frequentista, essa relação causal é avaliada por meio de testes de associação entre genótipo e fenótipo. Uma variável quantifica o número de indivíduos que possuem cada um dos três genótipos possíveis: ausência da variante genética (0), presença de uma variante genética em um alelo (1), variante genética em ambos os alelos (2); a outra variável diz respeito a um fenótipo (por exemplo, altura em centímetros, nível de colesterol, diagnóstico de bipolaridade, etc.). Para cada par de variáveis, genética e fenotípica, é realizado um teste sob a hipótese de que **não há** efeito genotípico, e avalia-se se há evidências amostrais suficientes para rejeitar essa hipótese. Caso haja evidência para rejeitar a hipótese, essa variante genética pode exercer influência na expressão da doença ou condição estudada.

Os indivíduos sobre os quais os dados são coletados são alocados em um grupo alvo *versus* um grupo controle, nos quais os indivíduos possuem ou não a característica de interesse, ou em níveis diferentes de um fenótipo contínuo.

A metodologia GWAS consiste em duas macro-etapas esquematizadas na Figura 2: um fluxo de trabalho em laboratório e, posteriormente, um fluxo de bioinformática.

A partir de um plano amostral, são selecionados indivíduos e coletadas suas amostras de extrato de DNA para sequenciamento ou genotipagem por *biochips* em laboratório. Os genomas são armazenados em arquivos de texto e binários, e encaminhados para o fluxo de bioinformática no qual os dados serão tratados. Os dados passam por um controle de qualidade, que visa remover unidades amostrais com problemas (isto é, que apresentam SNPs mal genotipados, DNA mal isolado, indivíduos aparentados, disparidade entre o sexo biológico informado e o coletado), e é realizada uma repartição desses dados em subgrupos homogêneos por ancestralidade, chamada de estratificação populacional, caso necessário. Em seguida, na etapa de Imputação, essa base é preenchida por referência à bancos de dados públicos em consórcios de genomas. Por fim, os dados tratados são submetidos às analises de associação por meio de ferramentas computacionais, como o PLINK (disponível em `https://zzz.bwh.harvard.edu/plink/`) e o RICOPILI (disponível em `https://sites.google.com/a/broadinstitute.org/ricopili/`).

Dessa última etapa, são obtidos os resultados da análise primária, compostos por

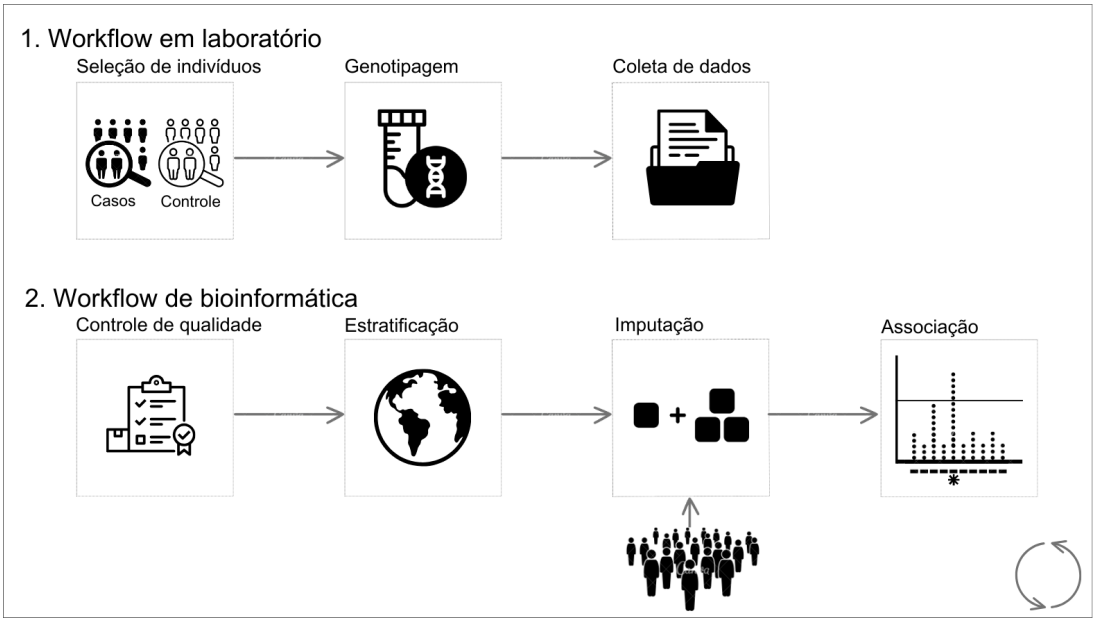

Figura 2: Visão geral da Metodologia GWAS

uma lista de SNPs e estimativas associadas. O valor P indica a probabilidade de que a associação observada entre uma variante genética e uma característica ou doença seja devida ao acaso, isto é, indica a significância da associação sob a hipótese de que o efeito daquele SNP é nulo. Comumente, os resultados são resumidos ao plotar o gráfico Manhattan (Figura 3) que relaciona a posição cromossômica do SNP a seu valor P, em comparação com um limiar de significância pré-estabelecido. Assim, as variantes nas posições cromossômicas superiores à linha tracejada rejeitam a hipótese de efeito nulo e, portanto, são consideradas significativas.

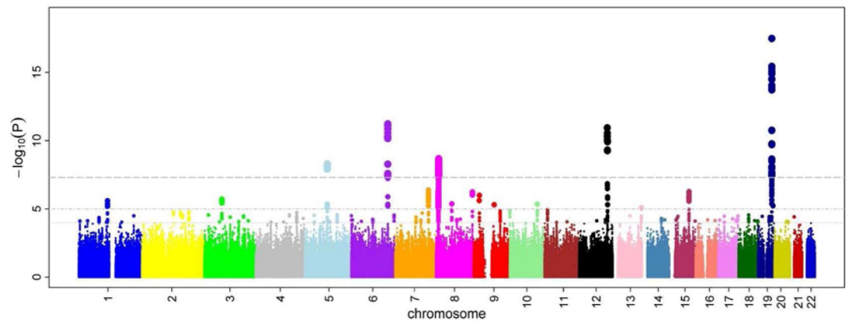

Figura 3: Exemplo de gráfico Manhattan para uma análise primária em GWAS. Fonte: Ikram MK et al. (2010), disponível em `https://doi.org/10.1371/journal.pgen.1001184`.

Além disso, como resultado das análises de associação, também é obtida uma estimativa $\hat{\beta}$ que quantifica o efeito do SNP no fenótipo. Ela é a estimativa do coeficiente β da Regressão Linear entre genótipo e fenótipos contínuos, ou calculada pela razão de *odds* na Regressão Logística entre genótipo e fenótipos binários.

Os resultados obtidos a partir desse método têm contribuído significativamente na área

de saúde geral e mental e são úteis para revelar o controle genético por trás de fenótipos complexos e condições patológicas (Pereira Ciochetti et al. 2023). É possível interpretar a lista de SNPs associados, juntamente com um estudo fisiológico e patofisiológico, e determinar uma anotação funcional a cada variante genética mais provável de ser causal e sua possível convergência em vias biológicas, como, por exemplo, em Shah et al. 2020 (estudo sobre insuficiência cardíaca), Presumey, Bialas e Carroll 2017 (sinapse neural em esquizofrenia), Corvol et al. 2015 (doença pulmonar em fibrose cística).

Após esse processo, por meio dos resultados sumários dos testes estatísticos, ainda se pode construir análises posteriores, conhecidas como pós-GWAS.

Em resumo, a partir desse estudo fundamental, substancialmente estatístico e computacional – pelos testes de associação e pelo uso de ferramentas como o PLINK para ser executado – são obtidas as localidades genômicas envolvidas em modular o fenótipo de interesse com seus respectivos efeitos, e se pode aprofundar na via biológica do mecanismo que causa a doença. Ainda, é possível explorar a modulação da gravidade em doenças monogênicas e a resposta a tratamentos fármacos.

1.1 Objetivos

O objetivo deste trabalho é entender os métodos usados em GWAS do ponto de vista de modelagem matemática, elencando seus respectivos usos e limitações, para ter uma melhor compreensão das abordagens estatísticas e computacionais utilizadas no desenvolvimento dessa ferramenta, que tem se mostrado importante para o estudo da influência genética na saúde humana.

Primeiramente, será apresentada a fundamentação teórica sobre Genética e, em seguida, conceitos fundamentais da Inferência Estatística, que inclui especialmente os temas de Estimação de Parâmetros, Teste de Hipóteses e Regressão Linear e Logística, aplicados no método de associações em GWAS. Em seguida, com base em uma revisão bibliográfica, é detalhada a estratégia para a modelagem matemática de GWAS, além de algumas análises posteriores feitas com as estatísticas sumárias, o procedimento de tratamento dos dados e a organização dos arquivos. Ainda, é disponibilizado relatório, com o passo a passo, resultados e conclusões das análises pós-GWAS executadas com o suporte do Laboratório de Fisiologia Genômica da Saúde Mental do ICB - USP, para uma aplicação prática dos métodos estudados. Por fim, são feitas discussões relevantes sobre o modelo e suas hipóteses, que levam à conclusão do trabalho.

2 Fundamentação Teórica

2.1 Genética

O cariograma (Figuras 4 e 5) é uma ferramenta fundamental na citogenética – a ciência que estuda a estrutura e função dos cromossomos. Trata-se de uma representação gráfica de um cariótipo, a coleção completa de cromossomos de uma célula, organizados em pares homólogos e dispostos por tamanho e forma.

Cada espécie tem um número característico de cromossomos. Em especial, os humanos possuem 46 cromossomos, organizados em 23 pares. Cada par é composto de um cromossomo herdado do pai e outro da mãe, ditos homólogos.

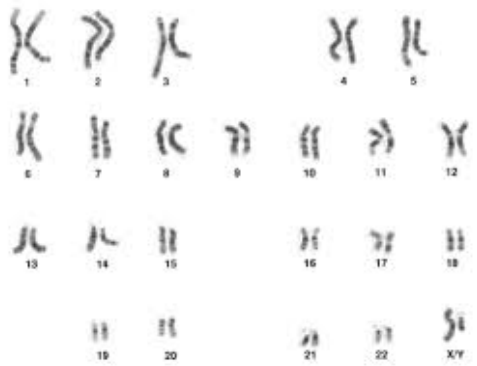

Figura 4: Cariograma. Representação fotográfica dos 23 pares de cromossomos de um humano.

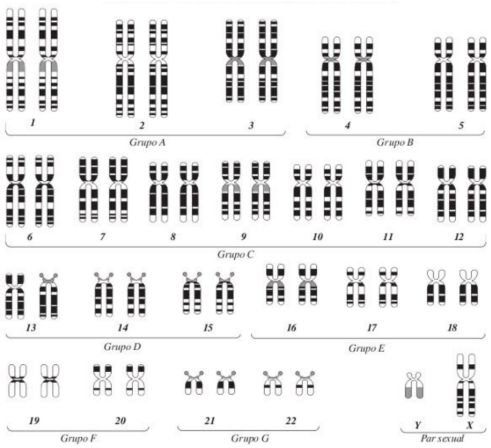

Figura 5: Representação esquematizada dos cromossomos, em que as "listras", ou bandas, indicam diferentes regiões cromossômicas.

Os cromossomos (Figura 6) são estruturas organizadas e compactas, compostas principalmente por DNA – a molécula que carrega a informação genética. O DNA é composto por duas cadeias de nucleotídeos que formam uma dupla hélice. Cada nucleotídeo é composto por uma base nitrogenada, sendo elas adenina (A), timina (T), citosina (C) ou guanina (G), que emparelham-se de forma específica, A-T e C-G.

Os genes, as unidades fundamentais da hereditariedade, são segmentos de DNA localizados nos cromossomos (Figura 7), e cada gene pode existir em diferentes formas, chamadas de alelos. Para cada ponto polimórfico em um gene, um indivíduo possui dois alelos, um herdado de cada progenitor.

A expressão gênica se dá por meio da síntese de proteínas. Nesse processo, a partir da sequência de DNA, a cada trinca de nucleotídeos, é associado um aminoácido. Esse conjunto de aminoácidos correspondentes à sequência formam, então, uma cadeia polipeptídica que se dobra em uma proteína funcional.

As proteínas desempenham uma ampla gama de funções biológicas: estruturais, en-

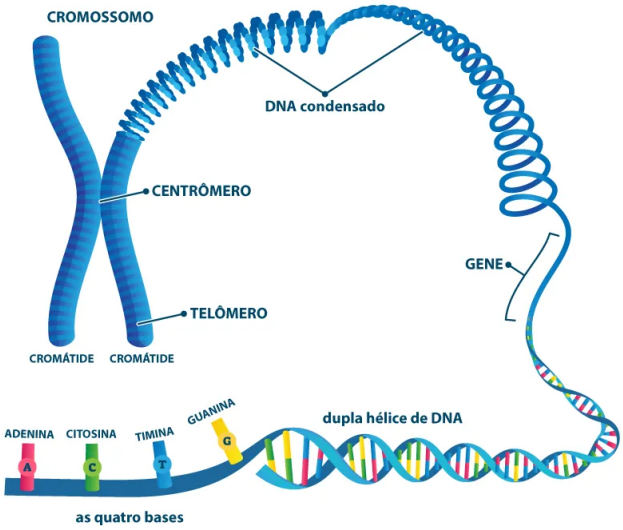

Figura 6: Superestrutura de um cromossomo. Fonte: Brasil Escola.

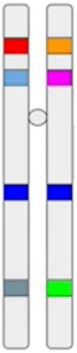

Figura 7: Um cromossomo, com alguns genes destacados em colorido.

zimáticas, regulatórias, defensivas e de transporte. Proteínas como colágeno e queratina formam a estrutura dos tecidos e órgãos, enzimas aceleram reações químicas vitais para o metabolismo, hormônios regulam processos fisiológicos, anticorpos protegem o organismo contra patógenos, hemoglobinas transportam moléculas essenciais como oxigênio.

Assim, nota-se como a variação genética é essencial para a diversidade biológica. Concomitantemente, mutações ou alterações no DNA podem ter efeitos variados, desde uma ausência de impacto, a variedade de características biológicas observáveis, até a causa de doenças graves. Atualmente, as variantes são classificadas seguindo as diretrizes internacionais do Colégio Americano de Genética e Genômica (ACMG), dentre benignas à patogênicas.

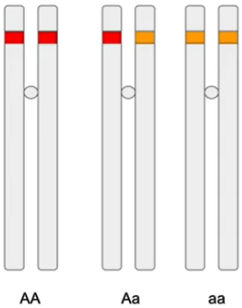

Figura 8: Alelos

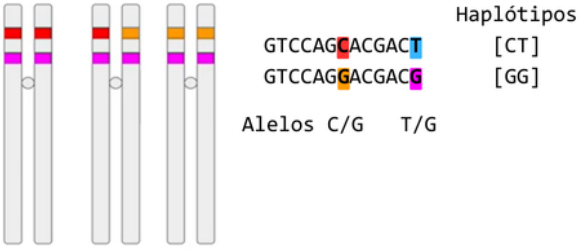

Figura 9: Haplótipos

Nesse sentido, torna-se evidente a importância do estudo dos Polimorfismos de Nucleotídeo Único (SNPs, do inglês *Single Nucleotide Polymorphisms*). Os SNPs consistem na alteração de um único nucleotídeo em relação a um genoma de referência. Essa alteração ocorre na linhagem de células germinativas, o que resulta na segregação dessa

variação dentro de uma população de organismos.

Para ser considerada um SNP, a variante deve estar presente em pelo menos 1% da população (Auton et al. 2015). Para ilustrar, em uma posição específica do genoma humano, a maioria da população pode apresentar o nucleotídeo A, enquanto a minoria apresenta C. Tal variação caracteriza a presença de um SNP nessa localização específica do genoma humano e, nesse caso, as duas variantes possíveis são denominadas alelos para essa posição (Figura 8). Além disso, um conjunto de SNPs no mesmo cromossomo é denominado haplótipo (Figura 9).

A maior parte de SNPs identificados por GWAS são localizados em regiões não codificadoras do genoma, mas eles servem como representativos para todos os SNPs no mesmo bloco de haplótipo. Uma hipótese é de que eles desempenham papel regulatório e podem causar mudanças na expressão gênica, mas não necessariamente na função proteica (Tak e Farnham 2015), pois há vários elementos envolvidos na regulação transcricional, como promotores, intensificadores e elementos estruturais nucleares que controlam a expressão de um gene. Pesquisadores estão fazendo progresso em compreender a relevância desses SNPs não codificantes e em como eles podem influenciar no risco de doenças.

Diante disso, é necessário ressaltar que o SNP descoberto por GWAS de uma determinada região não é necessariamente o causal, mas pode ser um marcador para ele.

No contexto de análise estatística, se faz importante evidenciar que o genoma humano possui 3,2 bilhões de bases (Makałowski 2001). Dentre esse conjunto de bilhões de bases, um genoma comum pode diferir do genoma de referência em cerca de 4,1 a 5,0 milhões (Auton et al. 2015). Essas variações são o que difere cada ser humano do outro, e a maior diferença da quantidade de variantes é observada entre populações de ancestralidades distintas.

As variações são classificadas dependendo de sua frequência populacional. Mais de 99% das variantes existentes consistem de Variações de Nucleotídeo Único (SNVs). Dentre as quais 99,98% são raras (com uma frequência alélica menor que 1%), e as demais 0,02% são comuns, no caso, SNPs (Bick et al. 2024).

Quantidade de bases no genoma humano	3,2 bilhões
Quantidade de bases diferentes do referencial	4,1 a 5 milhões
Taxa de variantes que são comuns (SNPs)	$\approx$ 0,02%

Tabela 1: Resumo da quantificação de nucleotídeos e polimorfismos de nucleotídeo único do genoma humano

2.2 Inferência Estatística (aplicada a GWAS)

Um modelo estatístico procura representar, em termos matemáticos, a complexidade das relações que envolvem uma população estudada. Para isso, podem ser usadas distribuições de probabilidade. Elas são definidas por um par de conjuntos $(\mathcal{X}, \mathcal{P})$, no qual $\mathcal{X}$

representa os valores possíveis para os dados – é o espaço amostral –, e $\mathcal{P}$ a distribuição de probabilidade sobre os dados.

Em virtude de muitos fenômenos observáveis seguirem padrões previsíveis de geração dos dados, foram caracterizadas distribuições conhecidas sobre o comportamento das variáveis aleatórias geradas. Algumas distribuições recorrentes em aplicações matemáticas são Bernoulli, Binomial, Uniforme e Normal, Poisson, Gamma, Qui-Quadrado e Beta. Em comum, toda distribuição é indexada por um parâmetro θ, que a especifica.

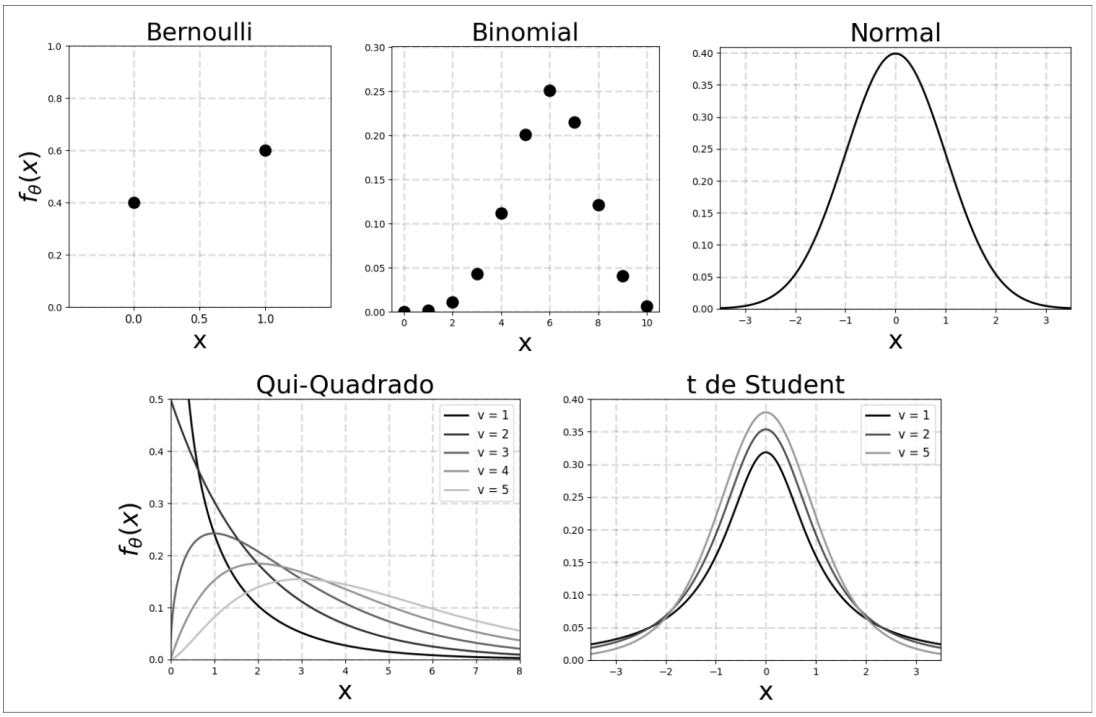

Figura 10: Representação gráfica das funções de probabilidade de diferentes distribuições amostrais: Bernoulli(0.6), Bin$(10, 0.6)$, N$(0, 1)$, χ^2(v) e t(v).

Por exemplo, ao testar um indivíduo para a presença de uma mutação genética, o resultado obtido será "ausente" ou "presente", em referência ao alelo mutado. Essa resposta binária segue uma distribuição conhecida chamada Bernoulli, e o parâmetro que a indexa é a probabilidade do evento.

Em um segundo exemplo, pode-se estudar a ocorrência de uma mutação genética em descendentes de um cruzamento mendeliano, que consiste numa repetição de ensaios de Bernoulli. A partir de dois indivíduos heterozigotos para um gene específico (Aa × Aa), a probabilidade de um descendente ser homozigoto recessivo (aa) é 25%. Cada descendente representa um ensaio independente, então, ao considerar 10 descendentes, a quantidade de indivíduos homozigotos recessivos nesse grupo segue uma distribuição Binomial, cujo parâmetro é (10, 25%), composto pelo número de descendentes, 10, e pela probabilidade de ser homozigoto, 25%.

Ainda, ao examinar a estatura humana – que é uma variável contínua – de um número suficientemente grande número de indivíduos, conclui-se que ela segue uma distribuição

Normal. Nessa distribuição, os valores se distribuem de forma simétrica, com a maior parte próxima da média e a menor progressivamente distante desta.

Essas variáveis seguem uma função de probabilidade, denotada por $f_\theta(x)$, que descreve a probabilidade de uma variável aleatória X assumir o valor x, sob determinada parametrização por θ (Figura 10). A seguir, elenco cinco distribuições conhecidas que merecem destaque no escopo deste trabalho, Bernoulli, Binomial, Normal, Qui-Quadrado e t de Student, com suas respectivas notações, funções de probabilidade e espaços paramétricos (onde os parâmetros estão definidos), de Morettin e O. Bussab 2017, além da sua utilidade em GWAS.

Distribuição de Bernoulli:

$$\begin{aligned} &X \sim \text{Bernoulli}(\theta): \\ &f_\theta(x) = \theta^x(1-\theta)^{1-x}, \\ &\theta \in (0,1),\ x \in \{0,1\} \end{aligned} \tag{1}$$

Essa é a distribuição que seguem as variáveis de resposta binária. Em GWAS, modela a variável de presença/ausência de determinada doença ou variante genética. Cada indivíduo pode possuir ($x = 1$) ou não ($x = 0$) um fenótipo de interesse.

Distribuição Binomial:

$$\begin{aligned} &X \sim \text{Bin}(n,p): \\ &f_{(n,p)}(x) = \binom{n}{x} p^x(1-p)^{n-x}, \\ &\theta = (n,p) \in \mathbb{N} \times (0,1) \end{aligned} \tag{2}$$

A distribuição Binomial em GWAS pode ser aplicada ao comparar a frequência de um alelo entre casos e controles, quando o fenótipo é binário. A contagem esperada de alelos em cada grupo genotípico segue essa distribuição.

Distribuição Normal:

$$\begin{aligned} &X \sim N(\mu,\sigma^2): \\ &f_{(\mu,\sigma^2)}(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \\ &\theta = (\mu,\sigma^2) \in \mathbb{R} \times \mathbb{R}_+ \end{aligned} \tag{3}$$

e a variável aleatória X pode ser normalizada da seguinte forma

$$\frac{X-\mu}{\sigma} \sim N(0,1). \tag{4}$$

Essa distribuição $N(0,1)$, com parâmetro de média zero e variância 1, é chamada de Normal Padrão.

Em GWAS, assume-se que fenótipos quantitativos, tratados como uma variável contínua, seguem distribuição Normal. Essa premissa é utilizada para realizar testes de associação usando Regressão Linear. Além disso, a normalidade também é assumida para o erro residual no modelo de Regressão Linear.

Distribuição Qui-Quadrado:

$$
\begin{aligned}
&X \sim \chi^2(\theta): \\
&f_\theta(x) = \begin{cases} \dfrac{1}{\Gamma(\theta/2)2^{\theta/2}} x^{\theta/2-1} e^{-y/2}, \ y > 0 \\ 0, \ y < 0 \end{cases} \\
&\text{em que } \Gamma(a) = (a-1)\Gamma(a-1) \text{ e } \Gamma(n) = (n-1)! \iff n \in \mathbb{N}, \\
&\theta \in \mathbb{N}
\end{aligned}
\tag{5}
$$

Vale notar que uma variável aleatória com distribuição Normal Padrão, ao quadrado, é uma v.a. com distribuição $\chi^2(1)$. Além disso, uma variável com distribuição Qui-Quadrado pode ser vista como a soma de θ Normais Padrão ao quadrado, independentes.

A Qui-Quadrado é utilizada como estatística de teste para avaliar se os valores observados diferem do esperado, considerando a distribuição amostral. Por exemplo, em GWAS, essa distribuição é aplicada para comparar a frequência genotípica entre grupos de casos e controles em fenótipos binários; ao avaliar o desvio do equilíbrio de Hardy-Weinberg (equilíbrio genético esperado em uma população) de um SNP; e em teste de heterogeneidade, para verificar se os efeitos do SNP são consistentes em diferentes populações.

Distribuição *t* de Student:

$$
\begin{aligned}
&X \sim \mathrm{t}(\theta): \\
&X = \frac{Z}{\sqrt{Y/\theta}}, \\
&\text{em que } Z \sim N(0,1), \ Y \sim \chi^2(\theta)
\end{aligned}
\tag{6}
$$

Essa distribuição aproxima-se de uma Normal Padrão quando θ é suficientemente grande (Morettin e O. Bussab 2017), por convenção, maior do que 20.

A distribuição t de Student pode ser usada para comparar médias entre grupos. Isso se dá pela estatística de teste t, que avalia se a média do fenótipo difere significativamente em cada grupo, levando em conta sua variabilidade. Em GWAS, é aplicável em testes de associação de SNPs com fenótipos quantitativos.

Com isso, a partir de uma amostra ou vetor X de valores observados, que seguem uma determinada distribuição, busca-se especificar o valor do parâmetro θ real da população. Esse processo é realizado por meio da Inferência Estatística (Bolfarine e Sandoval 2001), construindo um estimador $\hat{\theta}$ e calculando uma estimativa sobre o conjunto amostral (Figura 11). Essa forma de raciocínio é baseada na indução, pois a partir da observação do particular (uma amostra probabilística), infere-se sobre o geral (o parâmetro da distribuição).

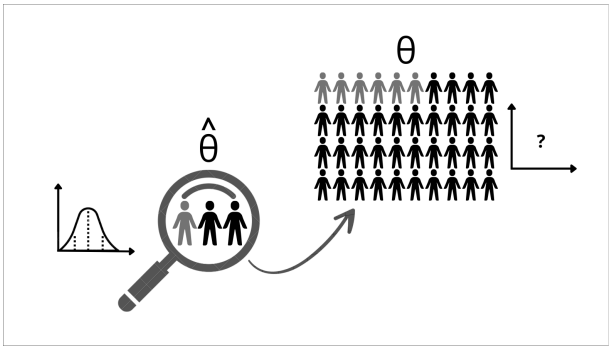

Figura 11: Ilustração de um processo de Inferência Estatística.

Por definição, o estimador é qualquer função da amostra; já estatística é o valor que o estimador assume para uma determinada amostra (Bolfarine e Sandoval 2001).

Para uma modelagem adequada, deve-se levar em conta fatores que influenciem na relação entre as variáveis, por exemplo, sob o contexto genético, parentesco entre indivíduos, pois estabelece relação de dependência entre os genótipos; fatores ambientais, pois agem como fatores causais em mudanças na expressão gênica; ancestralidade, sobre a qual é necessário agrupar as variáveis para análise, etc. Ademais, na prática, deve-se valer de uma coleta coerente dos dados para que a amostra tenha propriedades probabilísticas adequadas.

A partir disso, no campo da Inferência Estatística, surgem problemas de decisão, como justamente a escolha de um estimador adequado a partir de suas propriedades, ou a aproximação de uma estatística a um parâmetro populacional (como média amostral e média populacional, frequência relativa e probabilidade, etc.).

2.2.1 Estimação de Parâmetros

A estimação de parâmetros é uma técnica utilizada para inferir informações sobre uma população a partir de uma amostra de dados (Morettin e O. Bussab 2017). Com ela, busca-se encontrar valores, chamados de estimativas, para parâmetros desconhecidos, como a média ou a variância.

A estimação pode ser feita de forma pontual ou por intervalo. A estimativa pontual fornece um único valor como estimativa de um parâmetro, por exemplo, a média amostral

para estimar uma média populacional. Já a estimativa por intervalo fornece um intervalo no qual se admite que esteja o parâmetro populacional, com determinado nível de confiança.

Um estimador pode ser construído, por exemplo, pelo Método dos Momentos, ou pelo Método da Máxima Verossimilhança. A escolha por determinado estimador depende do contexto e das características da distribuição dos dados. Um bom estimador deve ser função de uma estatística suficiente (Bolfarine e Sandoval 2001), além disso, também pode ser classificado como não-viesado, consistente e eficiente.

Há medidas para avaliar estimadores, como o erro quadrático médio,

$$\mathrm{EQM}_\theta(\hat{\theta}) = E[\text{erro}] = E[(\hat{\theta} - \theta)^2] \tag{7}$$

(Bolfarine e Sandoval 2001) que mede a diferença quadrática entre a estimativa por $\hat{\theta}$ e o parâmetro θ; a acurácia do estimador que é avaliada com base no seu viés; a precisão, por sua vez, medida pela variabilidade das estimativas; e a suficiência. Esses conceitos são explicados em detalhes em Casella e Berger 2002 e Bolfarine e Sandoval 2001.

Um método utilizado especialmente em Regressão Linear é o Método dos Mínimos Quadrados. Esse estimador procura minimizar a soma dos erros quadráticos ao ajustar uma reta ao conjunto de pontos de uma relação observada entre X e Y. O erro quadrático para cada observação é dado por $(y_i - \hat{\theta}x_i)^2$. Assim, obtém-se o estimador

$$\hat{\theta}* = \arg\min_{\theta \in \Theta} \sum_{i=1}^{n} (y_i - \theta x_i)^2 \tag{8}$$

e esse ponto de mínimo é encontrado como raiz da derivada da função em relação a θ,

$$\sum_{i=1}^{n} (y_i - \hat{\theta}*x_i)(-2x_i) = 0 \implies \hat{\theta}* = \frac{\sum x_i y_i}{\sum x_i^2}, \tag{9}$$

que é um estimador pelo Método dos Mínimos Quadrados (Morettin e O. Bussab 2017).

2.2.2 Teste de Hipóteses

Nesse contexto, insere-se o Teste de Hipóteses, uma ferramenta que avalia hipóteses sobre parâmetros populacionais com base em amostras.

Para realizar um Teste de Hipóteses, define-se as hipóteses nula (H_0) e alternativa (H_1) sobre um parâmetro populacional. Em seguida, escolhe-se o estimador que será utilizado, sempre tendo em mente que é preciso saber a distribuição amostral desse estimador.

Além disso, deve ser determinado o erro (do tipo I) que se pode assumir, o que delimita uma região crítica. Subsequentemente, a amostra é coletada, e por fim, conforme o valor calculado para a estatística de teste ou valor P, toma-se a decisão de rejeitar H_0 ou não.

É inerente ao procedimento o fato de que sempre é cometido algum erro, classificados em dois tipos (Tabela 2). A única forma de não cometer um dos erros seria sempre rejeitar a hipótese, ou sempre aceitá-la, mas essa decisão é incoerente. Fora o caso em que a população é totalmente conhecida, o procedimento está sujeito a inexatidão.

	H_0 é verdadeira	H_0 é falsa
Rejeitar H_0	Erro tipo I	Acerto
Não rejeitar H_0	Acerto	Erro tipo II

Tabela 2: Tipos de erros em um teste de hipóteses.

Por convenção, é fixado que o erro mais grave de ser cometido é rejeitar H_0 quando H_0 é verdadeira, chamado erro tipo I, e sua probabilidade de ocorrência é definida por α. Busca-se definir o teste de modo a controlar a probabilidade de cometer o erro tipo I.

	$\theta \in \Theta_0$	$\theta \notin \Theta_0$
Rejeitar H_0	α	0
Não rejeitar H_0	0	γ

Tabela 3: Exemplo de função de perda de um teste de hipóteses onde $H_0 : \theta \in \Theta_0 \subset \Theta$.

Por definição (adaptada de Bolfarine e Sandoval 2001), um teste de hipótese estatística possui a função de decisão bijetora

$$d : \mathcal{X} \to \{a_0, a_1\}, \tag{10}$$

em que a_1 corresponde à ação de rejeitar a H_0, e a_0 não rejeitar a H_0. Essa função reparte o espaço amostral $\mathcal{X}$ em dois conjuntos disjuntos: uma região A_0 de não-rejeição de H_0,

$$A_0 = \{(x_1, \ldots, x_n) \in \mathcal{X} : \ d(x_1, \ldots, x_n) = a_0\} \tag{11}$$

e uma região de rejeição ou região crítica, que é seu complementar em relação a $\mathcal{X}$,

$$\mathcal{X} \backslash \{A_0\} = \{(x_1, \ldots, x_n) \in \mathcal{X} : \ d(x_1, \ldots, x_n) = a_1\}. \tag{12}$$

e essas regiões podem ser definidas ao fixar um valor para α.

A medida que é escolhido um valor menor para a probabilidade α, aumenta a probabilidade do erro tipo II: não rejeitar H_0 quando H_0 é falsa. Isto é, quanto menor o α, maior a probabilidade de deixar de identificar um efeito que realmente existe.

Como procedimento padrão, primeiro, é fixada uma probabilidade de cometer o erro tipo I, em geral utiliza-se $\alpha = 0,05$. Em seguida, é necessário calcular os limites que definem a região crítica, ou um valor P, à este nível de significância de 5%. E por fim, verifica-se se os valores observados (amostrais) pertencem à região crítica, ou se $P <$ 0,05, para rejeitar a hipótese nula. Em GWAS, utiliza-se o procedimento do valor P.

A seguir, um exemplo que será aplicado nas análises de associação.

Exemplo 1 (a) Seja X uma amostra aleatória de uma população, que segue uma distribuição Normal, com média θ desconhecida. Pelo teste, será determinado se a média populacional θ é igual ou diferente de um valor específico θ_0. Para isso, são formuladas as hipóteses:

$$\begin{aligned} H_0 &: \theta = \theta_0 \\ H_1 &: \theta \neq \theta_0. \end{aligned} \tag{13}$$

O teste é chamado bilateral quando a hipótese alternativa considera valores em ambas as direções, ou seja, valores maiores e menores do que a hipótese nula.

Caso a variância populacional σ^2 seja desconhecida, usa-se o desvio padrão amostral s para estimá-la. Nesse caso, a estatística de teste é dada por:

$$t = \frac{\overline{x} - \theta_0}{s/\sqrt{n}}, \tag{14}$$

em que $\overline{x}$ é a média amostral, s é o desvio padrão amostral e n é o tamanho da amostra.

A estatística de teste é um número real, calculado com os valores da amostra. A partir dela, encontra-se um valor P, que dita a decisão de aceitar ou rejeitar a hipótese. O valor P corresponde à probabilidade associada à área sob a curva de sua distribuição (Figura 12), considerando a região mais extrema em relação à estatística observada, conforme o tipo de teste.

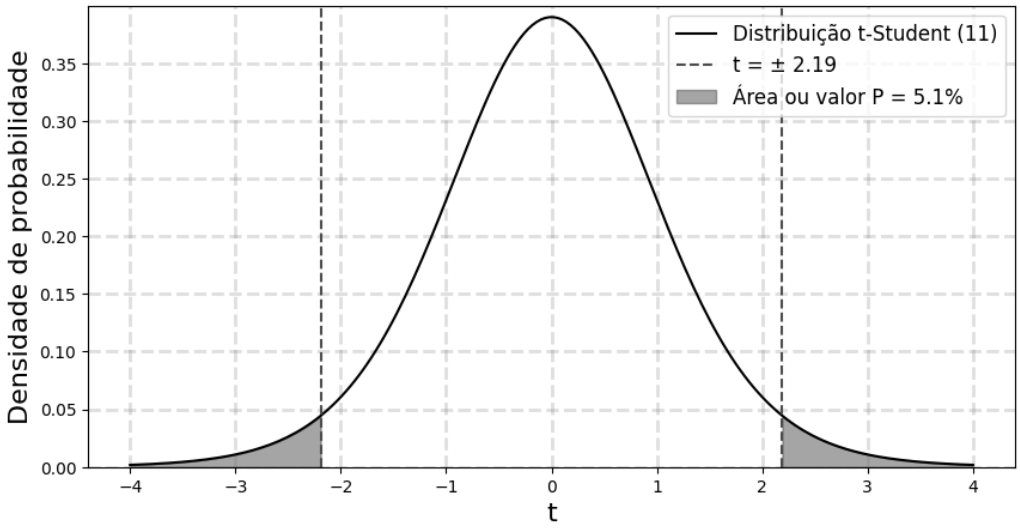

Figura 12: Exemplo de valor P bilateral. Distribuição t de Student com 11 graus de liberdade, destacando as regiões de cauda para $t = \pm 2.19$, em que o valor P é de 5,1% (2,55% em cada cauda).

Para encontrá-lo nesse caso, deve saber que a variável aleatória em (14) segue uma

distribuição t de Student com parâmetro $n-1$. Por fim, basta verificar se

$$P < \alpha \tag{15}$$

para rejeitar a hipótese nula.

Quando o tamanho amostral, n, é suficientemente grande, essa distribuição se aproxima de uma Normal.

Ainda, se a variância for conhecida e igual a σ^2, a estatística do teste seria

$$z = \frac{\overline{x} - \theta_0}{\sigma/\sqrt{n}} \tag{16}$$

(com σ ao invés da variância amostral s) que também segue uma Normal.

Exemplo 1 (b) Na outra abordagem, são definidos os limites da região crítica para os quais a área abaixo da função de probabilidade totalizam α (5%). Esses valores são facilmente encontrados para as distribuições Normal Padrão e t-Student, utilizando tabelas ou *software* estatístico como o R, disponível em `https://www.r-project.org/`.

Os limites para uma Normal Padrão são aproximadamente $z_1 = -1.96$ e $z_2 = 1.96$ (Figura 13), e bastaria verificar se $z < z_1$ ou $z > z_2$, para a estatística em (16).

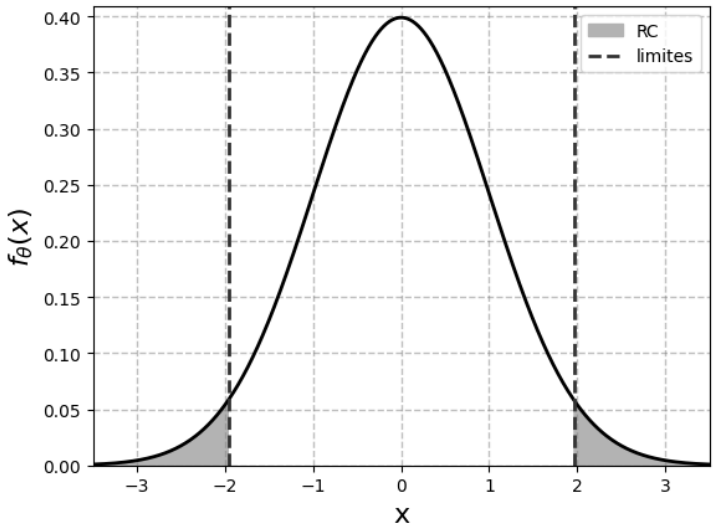

Figura 13: Região crítica (RC, em cinza) e região de não-rejeição (em branco) para uma distribuição Normal Padrão a nível de significância de 5%, pela decisão de um teste de hipóteses bilateral.

Portanto, em um teste bilateral, se o valor da estatística do teste estiver dentro da região crítica definida por um α fixado – no Exemplo 1.b, se z for menor do que -1.96 ou maior do que 1.96 –, rejeita-se a hipótese nula. Alternativamente, é possível encontrar o valor P e, se ele for menor que α, rejeita-se a hipótese nula. A vantagem da abordagem

pelo valor P, ao invés da região crítica, é que ele fornece uma medida contínua da evidência contra a hipótese nula.

De acordo com uma revisão de literatura para GWAS, sob a hipótese nula de que não há efeito da variante genética no fenótipo, costuma-se usar $\alpha = 5.10^{-8}$. Esse valor é escolhido baseado na correção de Bonferroni, que ajusta o nível de significância de 5% para testar a ordem de 10^6 variantes independentes (Marees et al. 2018). O método de Bonferroni leva em consideração que é testado um grande número de variantes genéticas simultaneamente, e ajusta o nível de significância para minimizar o risco de falsos positivos.

Note que o valor P é uma medida contínua com significado binário, por oferecer as possibilidade de rejeitar ou não rejeitar H_0. Ele não indica o quão próximo ou distante o parâmetro estimado está do valor hipotético, pois apenas serve como medida da probabilidade de cometer um erro tipo I. Assim, a partir dele decide-se rejeitar ou não a hipótese nula ao nível de significância estipulado, mas não pode ser inferido algum resultado quantitativo sobre o efeito genotípico do SNP. Para isso, é necessário estimar o tamanho do efeito, que será visto na Seção 2.2.3.

Resultado do teste	Efeito nulo	Efeito não nulo	Total
Rejeitar H_0^j	FP	VP	D
Não rejeitar H_0^j	VN	FN	$T - D$
Total	T_0	$T - T_0$	T

Tabela 4: Resultados possíveis em um teste de hipóteses em GWAS, para cada teste j, acerca da existência de efeito de uma variante no fenótipo. FP: Falsos positivos, VP: Verdadeiros positivos, D: descobertas ou variantes significativas, VN: Verdadeiros negativos, FN: Falsos negativos, T: Número total de hipóteses, T_0: Número de hipóteses nulas verdadeiras.

Foi apresentado que os resultados possíveis (Tabela 4) são rejeitar a hipótese nula ou não encontrar evidências suficientes para rejeitá-la. Entretanto, existe outra abordagem estatística, que não será tratada nesse trabalho: ao conduzir um teste *agnóstico*, pode-se rejeitar, aceitar ou se manter em dúvida sobre H_0.

2.2.3 Modelos de Regressão

Para avaliar quantitativamente a relação entre duas variáveis aleatórias distintas, em estatística, é usado o termo **regressão**, que significa justamente uma relação entre variáveis. Tal análise é possível a partir de Modelos de Regressão. A seguir, será apresentada a Regressão Linear Simples e a Logística, com base em Morettin e O. Bussab 2017 e Kutner et al. 2004.

Regressão Linear Simples Uma Regressão Linear descreve a relação linear entre duas variáveis, X e Y (Figura 14). Em particular, ela descreve que o valor esperado de Y para dado valor $X = x$ é uma função linear de x, e também é linear nos parâmetros.

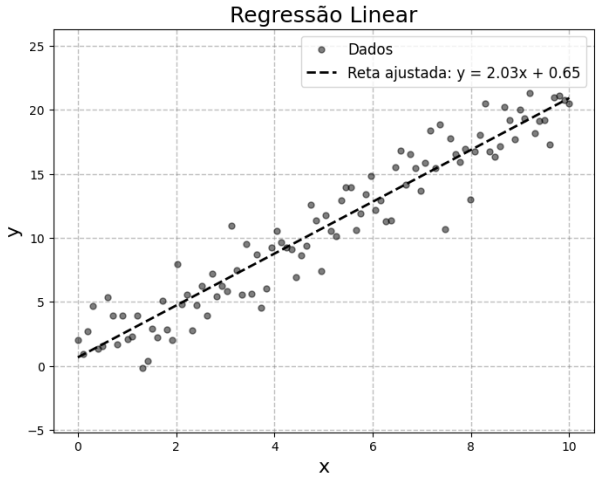

Figura 14: Exemplo de reta ajustada para 100 pontos da relação entre X e Y.

No problema de ajustar a relação entre X e Y por uma reta, implicitamente assume-se que há essa relação linear entre as variáveis, e é verificado quão bem o modelo se adequa. Uma regressão linear é dita "simples" por possuir unicamente uma variável preditora, X, e a relação é da forma

$$y_i = \alpha + \beta x_i + \epsilon_i. \tag{17}$$

para cada $y_i \in Y$ e $x_i \in X$. Y é um conjunto contínuo de valores para a variável dependente, e X o conjunto de valores da variável preditora, que pode ser contínuo ou discreto.

O intercepto α e o coeficiente angular β da reta são assumidos fixados. Como apresentados em Casella e Berger 2002, os estimadores $\hat{\alpha}$ e $\hat{\beta}$ que minimizam o erro quadrático médio são

$$\hat{\alpha} = \overline{y} - \hat{\beta}\overline{x} \tag{18}$$

e

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}, \tag{19}$$

com $\overline{x}$ e $\overline{y}$ representando as médias amostrais.

Essa ferramenta é utilizada nos testes de associação em GWAS, em que a variável X representa o genótipo, quantificando o número de cópias de uma variante genética presente em determinado alelo, e a variável assume os valores 0, 1 ou 2. A variável Y representa o nível de um fenótipo qualquer (frequência cardíaca, nível de glicose, altura, etc.) e seu espaço amostral é $\mathbb{R}$. Com isso, e assumindo que há uma relação linear entre as médias, estuda-se o quanto a presença de uma variante modifica a expressão da característica avaliada, por unidade (Figura 15).

Sob as suposições de linearidade das médias, aditividade do efeito, da normalidade, e da homocedasticidade dos grupos genotípicos, é obtida a estimativa para o efeito do

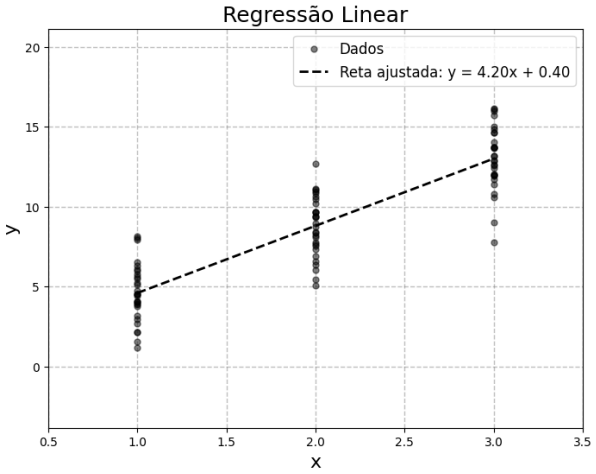

Figura 15: Exemplo de reta ajustada para 100 pontos cujo domínio dos dados em X é de 3 valores. Essa ilustração é característica de GWAS, ao analisar três diferentes genótipos (de alelos codificados como 0, 1 e 2) em relação a um fenótipo contínuo.

genótipo a partir da amostra, dada pelo coeficiente angular da reta, $\hat{\beta}$.

Em geral, na aplicação de GWAS, o tamanho do grupo para o qual $x = 2$ é consideravelmente menor que os demais, devido à baixa frequência de ocorrência das variantes genéticas de SNPs na população geral. Ainda assim, pelo Teorema do Limite Central, é possível assumir que a média dos três grupos segue uma distribuição Normal, pois o processo delineado de coleta à tratamento de dados busca garantir variáveis aleatórias independentes, e o tamanho amostral é grande.

> Pelo Teorema do Limite Central, se a amostra for suficientemente grande, a distribuição da média $\overline{X}$ terá distribuição Normal. Os parâmetros serão média da população, e a variância será dada pela variância da população dividida por n (tamanho amostral).

Regressão Logística No caso de a variável Y ser discreta, com $Y \sim \text{Bernoulli}(p)$ é utilizada regressão logística para ajustar os pontos, por

$$\log\left(\frac{p}{1-p}\right) = \alpha + \beta x_i \tag{20}$$

em que a razão

$$\frac{p}{1-p}$$

é chamada de *odds* (Casella e Berger 2002; Kutner et al. 2004) e seu logaritmo quantifica o efeito observado.

2.3 Modelo Aditivo

Comumente, os GWAS utilizam o modelo aditivo, que assume duas hipóteses principais, 1. a média dos grupos genotípicos é proporcional a quantidade cópias da variante (é dito que a média depende *aditivamente* do alelo 1, de modo que o efeito combinado é igual a sua soma), e 2. os desvios padrões dos grupos são iguais.

A suposição 2 diz respeito aos grupos serem homoscedásticos, isto é, a variabilidade em cada grupo ser similar entre si. Assim, a fim de estudar o efeito genotípico, a diferença observada entre os grupos genotípicos será atribuída mais à média dos grupos do que a diferenças na dispersão dentro dos grupos.

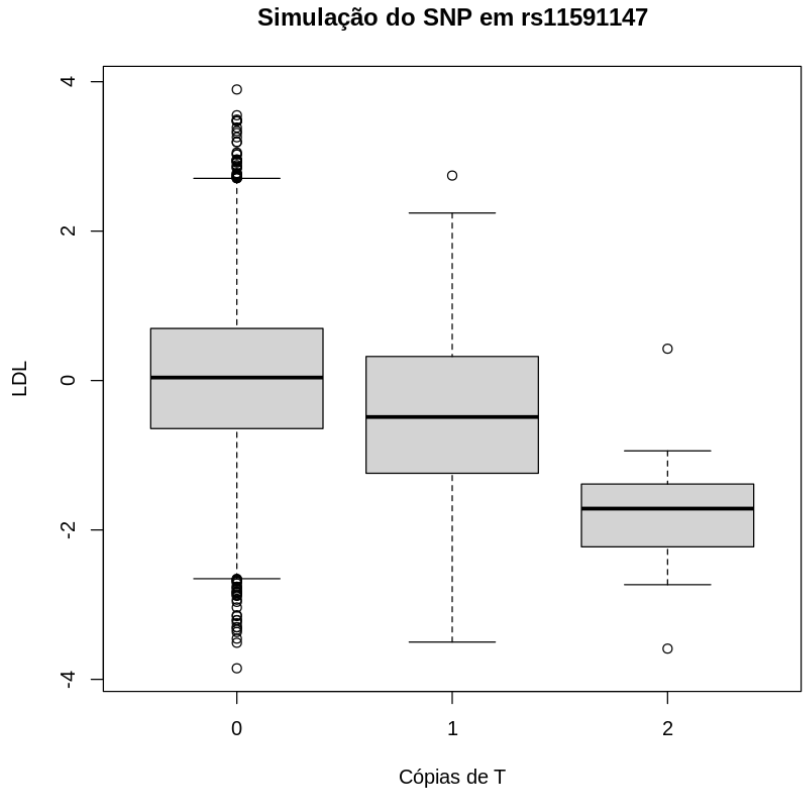

Figura 16: Box-plot do nível de colesterol LDL para o SNP na posição rs11591147. Os dados foram simulados: assume-se que as distribuições individuais com 0, 1 ou 2 cópias do alelo T no SNP seguem Normais com desvio padrão 1 e médias 0,02, -0,40 e -2,00 respectivamente, e que a frequência do alelo T é 4% (Pirinen 2023).

A escolha por esse modelo em específico se dá pela sua simplicidade e capacidade de capturar com precisão "suficiente" os efeitos genéticos. Além disso, o poder estatístico para detectar não-aditividade é baixo na prática (Marees et al. 2018). Contudo, como o modelo aditivo depende da suposição de que os efeitos dos alelos sejam independentes e lineares, é limitada a capacidade de capturar relações genéticas mais complexas.

Note no exemplo da Figura 16 que o nível de colesterol parece decrescer a cada cópia adicional do alelo T. É assumida, nessa simulação, uma distribuição Normal para cada

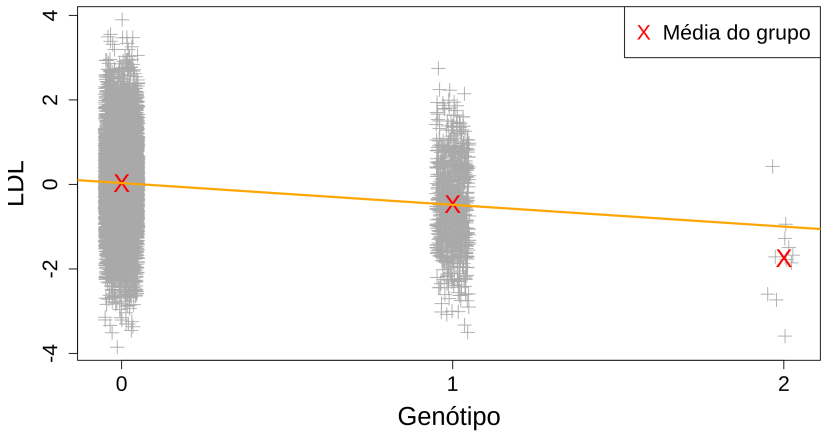

Figura 17: Regressão linear nos mesmos dados da simulação da Figura 16 (Pirinen 2023).

grupo genotípico, com desvio padrão constante e médias decrescentes à medida que a contagem do alelo T aumenta. Essa suposição permite que seja utilizada uma Regressão Linear Simples para ajustar a relação entre o genótipo e o fenótipo, e estimar o efeito $\hat{\beta}$.

Em geral, o terceiro grupo genotípico é menor que os dois primeiros (Figura 17), devido à baixa frequência dos SNPs na população. No entanto, isso não afeta consideravelmente a adequação ao modelo, pois são estudadas variantes comuns, e não raras.

O modelo aditivo não testa explicitamente efeitos dominantes ou recessivos, mas consegue captar esses padrões como efeitos lineares aproximados por

$$y = \mu + x\beta + \epsilon, \tag{21}$$

em que y é o fenótipo, x é o genótipo (0, 1 e 2), μ é a média do genótipo 0, e β é o efeito de cada cópia do alelo 1 na média do fenótipo. Esse modelo é ajustado aos dados, a partir do qual μ e β são estimados.

2.4 Modelo Biométrico

A teoria do Modelo Biométrico, que considera tanto os efeitos genéticos quanto os ambientais sobre características fenotípicas, é deduzida a partir da Herança Mendeliana. Nessa teoria, será apresentada a distribuição dos genótipos e dos seus efeitos na população, além dos elementos de variâncias genotípica, aditiva, e de dominância, que são considerados na estratégia para a modelagem matemática.

Humanos são organismos diploides, pois possuem dois conjuntos cromossômicos, dispostos em pares. Portanto, assume-se um sistema bialélico, no qual para cada variante genética, dois alelos diferentes existem numa população. Consideremos o alelo menos comum 0, e o alelo mais comum 1. Seja f a frequência do alelo 1 e, consequentemente, $(1 - f)$ a frequência do alelo 0.

Alelos	0	1
Freq.	f	$(1-f)$

Tabela 5: Frequência dos alelos de uma população

Com dois alelos, há 3 possíveis genótipos ($\mathcal{G}' = \{(0,0),(1,0),(1,1)\}$). Para facilitar análises quantitativas, codificamos os genótipos aditivamente: seja cada genótipo dado pelo número de cópias que contém do alelo 1 ($\mathcal{G} = \{0,1,2\}$).

Por hipótese de cruzamento aleatório, igual viabilidade dos alelos, e em um contexto sem seleção, migração e mutações, espera-se que a frequência desses genótipos permaneça inalterada, sendo respectivamente f^2, $2f(1-f)$, e $(1-f)^2$, e que as proporções genotípicas atinjam um equilíbrio estável. Sob essas condições, dizemos que a população está em equilíbrio de Hardy-Weinberg (HWE) (Mayo 2008). Nesse caso, note que o genótipo segue distribuição binomial, $\text{Bin}(2, f)$.

Pares de alelos	$(0,0)$	$(0,1)$ ou $(1,0)$	$(1,1)$
Genótipos	0	1	2
Frequências	f^2	$2f(1-f)$	$(1-f)^2$

Tabela 6: Frequências dos genótipos de uma população sob o HWE.

Vale salientar que desvios do HWE podem ser observados em populações recém miscigenadas, se há letalidade do genótipo, e/ou como resultado de dados mal coletados.

2.4.1 Teste para Equilíbrio de Hardy-Weinberg

Ao analisar uma amostra, é importante verificar se ela segue o equilíbrio de Hardy-Weinberg, para certificar-se que ela condiz com a previsão teórica, em certa proximidade aceitável, se for o caso esperado. Caso não siga, é provável que fatores como seleção natural, mutação, migração, deriva genética ou cruzamentos não aleatórios estejam atuando sobre a população. Essa verificação pode ser executada pela ferramenta de um teste estatístico.

Portanto, para testar se uma determinada amostra está em HWE, isto é, se segue uma distribuição Binomial indexada pelo parâmetro $\theta = (2, f)$, é executado teste de qui-quadrado de Pearson, com 1 grau de liberdade.

Considere uma amostra X de indivíduos, de tamanho n. Esse conjunto é repartido em X_0, X_1 e X_2 de tamanhos n_0, n_1 e n_2 (tais que $n = n_0 + n_1 + n_2$), de tal modo que cada subconjunto contenha exatamente as subpopulações com cada genótipo 0, 1 e 2.

A frequência alélica esperada do alelo 1 na população é dada por

$$f = \frac{n_1 + 2n_2}{2n} \tag{22}$$

Como visto, sob a hipótese de seguirem uma distribuição $\text{Bin}(2, f)$, o número esperado de pessoas com cada genótipo (em cada subpopulação) é

$$\begin{cases} x_0 = nf^2 \\ x_1 = 2nf(1-f) \\ x_2 = n(1-f)^2 \end{cases} \tag{23}$$

O teste estatístico Qui-Quadrado, que mede o desvio entre o observado na amostra e o esperado, é dado por

$$\chi^2 = \sum_{i=0}^{2} \frac{(n_i - x_i)^2}{x_i} \tag{24}$$

para $n_i \in N_i \subset N$ e $i \in \{0, 1, 2\}$.

Nas condições do HWE, temos que χ^2 segue aproximadamente uma distribuição qui-quadrado com 1 grau de liberdade.

De posse do valor da estatística, é conhecida a probabilidade de se obter um valor tão extremo quanto observado, como definido sob uma ótica frequentista. Se χ^2 é baixo (por exemplo, menor que 3,83 para um nível de 5%), conclui-se que há uma boa concordância entre as distribuições. Caso contrário, a população não está em equilíbrio.

2.4.2 Efeito Genotípico

Considere que uma variante genética tem um efeito, quantificado por um número real, em um determinado fenótipo. Por definição, esse é o efeito genotípico. Como visto, ele é estimado pelo β na regressão linear ou logística.

No Modelo Biométrico, o efeito genotípico do genótipo 2 é denotado por a; o efeito genotípico de 0 é $-a$; já o efeito de 1 é d, que quantifica uma dominância – assim, se $d = 0$, não há dominância, como no Modelo Aditivo.

Genótipos	0	1	2
Frequências	f^2	$2f(1-f)$	$(1-f)^2$
Efeitos	$-a$	d	a

Tabela 7: Efeitos genotípicos

Assuma apenas efeitos independentes e aditivos (aqueles cujo efeito combinado é igual a sua soma). A média populacional do efeito é expressa pela soma

$$\begin{aligned} \mu &= a(1-f)^2 + 2f(1-f)d - af^2 \\ &= a(1-2f) + 2d(1-f)f, \end{aligned} \tag{25}$$

que é uma função que depende dos efeitos dos genótipos, ponderados pelas frequências alélicas.

Portanto, cada variante genética que tem um efeito não nulo contribui para a média populacional daquele fenótipo.

Existe uma relação entre as variâncias

$$V_F = V_G + V_E \tag{26}$$

em que V_F é variância fenotípica, V_G é variância genotípica e V_E é variância de influências externas. Em especial, calcula-se V_G a partir da média populacional μ, do efeito daquele genótipo x_i, e da frequência genotípica f_i para um genótipo i:

$$\begin{aligned} V_G = \sigma^2 &= \sum_{i=0}^{2} f_i(x_i - \mu)^2 \\ &= f^2[-a - (a(1-2f) + 2(1-f)fd)]^2 \\ &\quad + 2f(1-f)[d - (a(1-2f) + 2(1-f)fd)]^2 \\ &\quad + (1-f)^2[a - (a(1-2f) + 2(1-f)fd)]^2 \\ &= 2f(1-f)[a + d(1-2f)]^2 \ + \ [2(1-f)fd]^2 \\ &:= V_A \ + \ V_D \end{aligned} \tag{27}$$

em que V_A é o componente de variância aditiva, e V_D é o componente de variância de dominância (Uffelmann et al. 2021; Falconer e Mackay 1995).

Variância genotípica: Variância observada nas diferenças fenotípicas entre os indivíduos, mas que é causada somente pelas diferenças genéticas entre os indivíduos de uma população.

Variância aditiva: É a parte previsível e herdável da variância fenotípica, explicada pelos efeitos médios dos alelos, que são transmitidos entre gerações.

Variância de dominância: Não é tão simples de quantificar, pois depende da combinação específica de um grupo de alelos, que podem interagir entre si. Por exemplo, os efeitos de um alelo podem se opor ao efeito de outro.

3 Métodos

3.1 Estratégias para a Modelagem Matemática de GWAS

Os GWAS executam múltiplos testes de hipóteses, sendo um para cada par: variante genética e fenótipo (Figura 18).

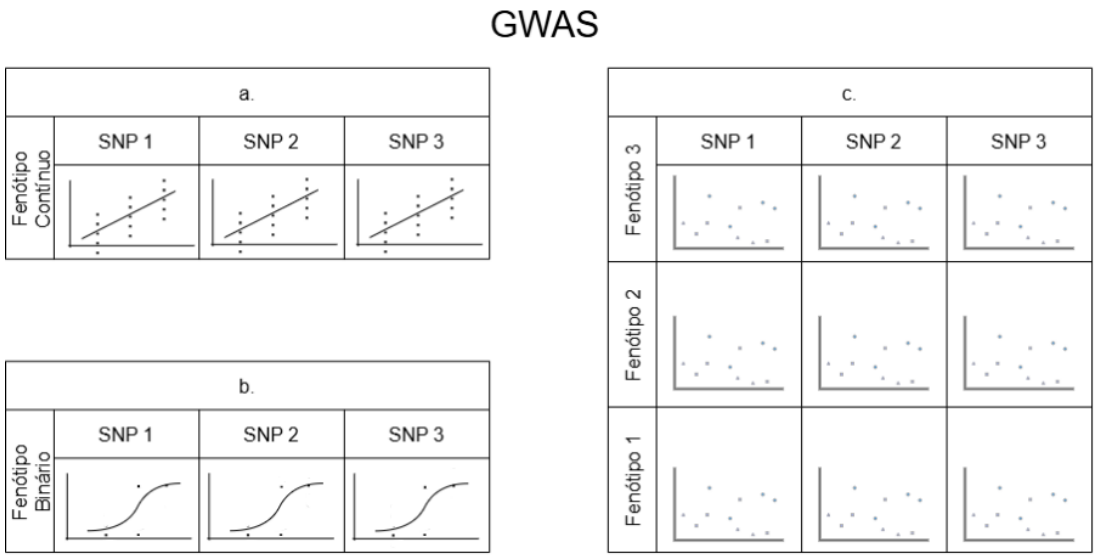

Figura 18: Painel que representa diferentes conjuntos de regressões, dentre haplótipos e fenótipos distintos, que são avaliados pelo GWAS. **a.** Para o estudo de um fenótipo contínuo (e.g. nível de colesterol), são avaliadas regressões lineares entre o fenótipo e cada SNP identificado. **b.** Para o estudo de um fenótipo binário (e.g. ausência/presença de uma doença), são avaliadas regressões logísticas entre o fenótipo e cada SNP identificado. **c.** GWAS multivariado, que relaciona componentes principais a cada SNP identificado, e o tipo de avaliação executada depende das propriedades de cada componente.

A análise de correlação entre características observáveis e SNPs identificados é feita por meio de hipóteses, que testam a existência de efeito do genótipo, a partir da relação entre milhares de dados de variantes genéticas identificadas em indivíduos da amostra populacional, e seus fenótipos avaliados. Pode-se correlacionar dentre grupos de casos *versus* controles, ou em níveis contínuos. Além disso, é assumido que os indivíduos são aleatoriamente selecionados da população, e o tratamento de dados busca garantir independência.

Para testar as associações, em geral são utilizados modelos de Regressão Linear se o fenótipo é modelado por uma variável contínua, como altura, pressão sanguínea e massa corporal, ou Regressão Logística se a variável é binária, por exemplo ao avaliar presença ou ausência de determinado fator. Para isso, supõe-se que as médias dos efeitos para cada grupo genotípico ($\{0, 1, 2\}$, codificados de acordo com a quantidade de cópias da variante nos alelos) estão de acordo com a propriedade aditiva, e são homoscedásticos. Desse modo, obtém-se uma estimativa para o efeito, que é submetida ao teste de hipóteses.

O poder estatístico para detectar não-aditividade é baixo na prática, por isso testes não-aditivos não são muito aplicados (Marees et al. 2018). Assim, a herança de SNPs é medida pela variância explicada por efeitos aditivos, vista nas seções anteriores.

Sob o modelo aditivo, duas cópias do alelo 1 tem o dobro do efeito que uma cópia, e nenhuma cópia tem zero efeito. Neste modelo,

$$y = \mu + \beta x + \epsilon, \tag{28}$$

y é a variável de resposta para o fenótipo, $x \in \{0, 1, 2\}$ é a variável preditora sobre um genótipo, μ é a média do genótipo 0, e β é o efeito de cada cópia do alelo 1 na média do fenótipo. Estes dois últimos parâmetros, μ e β serão estimados. O erro ϵ segue uma $N(0, \sigma^2)$, com variância σ^2 desconhecida, a ser estimada da amostra também.

A hipótese nula é

$$H_0 : \beta = 0, \tag{29}$$

ou seja, declara que o tamanho do efeito genotípico é nulo. No contexto apresentado, essa hipótese é plausível para a maior parte das variantes. E a hipótese alternativa é

$$H_1 : \beta \neq 0. \tag{30}$$

O resultado obtido ao executar o teste de associação é composto por

1. estimativas para os parâmetros e seus erros associados;

2. valores P, a partir da hipótese nula.

Os principais resultados da regressão são as estimativas dos parâmetros e seus erros. Além deles, o resultado do valor P é central em GWAS, pois o teste da hipótese nula de zero efeito do alelo no fenótipo – o que é uma hipótese que vale para a maioria das variantes no genoma – só pode ser decidido pelo valor P.

A partir de uma regressão dentre o genótipo de algum SNP em determinada posição genômica e um dos fenótipos avaliados, é obtido um valor P que indica, probabilisticamente, se há ou não correlação. Desse modo, o número de associações avaliadas e de P valores obtidos é o mesmo que o de fenótipos avaliados vezes o de variantes genéticas (Figura 18).

Comumente, é gerado um Manhattan *Plot* após associações de um GWAS (Figura 19). Cada ponto nele é advindo de um valor P obtido pelas regressões. Já o parâmetro de limiar, para o qual se rejeita ou aceita a hipótese de correlação, é indicado pela linha horizontal. Assim, os SNPs cujos valores P superam o limiar são categorizados como SNPs significantes para o fenótipo de interesse.

O parâmetro de limiar usualmente em 5.10^{-8} é o resultado da correção de Bonferroni aplicada para conseguir FWER (*family-wise error rate*, ou taxa de erro familiar, que quan-

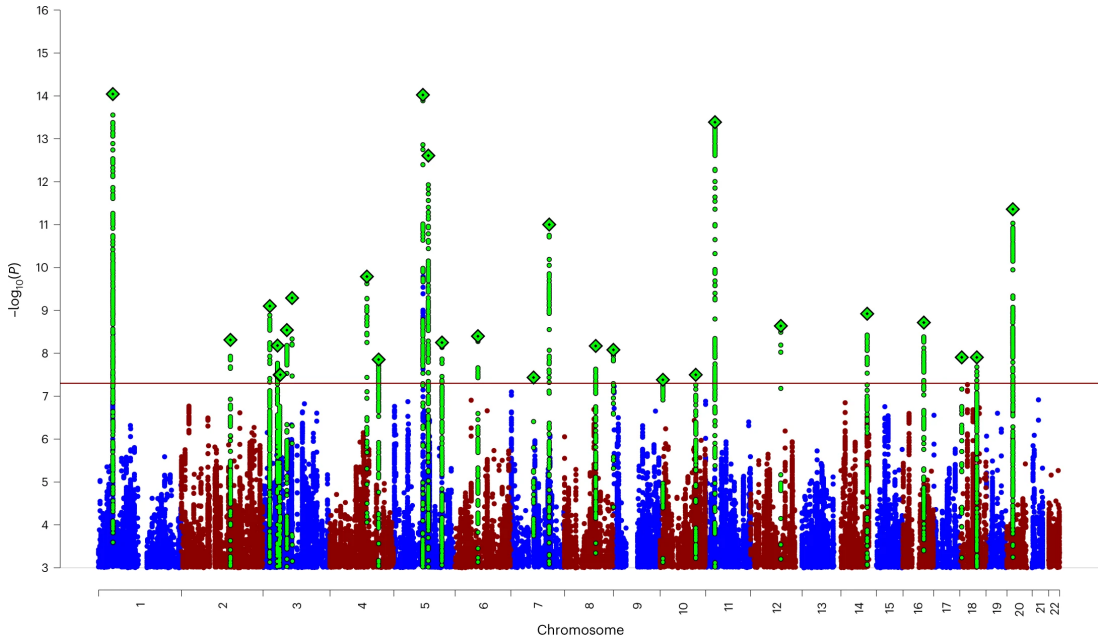

Figura 19: Manhattan *Plot*. No eixo x, a posição cromossômica de um SNP identificado. No eixo y, o valor P em escala logarítmica. Fonte: Demontis et al. 2023.

tifica a probabilidade de erros do tipo I, ao realizar testes de múltiplas hipóteses) fixada em $0,05$, ao testar 10^6 variantes independentes. Portanto, a depender da quantidade de variantes testadas, é recomendado recalcular esse parâmetro do limiar, para condizer ao mesmo erro fixado. Por isso, na Figura 19, o limiar representado pela linha em vermelho é constante em

$$-\log_{10}(5 \times 10^{-8}) \approx 7{,}3.$$

É crucial frisar que o valor P não representa a intensidade do efeito fenotípico, apenas indica se aquela posição cromossômica é, ou não, significante no contexto estudado. A intensidade do efeito é medida pela estimativa para o coeficiente β, ao ajustar os dados por regressão.

3.1.1 Modelo de Regressão Linear

O modelo a seguir será uma generalização daquele visto anteriormente (28), pois compreende o fato de que as variáveis são organizadas em matrizes e vetores, que representam os conjuntos de valores escalares e de dados individuais. Além disso, leva em conta um ajuste para variáveis de confusão, ou confundidores, na forma de co-variáveis.

Um modelo de regressão linear misto para GWAS é dado por

$$Y = W\alpha + X_s\beta_s + g + e \tag{31}$$

$$g \sim N(0, \sigma_A^2\psi) \tag{32}$$

$$e \sim N(0, \sigma_e^2 I) \tag{33}$$

em que

- Y é o vetor de valores fenotípicos reais (variável de resposta),
- W é a matriz de possíveis co-variáveis, como idade ou gênero,
- α é o vetor correspondente de efeitos,
- X_s é o vetor dos valores genotípicos para todos os indivíduos no SNP s (variável preditora),
- β_s é o valor fixado de efeito do SNP s,
- g é um efeito aleatório que captura o efeito poligênico de outros SNPs,
- e é o efeito aleatório de erros residuais,
- σ_A^2 é a variação genética aditiva [1] do fenótipo,
- ψ é a matriz padrão de relação genética,
- σ_e^2 é a variância residual, e
- I é uma matriz identidade

(Uffelmann et al. 2021).

A saída ao executar um GWAS contém: um vetor dos valores das estimativas para os tamanhos de efeitos, o desvio padrão de cada estimativa, e o vetor de seus valores P, para cada SNP.

Se a estimativa para $\hat{\beta}_s$ for positiva, isso significa que a presença dessa variante genética s está associada a um aumento no fenótipo (por exemplo, maior altura), enquanto um valor negativo indicaria uma associação com uma diminuição no fenótipo (por exemplo, menor altura). No entanto, o valor de $\hat{\beta}$ por si só não determina se o efeito genotípico observado é clinicamente relevante, quem faz isso é o valor P, obtido para a estimativa sob o teste de hipóteses.

3.1.2 Teste com duas amostras

Para analisar um fenótipo qualitativo, em um estudo de caso-controle, pode-se construir uma tabela de contingência para testar por associações, por meio de um teste estatístico discreto, como o de Fisher ou χ^2 (Gumpinger et al. 2018).

[1] Quando um traço é influenciado por múltiplos genes, a variação genética aditiva é a soma dos efeitos individuais desses genes sobre o fenótipo do organismo.

Nesse caso, não pode ser aplicada Regressão Linear, pois não é possível assumir que os termos de erro dentre fenótipos binários seguem uma Normal. Portanto, alternativamente, são comparadas *odds*, que podem ser estimadas. Elas são definidas por

$$\frac{p}{1-p} \tag{34}$$

em que p é a probabilidade do evento (no caso, a presença do fator estudado) acontecer, para $Y \sim \text{Bernoulli}(p)$. Ou seja, as *odds* dizem quantas vezes é mais provável de o evento acontecer do que de não acontecer.

Assim, a estimativa é calculada por contagem. Contudo, diferentemente da associação realizada por regressão, ela não leva em conta co-variáveis de confusão.

3.1.3 Modelo de Regressão Logística

Um modelo de regressão logística misto para GWAS é dado por

$$\log\left(\frac{p}{1-p}\right) = W\alpha + X\beta + g \tag{35}$$

em que

- p é a probabilidade de $Y = 1$, para $Y \sim \text{Bernoulli}(p)$,
- W é uma matriz de covariáveis,
- α é o vetor correspondente de efeitos,
- X é um vetor de genótipos, codificados como 0, 1 ou 2,
- β é o vetor de efeitos em cada SNP, e
- g é um vetor aleatório, que segue uma distribuição Normal multivariada com média 0 e variância desconhecida

(Milet et al. 2020).

3.2 Tratamento de dados

Na primeira etapa da metodologia de GWAS, os extratos de DNA são coletados para que o genoma completo dos voluntários seja sequenciado ou genotipado por biochips. Nesse processo, também é essencial construir uma amostra individual bem caracterizada, para possibilitar uma análise final sem tanta heterogeneidade.

Depois, os dados são tratados por meio de ferramentas computacionais, em especial o RICOPILI (*Rapid Imputation and Computational PipeLine for GWAS*), que facilitam os passos descritos a seguir.

Primeiro, é necessário um Controle de Qualidade, para garantir indivíduos independentes e evitar introduzir viés na análise, no qual são removidos polimorfismos de nucleotídeo único (SNPs) mal genotipados, amostras consideradas ruins por apresentarem DNA mal isolado, indivíduos aparentados, e é feita conferência de sexo biológico entre o informado e o coletado.

Ainda no sentido de tratamento da amostra, por meio da análise de Componente Principal, uma técnica de redução dimensional linear, pode ser realizada estratificação populacional se necessário. É um processo em que há divisão dos membros da população em subgrupos homogêneos por ancestralidade. Os estratos definem uma partição da população, pois cada indivíduo é atribuído a apenas um estrato. O objetivo da estratificação é assegurar a representatividade de cada estrato na amostra e permitir inferências sobre subgrupos específicos da população. A análise destes subgrupos deve ser considerada por conta da diferença nas frequências alélicas entre diferentes grupos étnicos.

Para controle de estratificação populacional, é utilizado o método de Escalonamento Multidimensional de Análise de Componente Principal, detalhado em Marees et al. 2018 e Pirinen 2023, que identifica grupos mais similares entre si que o esperado, geralmente refletindo a ancestralidade. Isso ocorre porque algumas variantes podem ser mais comuns em determinados grupos devido a sua ancestralidade, além disso, eventuais características fenotípicas distintas entre grupos podem não ser relacionadas às variantes genéticas. De acordo com esse método, a partir da média da proporção de alelos entre todo par de indivíduos da amostra, são calculadas coordenadas quantitativas da variação genética para cada indivíduo, usadas para agrupá-los.

Em seguida, é realizada a etapa de Imputação. A partir de bancos de dados em consórcios de genomas, é feita uma previsão de variantes genéticas associadas àquelas identificadas, a partir do desequilíbrio de ligação entre elas. O desequilíbrio de ligação (LD) é medido pela associação não aleatória de alelos em diferentes loci. Ele descreve uma situação em que alguns conjuntos de alelos ocorrem mais ou menos frequentemente numa população do que era esperado pela sua combinação aleatória, sendo influenciado principalmente pela distância física entre os genes, que permanecem juntos quando acontece a meiose.

No contexto da informática, com o objetivo de otimizar capacidade computacional, são usados *clusters* de computadores ou ambientes na nuvem que podem distribuir o trabalho de processamento de todos esses dados para vários computadores. Já pipelines analíticas podem ser rodadas em paralelo (Uffelmann et al. 2021).

3.3 Base de dados

3.3.1 Formatos

A maioria dos GWAS é conduzida utilizando recursos pré existentes, como coortes de informação em larga escala, por exemplo o UK Biobank (disponível em `https://www.ukbiobank.ac.uk/`) e o Estonian Biobank (`https://genomics.ut.ee/en/content/estonian-biobank`). Nesses biobancos, os dados de indivíduos são fortemente fenotipados por questionários (Uffelmann et al. 2021).

Por se tratar de dados de genoma humano, uma informação altamente pessoal e referente a saúde, seu acesso requer um contrato legal por escrito entre a organização do pesquisador e cada voluntário.

Os dados de GWAS são armazenados e tratados por um formato comum fornecido por um programa de linha comando, o PLINK. Para exportar dados para o formato PLINK basta acessar `https://www.cog-genomics.org > plink`.

Em PLINK textual, são armazenados

1. Em um arquivo *.ped, a informação dos indivíduos e seus genótipos
2. E em um arquivo *.map, a informação dos marcadores genéticos

Já no PLINK binário,

1. Em um arquivo binário *.bed, os identificadores individuais e resultados de genotipagem de todos os pacientes e controles saudáveis
2. Em texto, *.fam, a informação dos indivíduos. Esse arquivo contém

   FID: ID da família,

   IID: ID individual,

   PID: ID paterno,

   MID: ID materno,

   Sexo: 1 (masculino), 2 (feminino) ou 0 (desconhecido)

3. E em *.bim, a posição física dos SNPs, ou marcadores genéticos. Por sua vez, organiza

   CHR: cromossomo,

   POS: posição,

   CM: centimorgan,

   A1: alelo 0,

   A2: alelo 1

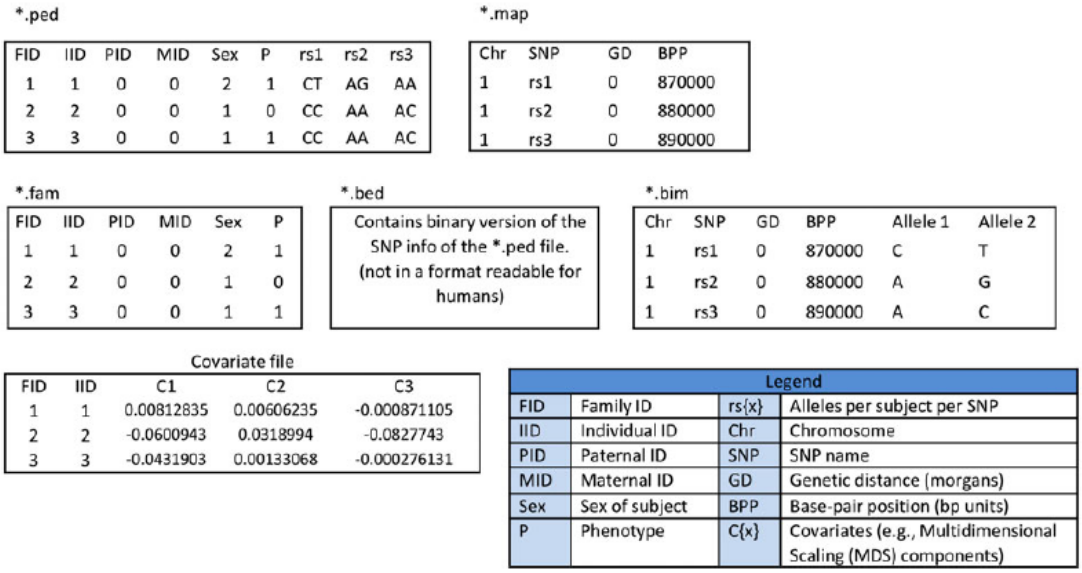

Figura 20: Exemplos de arquivos comumente utilizados para o PLINK. Fonte: Marees et al. 2018.

Também há o formato VCF (*Variant Call Format*) para dados de GWAS. Ele é constituído por combinações de genótipos de cada amostra em que

- 0/0 é homozigoto para o alelo de referência,
- 0/1 é heterozigoto,
- 1/1 é homozigoto para o alelo alternativo, e
- ./. genótipo indeterminado (*missing*)

(Marees et al. 2018).

3.4 Ferramentas

3.4.1 R

Para análise estatística e modelagem de dados, pode-se fazer uso da ferramenta R. Ela é capaz de lidar com grandes volumes de dados, padrão em pesquisas genômicas.

São oferecidos pacotes que permitem a criação de gráficos sofisticados e personalizados, por exemplo o *ggplot2*. Essa funcionalidade é útil, pois visualizações intuitivas colaboram para interpretar dados complexos em bioinformática.

Além disso, o R possibilita reprodutibilidade de códigos, possui uma grande comunidade ativa e recursos abundantes, como pacotes específicos para bioinformática e que permitem integração com outras ferramentas.

3.5 Análises Posteriores com Estatísticas Sumárias

A partir das estatísticas sumárias resultantes das associações vistas anteriormente, são calculadas pontuações que veremos a seguir, para análises posteriores.

Dentre as análises pós-GWAS, destacam-se a de Escore Poligênico, que consiste numa soma ponderada de alelos de risco que um indivíduo possui, em que os pesos são seu respectivo tamanho de de efeito, geralmente estimado na forma de um coeficiente angular β de uma regressão linear ou logística. Também há a Correlação Genética Global, que verifica a proporção de variância que duas variáveis de interesse compartilham pela genética, em um intervalo de 0 a 1, sendo 0 nenhuma covariância e 1 uma singularidade; ela requer somente o resumo estatístico de GWAS e não é viesada por sobreposição amostral. E por fim, a Correlação Genética Local, que analisa a proporção de variância que duas variáveis alvo compartilham por causa da genética em um segmento particular do genoma, particionado utilizando LD (Marees et al. 2018; Uffelmann et al. 2021).

> As estatísticas sumárias contém: lista de todos os SNPs testados, valores P, estimativas dos tamanhos dos efeitos e seus erros; informação de IDs dos SNPs, localização dos SNPs, *build*[2], frequência do alelo menor (MAF) e tamanho da amostra.

3.5.1 Escore Poligênico

Escores de risco poligênico dizem respeito ao risco genético individual, relativo a outros indivíduos – e não absoluto –, para uma doença (Choi, Mak e O'Reilly 2020).

São necessárias apenas duas bases de dados, que são amostras independentes entre si:

1. Estatísticas sumárias de GWAS;
2. Dados alvo

As estatística sumárias são normalmente disponíveis com total acesso, sem restrição de compartilhamento; já os dados alvos são de acesso dos pesquisadores perfomando a análise de PRS, e não necessariamente precisam ser disponibilizados publicamente. As análises de predição de risco (PRS) são feitas nos dados alvo, que consistem normalmente de genótipos e fenótipos dos indivíduos.

O escore de risco poligênico de um indivíduo j é definido por

$$\sum_{i=1}^{m} \hat{\beta}_i x_{ij}, \tag{36}$$

em que

[2] *Build* é a versão de referência do genoma humano, construída pelo *Genome Reference Consortium*.

- m é o número de SNPs,
- $\hat{\beta}_i$ é a estimativa de peso por alelo para o SNP, resultante do GWAS, e
- $x_{ij} \in \{0, 1, 2\}$ é a quantidade de alelos do SNP i no indivíduo j, associados ao traço.

Uma das principais limitações enfrentadas atualmente é de que a precisão da predição de risco decai à medida que aumenta a distância quanto à ancestralidade entre a coorte de descoberta e a alvo. Há uma perspectiva de melhora no aumento da diversidade em coortes de descoberta de GWAS, mas ainda há populações pouco representadas – em geral, populações não europeias.

3.5.2 Correlação Global

Até a criação desse método, não se fazia muito clara a distinção entre uma inflação estatística por viés, ou por um sinal verdadeiro de poligenia, em que um grupo de genes acumula seus efeitos para influenciar em fenótipos. Foi observado que variantes em desequilíbrio de ligação com uma variante causal mostram estatísticas de teste elevadas em análises de associação, proporcionais ao desequilíbrio de ligação com a variante causal. Portanto, para solucionar esse problema, a Regressão de Escore LD quantifica a contribuição de cada SNP para a herança de um traço, por meio da relação entre as estatísticas de teste e o desequilíbrio de ligação das variantes (B. K. Bulik-Sullivan et al. 2015).

Os fenótipos são modelados pela equação

$$\phi = X\beta + \epsilon \tag{37}$$

em que

- ϕ é um vetor $N \times 1$ de fenótipos quantitativos,
- X é uma matriz $N \times M$ de genótipos normalizados para média zero e variância 1,
- β é um vetor $M \times 1$ de tamanhos de efeitos genotípicos normalizados, e
- ϵ é um vetor $N \times 1$ de efeitos ambientais

As três últimas variáveis são aleatórias e mutualmente independentes.

Para incorporar desequilíbrio de ligação no modelo, é definido

$$r_{jk} := \mathbb{E}[X_{ij}X_{ik}] \tag{38}$$

que denota a correlação amostral entre genótipos codificados aditivamente, nas variantes j e k, para cada indivíduo independente i.

E o Escore LD da variante j é dado por

$$l_j = \sum_{k=1}^{M} r_{jk}^2 \tag{39}$$

que mede a quantidade de variação genética dada por j.

Assim, como detalhado no material suplementar de B. K. Bulik-Sullivan et al. 2015, o valor esperado da estatística qui-quadrado é decomposto em

$$E[\chi^2] \approx \frac{Nh_g^2}{M} l_j + 1 \tag{40}$$

que se torna uma igualdade se somada a inflação por viés. O primeiro termo representa a contribuição da variância explicada pelos SNPs, com a herdabilidade (h_g^2) escalada pelo Escore LD e pelo tamanho amostral de genótipos e fenótipos.

Portanto, ao fazer a regressão de χ^2 contra o Escore LD, o intercepto menos 1 estima o viés médio de confusão que infla a estatística de teste (B. K. Bulik-Sullivan et al. 2015).

3.5.3 Correlação Local

Conhecido como LAVA (acrônimo para *Local Analysis of [co]Variant Association*, isto é, Análise Local de Associação Covariante), é um *framework* que visa análise local de correlação genética, denotada na literatura por r_g.

O objetivo da análise de uma correlação genética, em geral, é a identificação de pares de fenótipos que podem ser resultantes de pleiotropia ao longo do genoma, ou seja, que compartilhem de uma mesma base genética.

Tradicionalmente essa correlação é estudada numa escala global. A definição de uma região local permite certa customização, detalhada em Werme et al. 2022.

Além de testar o padrão de correlações genéticas entre dois fenótipos, esse método também avalia heranças locais e analisa as relações genéticas condicionais entre múltiplos fenótipos. Isso é feito por meio de correlação parcial e regressão múltipla. Permite, assim, elucidar relações complexas de genética multivariada.

Seu modelo utiliza as seguintes variáveis

- Y_f é o vetor de fenótipos contínuos centrado na média, para cada fenótipo f,
- X é uma matriz normalizada de genótipos contendo K SNPs,
- α_f é o vetor de efeitos conjuntos dos SNPs, levando em conta LD, e
- $\epsilon_f \sim N(0, \eta_p^2)$ é um vetor residual

Similarmente ao que já vimos em GWAS, a relação entre Y_f e X é dada pelo modelo de regressão linear múltipla

$$Y_f = X\alpha_f + \epsilon_f \tag{41}$$

A estratégia para encontrar r_g é reconstruir um modelo de regressão múltipla. Primeiro, α_f, de (41), é estimado por

$$\hat{\alpha}_f = (X^T X)^{-1} X^T Y_f. \tag{42}$$

Além disso, usando um *dataset* de referência de genotipagem de LD, é computada uma matriz S dada por $cor(X)$, de posto K,

$$S = X^T X. \tag{43}$$

Seja $\hat{\beta}_f$ a estimativa do vetor dos efeitos de SNPs, que não leva em conta LD, obtida das estatísticas sumárias de GWAS para Y_f:

$$\hat{\beta}_f = X^T Y_f. \tag{44}$$

A partir de (43) e (44), temos (42) reescrita como

$$\hat{\alpha}_f = S^{-1}\hat{\beta}_f \tag{45}$$

e então pode-se estimar os efeitos de SNPs conjuntos α_f, todo esse procedimento sem nenhum dado a nível individual, apenas utilizando estatísticas sumárias e dados de LD.

Calculado $\hat{\alpha}_f$, é então estimada a variância fenotípica residual local (η_f), e por fim a variância fenotípica explicada pelos SNPs no locus (h^2), detalhado em Werme et al. 2022.

Abordagens multivariadas como essa produzem estimadores de parâmetros não viesados com taxas de erro tipo 1 bem controladas para fenótipos binários e contínuos (Werme et al. 2022).

4 Aplicação prática

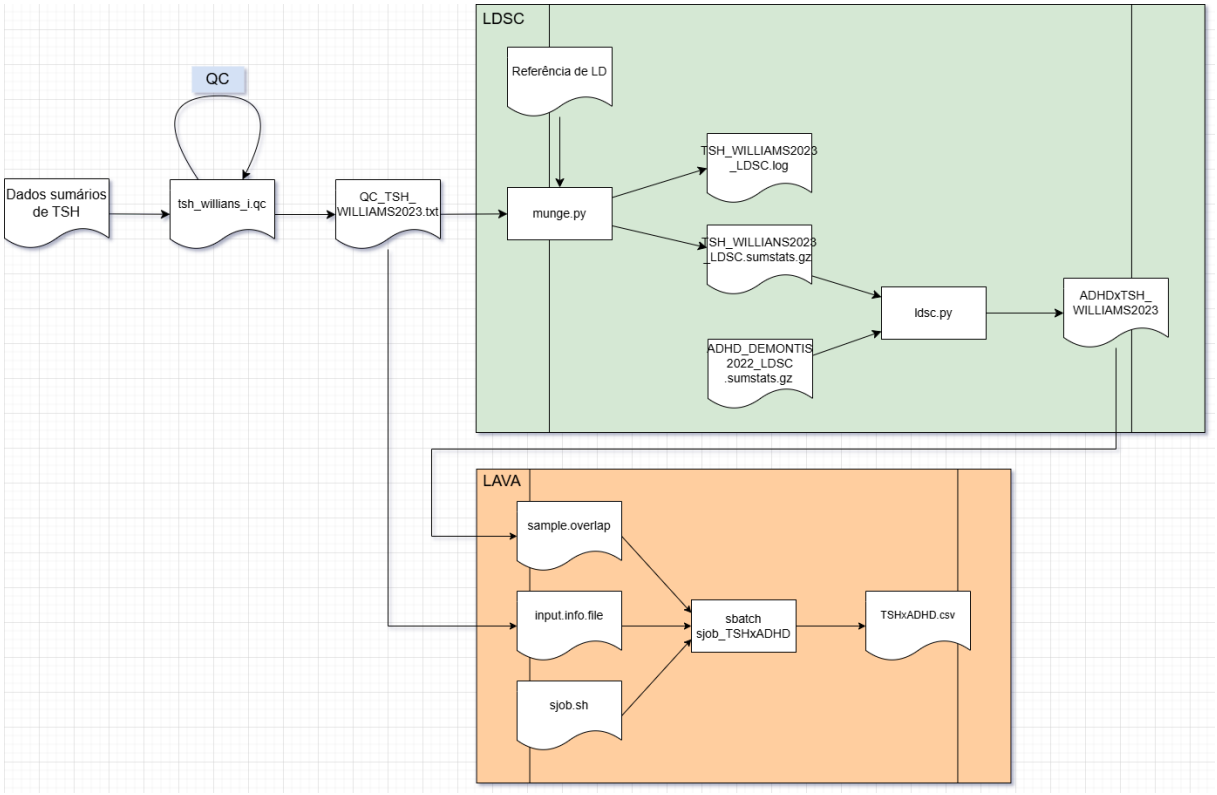

Figura 21: Fluxo da aplicação prática para os dados sumários de níveis de TSH. O mesmo processo é feito para a base de Hipotireoidismo.

O hormônio TSH regula a produção de hormônios da glândula tireoideana, essenciais para diversas funções metabólicas do organismo. O artigo Williams et al. 2023 apresenta os resultados e disponibiliza as estatísticas sumárias de um GWAS sobre níveis do hormônio TSH. O tamanho da amostra nesse estudo é de 247.107 indivíduos. Seu arquivo original com as estatísticas sumárias é nomeado e referenciado por `GCST90296333_rsID.txt`.

Já o artigo Mathieu et al. 2022 trata sobre hipotireoidismo, uma patologia na qual a concentração de hormônios produzidos pela tireoide é insuficiente. Nesse estudo, a amostra possui 51.194 casos e 443.383 controles, com tamanho total de 494.577 indivíduos. O arquivo original é nomeado como `categorical-20002-both_sexes-1226.tsv`.

No Laboratório de Genômica Fisiológica da Saúde Mental (*PhysioGen Lab*) do ICB - USP, foram executados os passos são documentados nessa seção, e obtidos os resultados apresentados juntamente com as conclusões. Primeiro, foi observado como os dados públicos de GWAS disponibilizados por Williams et al. 2023 e Mathieu et al. 2022 são organizados nos arquivos. Após, foi realizado o Controle de Qualidade desses dados. Por fim, executadas as análises de Correlação Global (LDSC) e Correlação Local (LAVA), em relação a uma base de dados sobre Transtorno de Déficit de Atenção com Hiperatividade (TDAH, ou ADHD em inglês).

4.1 Estrutura dos Dados

4.1.1 TSH

Os dados de cada linha dos arquivos dizem respeito a um SNP e são dispostos em colunas. O arquivo `GCST90296333_rsID.txt` possui 57.530.422 linhas, e o seguinte cabeçalho

```
CHR  BP  SNP  effect_allele  other_allele  beta  standard_error
   effect_allele_frequency  p_value  ci_lower  ci_upper  variant_id  n
```

que deve ser rearranjado para a ordem padrão do laboratório,

```
SNP  CHR  BP  A1  A2  BETA  SE  P  EAF  MAF  N  INFO
```

em que

1. `SNP` é o identificador da posição do SNP no cromossomo,
2. `CHR` $\in \{1, 2, \ldots, 23\}$, é o número do cromossomo,
3. `BP` é o par de bases,
4. `A1` $\in \{A, C, T, G\}$ é o alelo de efeito, ou seja, o alelo que possui uma variante significativa,
5. `A2` $\in \{A, C, T, G\}$ é o outro alelo,
6. `BETA` $\in \mathbb{R}$ é o valor de efeito estimado,
7. `SE` é o erro padrão cometido ao estimar o `BETA`,
8. `P` é o valor P,
9. `EAF` $\in [0, 1]$, é a frequência do alelo de efeito,
10. `MAF` $\in [0, 1]$, é a frequência do alelo com menor frequência (pode ser igual a `EAF` ou seu complementar),
11. `N` é o tamanho amostral, e
12. `INFO` é a qualidade de imputação do SNP.

Isso é feito em um novo arquivo chamado `tsh_willians_1.qc`, por meio do comando no terminal

```
tail -n +2  GCST90296333_rsID.txt | awk -v OFS="\t" '{print $3, $1, $2,
   toupper($4), toupper($5), $6, $7, $9, $8, ($8 < 1-$8 ? $8 : 1 - $8),
   $13}' > tsh_willians_1.qc
```

Por enquanto, o arquivo deve se manter sem cabeçalho, mas na nova ordem.

Note que nesse caso, o arquivo original não possui a coluna `INFO`, além disso são descartadas as colunas de `ci_lower`, `ci_upper` e `variant_id`, e é adicionada a coluna de `MAF`.

4.1.2 HYPO

Um processo análogo é feito para o arquivo `categorical-20002-both_sexes-1226.tsv`, que possui 28.987.534 linhas e o seguinte cabeçalho

```
chr pos ref alt af_cases_meta_hq af_controls_meta_hq beta_meta_hq
    se_meta_hq neglog10_pval_meta_hq neglog10_pval_heterogeneity_hq
    af_cases_meta af_controls_meta beta_meta se_meta
   neglog10_pval_meta neglog10_pval_heterogeneity af_cases_AFR
   af_cases_CSA af_cases_EUR af_controls_AFR af_controls_CSA
   af_controls_EUR beta_AFR beta_CSA beta_EUR se_AFR se_CSA se_EUR
   neglog10_pval_AFR neglog10_pval_CSA neglog10_pval_EUR
   low_confidence_AFR low_confidence_CSA low_confidence_EUR
```

As colunas necessárias para o padrão,

```
SNP CHR BP A1 A2 BETA SE P EAF MAF N INFO
```

são selecionadas e reordenadas em um novo arquivo chamado `hypo_mayxed_UKBB_1.qc`, pelo comando

```
tail -n +2 categorical-20002-both_sexes-1226.tsv | awk -v OFS="\t" '{
   print $1, $2, toupper($4), toupper($3), $25, $28, $31, EAF=($19*20563
    + $22*399910)/(420473), (EAF < 1-EAF ? EAF : 1 - EAF), 420473}' >
   hypo_mayxed_UKBB_1.qc
```

4.2 Controle de Qualidade

Será feito o controle de qualidade para os arquivos de Williams et al. 2023 e Mathieu et al. 2022.

A cada passo do controle de qualidade, os dados são armazenados em um novo arquivo no mesmo diretório, por exemplo `tsh_willians_i.qc` para `i`$\in \{1, 2, 3, \dots\}$. Também é registrado, à parte, o número de linhas daquele arquivo, para conferência futura.

Deve-se garantir que só existam os símbolos A, C, T ou G nas colunas A1 e A2. Para isso, são removidas as linhas em que existam símbolos diferentes, por meio do comando:

```
cat tsh_willians_1.qc | awk '($4=="A" || $4=="T" || $4=="C" || $4=="G
   ") && ($5=="A" || $5=="T" || $5=="C" || $5=="G") {print}' >
   tsh_willians_2.qc
```

São removidos SNPs de cromossomos sexuais, que é o cromossomo de número 23:

```
cat tsh_willians_2.qc | awk '$2>=1 && $2<=22 {print}' > tsh_willians_3.
   qc
```

São removidos eventuais SNPs duplicados:

```
cat tsh_willians_3.qc | awk '{seen[$1]++; if(seen[$1]==1){ print}}' >
   tsh_willians_4.qc
```

São removidos SNPs ambíguos:

```
cat tsh_willians_4.qc | awk '!( ($4=="A" && $5=="T") || ($4=="T" && $5
   =="A") || ($4=="G" && $5=="C") || ($4=="C" && $5=="G"))  {print}' >
   tsh_willians_5.qc
```

E remove-se os SNPs cuja frequência do alelo menos frequente (MAF) seja menor que 1%, o que impede a presença de variantes raras, e mantém apenas os SNPs comuns:

```
cat tsh_willians_5.qc | awk '$10 >= 0.01 {print}' > tsh_willians_6.qc
```

Também deveriam ser removidos os SNPs com qualidade de imputação menor que 80%, mas o arquivo desse exemplo não possui a coluna INFO.

4.2.1 TSH

Ao final, o arquivo possui 7.526.341 linhas, um número substancialmente menor que as cerca de 57 milhões iniciais.

Agora, o arquivo é renomeado no formato padrão do laboratório, `QC_TSH_WILLIAMS2023.txt` e é inserido o cabeçalho na primeira linha.

4.2.2 HYPO

Esse procedimento é repetido para o arquivo `hypo_mayxed_UKBB_i.qc`, e observa-se uma diminuição de 28.987.534 linhas para 7.098.780.

Um adendo, a coluna com o valor P nesse arquivo apresentava $-\log_{10}(p)$ ao invés do valor absoluto de P. Portanto foi necessária a conversão por $1/10^i$, sendo i o índice da coluna.

```
awk 'BEGIN {OFS="\t"} {print $1, $2, $3, $4, $5, $6, $7, $8, $9, $10,
   $11, 1/10^$8}' hypo_mayxed_UKBB_6.txt > hypo_mayxed_UKBB_7Pnew.txt
```

Concluído o processo de tratamento de dados, podem ser executados os métodos de correlação.

4.3 Correlação Global (LDSC)

O LDSC é uma ferramenta para cálculo da herdabilidade e correlação genética a partir de dados das estatísticas sumárias de GWAS.

O objetivo desta prática é executar o método seguindo os tutoriais disponíveis em `https://github.com/bulik/ldsc/wiki`, referentes aos usos do LDSC, no subsistema Ubuntu em Windows. Nesse mesmo repositório, também estão disponibilizados scripts em Python que serão utilizados.

Para a execução, é necessário ter o ambiente Ubuntu 20.04, e os *softwares* Anaconda 3 para Linux e PLINK instalados. Quantos aos arquivos, serão usados os que passaram pelo Controle de Qualidade, visto na seção anterior, e um arquivo para referência

de desequilíbrio de ligação (LD) para os cromossomos, chamado `1kg_eur.tar.bz2`, disponibilizado pelo Broad Institute em `https://data.broadinstitute.org/alkesgroup/LDSCORE/1kg_eur.tar.bz2`.

É esperada uma alta correlação entre TSH e hipotireoidismo (como apresentado em Williams et al. 2023), justamente porque os níveis do hormônio determinam o diagnóstico. Ao invés disso, cada uma das amostras será relacionada com uma base de Transtorno do Déficit de Atenção com Hiperatividade (TDAH, ou ADHD em inglês), referenciada pelo arquivo `ADHD_DEMONTIS2022.txt`, de Demontis et al. 2023.

Cada arquivo advindo do Controle de Qualidade deve ser rearranjado para o seguinte cabeçalho

```
SNP   A1   A2   N   OR/BETA   P VALUE   MAF(opcional)   SNP IQ(opcional)
```

que denota respectivamente um identificador do SNP, alelo 1 (de efeito) e alelo 2 (outro alelo), número amostral, efeito, valor P, menor frequência alélica e uma assinatura.

4.3.1 Estimação de Escore LD

Primeiro, é necessário ativar o ambiente ldsc, pelo comando

```
    conda activate ldsc
```

no terminal.

Em seguida, será rodado um script em Python, chamado "munge". Esse script é responsável por uniformizar os arquivos no formato do LDSC, tanto na amostra alvo (e.g., `TSH_WILLIANS2023`) quanto na base (e.g., `ADHD_DEMONTIS2022`). Ele garante que os arquivos contenham apenas informações consistentes e de alta qualidade para que sejam utilizadas nas análises subsequentes, ou seja, também representa uma forma de controle de qualidade.

Esse script requer as colunas organizadas e nomeadas pelo cabeçalho:

```
SNP   A1   A2   N   OR/BETA   P VALUE   MAF*   SNP IQ*
```

E então, são executados os comandos

```
./munge_sumstats.py \
--sumstats /path/TSH_WILLIANS2023_LDSC.txt \
--ignore EAF \
--out munge_data/TSH_WILLIANS2023_LDSC \
--chunksize 500000 \
--merge-alleles w_hm3.snplist
```

para a amostra alvo, e

```
./munge_sumstats.py \
--sumstats /path/ADHD_DEMONTIS2022.txt \
--ignore EAF \
--out munge_data/ADHD_DEMONTIS2022_LDSC \
```

```
--chunksize 500000 \
--merge-alleles w_hm3.snplist
```

para a amostra base.

Esse comando `./munge_sumstats.py` chama o script, seguido de alguns parâmetros:

- `--sumstats` referencia o arquivo. No caso, o de TSH resultante do controle de qualidade, com o cabeçalho específico para LDSC, nomeado por `TSH_WILLIAMS2023_LDSC`;
- `--out` indica o caminho da pasta em que serão gerados os novos arquivos;
- `--chunksize` define o tamanho do bloco para processamento. Em vez de processar todo o arquivo de uma vez, o script processará o arquivo em blocos de 500.000 linhas por vez. Isso é útil para reduzir o uso de memória e acelerar o processamento, especialmente quando se trabalha com grandes arquivos de dados;
- `--merge-alleles` indica que o script deve mesclar os alelos com base em uma lista fornecida, para garantir que os alelos sejam compatíveis ou corretamente atribuídos durante esse processo de formatação.

Após 33,09s é obtido como saída um *log*, em `.log`, e um arquivo zipado com os dados formatados, em `.sumstats.gz`.

```
*********************************************************************
* LD Score Regression (LDSC)
* Version 1.0.1
* (C) 2014-2019 Brendan Bulik-Sullivan and Hilary Finucane
* Broad Institute of MIT and Harvard / MIT Department of Mathematics
* GNU General Public License v3
*********************************************************************
(log parcialmente omitido)

Metadata:
Mean chi^2 = 1.676
Lambda GC = 1.198
Max chi^2 = 1424.989
4051 Genome-wide significant SNPs (some may have been removed by
   filtering).

Conversion finished at Wed Nov 20 22:59:06 2024
Total time elapsed: 33.09s
```

Deve ser verificado se a média da qui-quadrado é maior que 1,02. Caso contrário, é indicado que a summary a ser rodada não é adequada para utilizar a regressão com Escore LD. Nesse, caso, obtivemos 1,67.

Agora, com o novo arquivo formatado, é executado o script "ldsc", que realiza a análise de correlação genética entre os dois traços (TSH e TDAH), pelo método LDSC:

```
./ldsc.py \
--rg munge_data/TSH_WILLIANS2023_LDSC.sumstats.gz,munge_data/
   ADHD_DEMONTIS2022_LDSC.sumstats.gz \
--ref-ld-chr eur_w_ld_chr/ \
--w-ld-chr eur_w_ld_chr/ \
--out  rg_cross-trait/ADHDxTSH_WILLIANS2023_Yago
```

em que `./ldsc.py` chama o script, com os seguintes parâmetros

- `--rg` referencia os dois arquivos formatados que serão submetidos à análise;
- `--ref-ld-chr` especifica o caminho para os arquivos de referência de LD. O script usará esses arquivos para realizar cálculos de LD na análise;
- `--w-ld-chr` especifica o caminho para os arquivos de pesos de LD, que contêm os pesos utilizados para ajustar a análise de correlação genética;
- `--out` indica o caminho do novo arquivo gerado.

4.3.2 Resultados de TSH x TDAH

Após 13,33s, obtemos os resultados a seguir, interpretados.

```
 Herdability of phenotype 1
 --------------------------
 Total Observed scale h2: 0.1244 (0.0181)
 Lambda GC: 1.2005
 Mean Chi^2: 1.6853
 Intercept: 1.0815 (0.0583)
 Ratio: 0.1189 (0.0851)
```

A herdabilidade é a proporção da variação fenotípica que pode ser explicada pela variação genética, e a escala h^2 é o valor da herdabilidade observada para o fenótipo 1. O valor obtido sugere que aproximadamente 12,44% da variação fenotípica do fenótipo 1 pode ser atribuída à variação genética. O número entre parênteses ($1{,}81 \times 10^{-2}$) representa o erro padrão dessa estimativa.

Lambda GC é o fator de correção para o desvio da distribuição de qui-quadrado devido a efeitos não genéticos, como problemas na imputação. O valor 1,20 sugere que há um pequeno ajuste necessário para considerar esse viés.

Mean χ^2 é o valor médio da estatística qui-quadrado para todas as variantes analisadas. Um valor próximo de 1 indica que a análise não está com viés.

O intercepto da regressão é 1,08, com erro padrão de $5{,}83 \times 10^{-2}$. Isso indica o ajuste que está sendo feito para a estimativa da herdabilidade.

Ratio é a razão entre o intercepto e o valor esperado, com erro padrão. Ela indica o ajuste entre a variação genética observada e a estimada pelo modelo.

```
Herdability of phenotype 2/2
---------------------------
Total Observed scale h2: 0.0946 (0.0045)
Lambda GC: 1.3581
Mean Chi^2: 1.4502
Intercept: 1.0295 (0.0095)
Ratio: 0.0655 (0.0212)
```

A herdabilidade do fenótipo 2 é $9{,}46\times10^{-2}$, o que significa que 9,46% da variação fenotípica do fenótipo 2 é atribuída à variação genética. A estimativa é mais precisa, com um erro padrão de $4{,}5\times10^{-3}$.

O *Lambda GC* (1,35) sugere que o fenótipo 2 apresenta um pouco mais de viés relacionado a efeitos não genéticos ou problemas no controle da qualidade da amostra.

O intercepto para o fenótipo 2 é 1,03, com erro padrão de $9{,}5\times10^{-3}$, o que indica um bom ajuste do modelo.

A razão (*ratio*) de $6{,}55\times10^{-2}$ mostra que o ajuste entre a variação genética observada e a estimada para o fenótipo 2 é menor do que para o fenótipo 1.

```
Genetic Covariance
------------------
Total Observed scale gencov: -0.0078 (0.003)
Mean z1*z2: -0.0405
Intercept: -0.0018 (0.0069)
```

A covariância genética observada entre os dois fenótipos é muito próxima de zero ($-7{,}8\times10^{-3}$), com erro padrão de 3×10^{-3}. Isso sugere que há pouca ou nenhuma covariância genética entre os dois fenótipos, ou seja, as variantes genéticas que afetam um fenótipo não têm um efeito consistente no outro.

Já *Mean* z_1z_2 é o valor médio do produto da estatística de teste z para as variantes genéticas entre os dois fenótipos, que também sugere uma correlação genética muito pequena.

O intercepto para a covariância genética é $-1{,}8\times10^{-3}$, próximo de zero, com erro padrão de $-6{,}9\times10^{-3}$, indicando que o modelo não encontrou uma covariância genética substancial.

```
Genetic Correlation
-------------------
Genetic Correlation: -0.0723 (0.0275)
Z-score: -2.6265
P: 0.0086
```

A correlação genética entre os fenótipos 1 e 2 é $-7{,}23\times10^{-2}$, com erro padrão de $2{,}75\times10^{-2}$. O valor negativo sugere que existe uma correlação inversa, mas muito pequena, entre os dois fenótipos. A significância estatística dessa correlação é dada pelo escore Z e pelo valor P.

O *Z-score* (ou escore Z) indica a significância da correlação genética em comparação a uma região crítica construída para a distribuição Normal Padrão. Um valor de $-2{,}62$ é significativo em termos estatísticos e sugere que a correlação genética entre os dois fenótipos é substancial, mesmo que pequena.

Similarmente, o valor P indica a probabilidade de que a correlação genética observada seja devida ao acaso. Um valor de $8{,}6\times10^{-3}$ é menor que o nível tradicional de significância de 5%, indicando que a correlação genética entre os dois fenótipos é estatisticamente significativa.

4.3.3 Conclusão de TSH x TDAH

Ambos os fenótipos têm herdabilidades significativas, mas o fenótipo 1 tem uma herdabilidade maior (12,44%) em comparação com o fenótipo 2 (9,46%).

A correlação genética entre os dois fenótipos (TSH e TDAH) é estatisticamente significativa ($P = 8{,}6\times10^{-3}$), mas negativa e pequena ($-7{,}23\times10^{-2}$). Isso sugere que os fatores genéticos que afetam TSH e TDAH são, em certa medida, diferentes, embora não completamente independentes.

A covariância genética entre os dois fenótipos é muito baixa, o que reforça a ideia de que as variantes genéticas que influenciam cada fenótipo têm um efeito mínimo no outro.

Portanto, embora haja uma correlação genética significativa, os efeitos genéticos entre os dois fenótipos são relativamente fracos e complexos.

4.3.4 Resultados de HYPO x TDAH

Após 33,22s, obtivemos o log:

```
*********************************************************************
* LD Score Regression (LDSC)
* Version 1.0.1
* (C) 2014-2019 Brendan Bulik-Sullivan and Hilary Finucane
* Broad Institute of MIT and Harvard / MIT Department of Mathematics
* GNU General Public License v3
*********************************************************************
Call:
./munge_sumstats.py \
--out munge_data/hypo_mayxed_UKBB_LDSC \
--merge-alleles w_hm3.snplist \
--chunksize 500000 \
--sumstats /home/yago/ldsc/HYPOs/hypo_mayxed_UKBB_LDSC.txt \
```

```
--ignore EAF

Interpreting column names as follows:
N:      Sample size
A1:     Allele 1, interpreted as ref allele for signed sumstat.
P:      p-Value
BETA:   [linear/logistic] regression coefficient (0 --> no effect; above
    0 --> A1 is trait/risk increasing)
A2:     Allele 2, interpreted as non-ref allele for signed sumstat.
SNP:    Variant ID (e.g., rs number)

Reading list of SNPs for allele merge from w_hm3.snplist
Read 1217311 SNPs for allele merge.
Reading sumstats from /home/yago/ldsc/HYPOs/hypo_mayxed_UKBB_LDSC.txt
   into memory 500000 SNPs at a time.
Read 7098779 SNPs from --sumstats file.
Removed 5917851 SNPs not in --merge-alleles.
Removed 1 SNPs with missing values.
Removed 0 SNPs with INFO <= 0.9.
Removed 0 SNPs with MAF <= 0.01.
Removed 0 SNPs with out-of-bounds p-values.
Removed 0 variants that were not SNPs or were strand-ambiguous.
1180927 SNPs remain.
Removed 0 SNPs with duplicated rs numbers (1180927 SNPs remain).
Removed 0 SNPs with N < 280315.333333 (1180927 SNPs remain).
Median value of BETA was -5.651e-06, which seems sensible.
Removed 181 SNPs whose alleles did not match --merge-alleles (1180746
   SNPs remain).
Writing summary statistics for 1217311 SNPs (1180746 with nonmissing
   beta) to munge_data/hypo_mayxed_UKBB_LDSC.sumstats.gz.

Metadata:
Mean chi^2 = 1.485
Lambda GC = 1.214
Max chi^2 = 754.615
2572 Genome-wide significant SNPs (some may have been removed by
   filtering).

Conversion finished at Tue Nov 26 17:04:33 2024
Total time elapsed: 33.22s
```

e após 14,07s, os resultados a seguir.

```
Herdability of phenotype 1
---------------------------
Total Observed scale h2: 0.0478 (0.0056)
Lambda GC: 1.2136
Mean Chi^2: 1.4507
```

```
Intercept: 1.0261 (0.0197)
Ratio: 0.0579 (0.0436)

Herdability of phenotype 2/2
---------------------------
Total Observed scale h2: 0.0946 (0.0045)
Lambda GC: 1.3581
Mean Chi^2: 1.4503
Intercept: 1.0295 (0.0095)
Ratio: 0.0654 (0.0211)

Genetic Covariance
------------------
Total Observed scale gencov: 0.0091 (0.0023)
Mean z1*z2: 0.0518
Intercept: 0.003 (0.007)

Genetic Correlation
-------------------
Genetic Correlation: 0.1358 (0.037)
Z-score: 3.6684
P: 0.0002
```

4.3.5 Conclusão de HYPO x TDAH

Os resultados indicam uma correlação genética significativa ($P = 2{\times}10^{-4}$), mas moderada ($0{,}13 \pm 0{,}03$), entre o hipotireoidismo e o TDAH, sugerindo que ambos compartilham alguns fatores genéticos subjacentes. As herdabilidades observadas dos dois fenótipos ($h_1^2 = 4{,}78{\times}10^{-2}$ e $h_2^2 = 9{,}46{\times}10^{-2}$) indicam que fatores genéticos explicam apenas uma pequena parte da variabilidade de cada condição. Já os valores de Lambda GC (1,21) e o intercepto ($1{,}02 \pm 0{,}02$) revelam uma possível confusão devido a fatores externos. Conclui-se que a associação genética é consideravelmente fraca no contexto global.

4.4 Correlação Local (LAVA)

O LAVA é uma ferramenta para análise da correlação genética local, e deve ser instalado pelo tutorial descrito em `https://github.com/josefin-werme/LAVA`.

É necessário preparar três arquivos para a execução do programa

1. sample.overlap: Arquivo a ser montado a partir do resultado de intercepto da covariância genética, do LDSC, para os fenótipos.

```
            ADHD    HypoMax
   ADHD     1       0.003
   HypoMax  0.003   1
```

2. input.info.file: Arquivo com número de casos, número de controles, número total, e o caminho para cada arquivo de dados sumários.

```
phenotype    cases    controls     total    filename
ADHD         38691    186843       225534   /path/ADHD_DEMONTIS2022.txt
HypoMax      20563    399910       420473   /path/hypo_mayxed_UKBB_8.
   txt
```

3. sjob.sh: Arquivo com as instruções para delegar uma tarefa no escalonador de tarefas Slurm, utilizado em um ambiente virtual para criação de máquinas virtuais.

O arquivo sjob.sh chama o script do programa LAVA, os arquivos de dados sumários dos fenótipos estudados, e os demais arquivos advindos da própria instalação

```
srun Rscript /home/physiogenlab/LAVA/LAVA_Rscript.r "/home/physiogenlab/
   LAVA/g1000_eur" \
"/home/physiogenlab/LAVA/LAVA_2500loci.txt" \
"/home/yago/LAVA_HYPOMAYXEDxADHD/input.info.file" \
"/home/yago/LAVA_HYPOMAYXEDxADHD/sample_overlap" \
"ADHD;HypoMayx" \
"/path/LAVA_HYPOMAYXEDxADHD/results/26.11.2024-ADHDxHypoMayx")
```

Após essa preparação dos arquivos necessários, deve-se rodar a análise a partir do seguinte comando

```
sbatch sjob_HYPOMAYXEDxADHD.sh
```

E os resultados podem ser visualizados em uma planilha do Excel ou Google Sheets. Nas Figuras 22 e 23, são destacado apenas os locus significativos; os demais apresentaram um valor P maior que o nível de significância de 5%.

4.4.1 Resultados de TSH x TDAH

locus	chr	start	stop	n,snps	n,pcs	phen1	phen2	rho	rho,lower	rho,upper	r2	r2,lower	r2,upper	p	P-Bonf (p*111)
945	6	21295865	22199713	1733	227	ADHD	TSH	-0,56	-0,91	-0,27	0,32	0,07	0,82	3,24E-04	3,59E-02

Figura 22: Resultado da correlação local entre os fenótipos de nível de TSH e diagnóstico de TDAH.

A análise LAVA traz informações sobre a relação entre dois fenótipos (TSH e TDAH), em 111 locus. Em destaque na Figura 22, são exibidos os resultados de um locus específico, o único significativo dentre todos os analisados. O cabeçalho e seus valores correspondentes são explicados:

- `locus`: O número 945 é um identificador único para o locus analisado.
- `chr`: O locus está localizado no cromossomo 6.

- `start` e `stop`: Indicam as coordenadas genômicas iniciais e finais do intervalo onde o locus está situado (de 21.295.865 a 22.199.713, no cromossomo 6).
- `n,snps`: O número de SNPs analisados neste locus, 1.733.
- `n,pcs`: O número de componentes principais (PCs) usados na análise, 227. Os PCs são frequentemente usados em estudos genéticos para reduzir a dimensionalidade dos dados e capturar outras fontes de variação principais.
- `phen1`: O primeiro fenótipo analisado, que é o transtorno de déficit de atenção e hiperatividade (TDAH).
- `phen2`: O segundo fenótipo, que é o nível de hormônio estimulante da tireoide (TSH).

Agora, as estatísticas principais:

- `rho`: É o coeficiente de correlação entre os dois fenótipos (TDAH e TSH) para o locus analisado, e vale $-0{,}56$. Esse valor negativo sugere uma correlação inversa entre os dois fenótipos no locus em questão, ou seja, a genética que influencia positivamente um fenótipo tende a influenciar negativamente o outro.
- `rho,lower` e `rho,upper`: Os intervalos de confiança para o coeficiente de correlação. O intervalo de confiança de $-0{,}91$ a $-0{,}27$ indica que a correlação é estatisticamente significativa.
- `r2`: O valor de R^2, quantifica a proporção da variabilidade conjunta explicada pela correlação entre os fenótipos. O valor de 0,31 sugere que aproximadamente 31% da variabilidade conjunta dos fenótipos é explicada pela correlação.
- `r2,lower` e `r2,upper`: Os intervalos de confiança para R^2, e variam entre 0,07 e 0,82. Embora a média de R^2 seja 0,32, como o intervalo de confiança é amplo, a explicação da variabilidade pode variar consideravelmente dependendo das condições específicas da amostra.
- `p`: O valor P de significância estatística para o teste de correlação, que é $3{,}24\times10^{-4}$ (corrigido por Bonferroni, $3{,}59\times10^{-2}$). O valor P menor do que 0,05 define que a correlação observada é significativa, ou seja, a chance de que essa correlação tenha ocorrido por acaso é muito baixa, de acordo com o nível convencionado.

4.4.2 Conclusão de TSH x TDAH

A análise LAVA revelou uma correlação inversa significativa (P-Bonf = $3{,}59\times10^{-2}$) entre os fenótipos TDAH e TSH em um locus específico localizado no cromossomo 6 (ID: 945). Como o coeficiente de correlação é negativo ($\rho = -0{,}56$), as variações genéticas que

influenciam positivamente em um dos fenótipos influenciam negativamente o outro. O R^2 de 0,31 sugere que cerca de 31% da variabilidade conjunta é explicada por essa relação, contudo, o intervalo de confiança indica incerteza, devido a sua amplitude e por englobar valores muito próximos de zero ([0,07; 0,82]). Portanto, apesar do valor P, a conclusão sobre a relevância desse locus é incerta; seriam necessários outros estudos para explorá-lo.

4.4.3 Resultados de HYPO x TDAH

Os resultados são exibidos na Figura 23.

locus	chr	start	stop	n,snps	n,pcs	phen1	phen2	rho	rho,lower	rho,upper	r2	r2,lower	r2,upper	p	P-Bonf (p*159)
11	1	8580988	9475966	1725	252	ADHD	HypoMayx	0,68	0,38	1,00	0,46	0,14	1,00	4,47E-05	7,11E-03
1090	6	166853326	167714865	2478	217	ADHD	HypoMayx	-0,52	-0,81	-0,27	0,27	0,07	0,66	1,51E-04	2,40E-02
381	2	202504863	203639686	1770	187	ADHD	HypoMayx	0,57	0,28	0,92	0,32	0,08	0,84	3,05E-04	4,85E-02

Figura 23: Resultado da correlação local entre os fenótipos de hipotireoidismo e diagnóstico de TDAH.

4.4.4 Conclusão de HYPO x TDAH

De um total de 159 loci analisados, os resultados obtidos destacaram três loci com associações significativas (P-Bonf $< 0{,}05$) entre os fenótipos TDAH e hipotireoidismo. Dos loci identificados, um apresenta coeficiente de correlação negativo ($\rho_{1090} = -0{,}52$) e os outros dois, positivos ($\rho_{11} = 0{,}68$ e $\rho_{381} = 0{,}57$), portanto há influências em ambas as direções dentre diferentes loci na expressão de cada fenótipo. O valor de R^2 para cada locus sugere que cerca de 46%, 27% e 32% da variabilidade conjunta é explicada pela relação, contudo, o intervalo dos loci 1090 e 381 contém valores muito próximos de zero. Logo, apenas o locus 11 parece ter maior relevância; seriam necessários outros estudos para explorá-los.

5 Discussões

5.1 Hipóteses do Modelo

A modelagem de GWAS baseia-se em diversas hipóteses, e é necessário que sejam válidas, para a obtenção de resultados coerentes. Entre as principais suposições, estão: a independência das variáveis, a homoscedasticidade dos grupos genotípicos, e os efeitos entre genótipo e fenótipo serem aditivos, para associação por Regressão Linear. Contudo, pode ser difícil comprovar a validade dessas hipóteses na prática, e caso alguma não seja satisfeita, o experimento está sujeito a gerar resultados inválidos e levar a interpretações errôneas.

Dizemos que grupos de amostras são homocedásticos quando a variância é a mesma em todos eles, e essa propriedade deve ser assumida para que os resultados de uma Regressão sejam válidos. A homoscedasticidade é violada quando os indivíduos de um grupo genotípico possuem maior ou menor variabilidade quanto à média do fenótipo, em comparação aos demais grupos. Para resolver esse problema, uma sugestão é adotar testes não paramétricos ou modelos que possam lidar melhor com variâncias heterogêneas.

Além disso, diferentes loci podem interagir entre si, resultando em um efeito fenotípico não-aditivo, ou com efeito de dominância, o que viola a hipótese da linearidade. Para isso, uma solução seria testar a diferença entre as médias, ao invés de ajustar uma reta por Regressão Linear, para fenótipos contínuos. Assim, o teste retornaria um valor P referente a hipótese da diferença entre as médias de 0 e 1 ser a mesma que entre 1 e 2. Desse modo, seria evitado estudar a inclinação β de uma reta que não poderia ser ajustada por violação da aditividade.

Esses desafios enfatizam a necessidade de selecionar e validar modelos estatísticos apropriados, além de realizar ajustes ou correções antes da interpretação dos resultados. Consequentemente, seriam evitadas conclusões equivocadas sobre associações que não têm relação causal com o fenótipo, bem como a perda de associações reais.

5.2 Testes para Associação e Correlação

A regressão apresenta diversas vantagens, por possuir modelos paramétricos explícitos; algoritmos estáveis para estimação de parâmetros; fácil incorporação de covariantes como idade, sexo e ancestralidade; softwares confiáveis e amplamente disponíveis e bem documentados (Cantor, Lange e Sinsheimer 2010). Contudo, deve-se manter uma postura crítica aos conduzir os testes estatísticos para análise de associação e correlação com esse método.

Ao procurar por uma quantidade exacerbada de correlações, há o risco de identificar relações que são apenas advindas do acaso, o que pode comprometer a interpretação dos resultados. Além disso, as variantes identificadas por GWAS geralmente correspondem a

uma pequena fração das variantes realmente associadas, resultando em um poder preditivo limitado (Choi, Mak e O'Reilly 2020). Essa limitação também é relevante no contexto de Escores de Risco Poligênico (PRS), cujo objetivo é fazer previsões de risco, e para análise de PRS as descobertas ainda podem apresentar efeitos muito pequenos para servir de aplicação prática significativa.

5.3 Tamanho Amostral

O tamanho das amostras estudadas é fundamental para a descoberta de variantes genéticas significativas. Em especial, para aquelas cuja expressão de traços em patógenos sejam de fenótipos complexos ou de eitilogia multifatorial, como é o caso do TDAH (Silva et al. 2023), pois possuem diversas variantes associadas, com pesos relativamente baixos. Por isso, como visto, centenas de milhares ou milhões de SNPs são coletados e inspecionados, com o objetivo de conferir maior poder para detectar associações (Gumpinger et al. 2018).

Contudo, é necessário ter cuidado com o tamanho da amostra, pois em tentativas equivocadas de aumentar o número de casos, pode levar à sobre-amostragem. Esse fenômeno ocorre quando a proporção amostral de indivíduos afetados por uma doença ou condição rara é distorcida, desviando-se da frequência real da população. Nesse cenário, é comprometida a representatividade da amostra e consequentemente a validade dos resultados obtidos.

E ainda, quando o número de testes realizados é elevado, se aumenta o risco de descobertas falsas. Para tanto, é necessária a correção do nível α, ou a adoção de outras estratégias que visem controlar esse problema. Por exemplo, um método de meta-análise apresentado por Lin, Liang e Yang 2022, com integração numérica, que controla para falsos positivos mais eficientemente que o método de Fisher, assumindo independência.

5.4 Erro e Viés

O viés é um componente que existe, em algum grau, em qualquer experimento real conduzido. Sobretudo, o viés e a variância contribuem em conjunto para o risco em tomar uma decisão estatística. Assim, para a escolha do estimador, deve existir um balanço ou equilíbrio entre ambos, como sugere a literatura recente.

Similarmente, também é inevitável o erro cometido pelo procedimento de decisão de um testes de hipóteses. Apenas consegue-se garantias para controlar o erro tipo I, considerado o mais grave, a um nível arbitrário que é julgado eficaz pelo pesquisador. Em geral, utiliza-se o nível de 5%, mas esse valor pode ser reponderado, tendo em vista a quantidade de testes executados, e a relação inversa entre erros tipo I e tipo II.

Além disso, há fatores que contribuem para o erro de interpretação, por atuarem desviando o resultado observado do esperado. Nota-se especialmente termos de confusão,

heterogeneidade da população, dependência entre variáveis e sobre-amostragem. O conhecimento sobre a existência desses fatores corrobora a necessidade de um plano amostral e de um processo de tratamento de dados que vise minimizá-los, como já é delineado e implementado na metodologia de GWAS.

Por outro lado, há um viés eurocêntrico nas pesquisas, pois historicamente percebe-se uma predominância do uso de amostras genéticas para GWAS advindas de populações de origem europeia, o que limita a aplicabilidade dos resultados para outros grupos étnicos. Esse viés pode levar a descobertas de variantes genéticas que não são representativas ou que têm efeitos diferentes em populações não europeias. Dessa forma, são perpetuadas as disparidades étnicas na área da saúde, e portanto, faz-se necessária maior diversidade nos bancos de dados, experimentos e estudos.

5.5 Valor P e Estimativa de Efeito

Como reiterado, é necessário distinguir o significado e as definições dos resultados dos testes de associação. A estimativa para o efeito β, acompanhada de erros de estimação que conferem um intervalo de confiança, fornecem uma faixa na qual o verdadeiro valor do efeito se encontra com certa garantia. Além disso, o valor P determina a significância estatística do efeito. Juntas, essas informações ajudam a definir a relevância biológica do efeito observado.

O valor P indica a probabilidade de observar os dados ou algo mais extremo, assumindo que não há associação (sob a hipótese nula). Portanto, um valor suficientemente pequeno indica uma associação estatisticamente significativa, mas não garante que o efeito genotípico seja de fato causal. Para isso, é necessário considerar o contexto biológico ou outros fatores como confundidores.

Há críticas quanto ao significado e a dificuldade na interpretação do valor P, o que instiga uma abordagem Bayesiana para modelar GWAS. Essa questão pode ser mais aprofundada em obras como Ball 2013. Ademais, outro ponto interessante é a abordagem de teste agnóstico, ao invés do teste de hipóteses usual, pois ele possibilita controle simultâneo de erros tipo I e tipo II.

5.6 Causalidade

A correlação genética não fornece exatamente uma informação causal entre dois traços. Essa correlação pode ter diversas causas, como pleiotropia vertical (um traço causa outro), pleiotropia horizontal (uma variante influencia diretamente dois traços), LD induzindo pleiotropia horizontal (duas variantes diferentes que estão em LD influenciam um dos traços), poligenia induzindo pleiotropia (múltiplas variantes influenciam ambos os traços), ou uma mistura desses padrões (Werme et al. 2022).

Essa é uma área valiosa para estudo, pois possibilita elucidar vias biológicas compartilhadas, por exemplo, o complexo principal de histocompatibilidade (MHC; chr6:26-24Mb, 21 loci), uma região que exerce importante papel no sistema imune, auto-imunidade e no sucesso reprodutivo, e possui extensiva pleiotropia. É um tema que também viabiliza informar o significado funcional de resultados de GWAS e um melhor entendimento sobre a etiologia de doenças e traços complexos (Werme et al. 2022).

6 Conclusão

A colaboração e o compartilhamento de dados são o alicerce para o sucesso de GWAS, evidenciado pela prática comum de disponibilizar dados como estatísticas sumárias, o que motiva análises posteriores. Além disso, são acessíveis as ferramentas computacionais para conduzir estudos genéticos, muitas das quais são gratuitas e acompanham tutoriais em repositórios públicos.

Em GWAS, os métodos estatísticos utilizados são testes de associação, modelos lineares e logísticos mistos, análise de componentes principais e correção para múltiplos testes. A aplicação desses métodos permite identificar variantes genéticas associadas a características fenotípicas. Os testes de associação são realizados entre grupos genotípicos e fenótipos, utilizando modelos de Regressão Linear misto para fenótipos contínuos, e de Regressão Logística misto para fenótipos binários.

No entanto, esses métodos apresentam algumas limitações, como a necessidade de amostras grandes, risco de erros tipo I, desafios relacionados à heterogeneidade genética, dificuldade em capturar relações genéticas complexas e a validação das hipóteses do modelo. Embora os cálculos de probabilidade sejam facilmente realizados com os softwares apropriados, é essencial compreender os métodos utilizados para interpretar corretamente os resultados. Assim, com a compreensão plena dos métodos, é possível analisar erros, considerar pontos positivos e identificar necessidades de melhoria nas ferramentas computacionais existentes.

A integração entre pesquisadores matemáticos e biólogos é produtiva para a criação de projetos e descobertas na área. Como perspectiva futura, propõe-se a execução de testes que comparem as diferenças entre as médias dos grupos genotípicos, ao invés do teste usual quanto ao coeficiente de uma regressão. Além disso, também parece promissor o uso de testes não paramétricos, pois pode evitar falhas decorrentes de assumir suposições que não sejam válidas, acerca da distribuição da população. Essas estratégias buscam garantir que a análise estatística, dentro das limitações das modelagens e ferramentas computacionais, seja aplicada de maneira coerente e alinhada às dinâmicas da genética na saúde.

Portanto, é preciso manter uma postura crítica na execução das associações e no entendimento das características da população de estudo. Isso inclui uma análise cuidadosa das propriedades amostrais e dos estimadores, para o planejamento e a execução eficaz dos experimentos.

Referências

Bick, Alexander G. et al. (mar. de 2024). "Genomic data in the All of Us Research Program". Em: *Nature* 627.8003, pp. 340–346. ISSN: 1476-4687. DOI: 10.1038/s41586-023-06957-x. URL: https://doi.org/10.1038/s41586-023-06957-x.

Demontis, Ditte et al. (fev. de 2023). "Genome-wide analyses of ADHD identify 27 risk loci, refine the genetic architecture and implicate several cognitive domains". Em: *Nature Genetics* 55.2, pp. 198–208. ISSN: 1546-1718. DOI: 10.1038/s41588-022-01285-8.

Pereira Ciochetti, Nicolas et al. (2023). "Genome-wide association studies: utility and limitations for research in physiology". Em: *The Journal of Physiology* 601.14, pp. 2771–2799. DOI: https://doi.org/10.1113/JP284241.

Pirinen, Matti (2023). "Genome Wide Association Studies". Course at the University of Helsinki. URL: https://www.mv.helsinki.fi/home/mjxpirin/GWAS_course/.

Silva, Bruna Santos da et al. (jan. de 2023). "An overview on neurobiology and therapeutics of attention-deficit/hyperactivity disorder". Em: *Discover Mental Health* 3.1, p. 2. ISSN: 2731-4383. DOI: 10.1007/s44192-022-00030-1.

Williams, Alexander T et al. (out. de 2023). "Genome-wide association study of thyroid-stimulating hormone highlights new genes, pathways and associations with thyroid disease". Em: *Nature Communications* 14.1, p. 6713.

Lin, Yin-Chun, Yu-Jen Liang e Hsin-Chou Yang (jul. de 2022). "Evaluating statistical significance in a meta-analysis by using numerical integration". en. Em: *Comput Struct Biotechnol J* 20, pp. 3615–3620.

Mathieu, Samuel et al. (ago. de 2022). "Genetic association and Mendelian randomization for hypothyroidism highlight immune molecular mechanisms". en. Em: *iScience* 25.9, p. 104992.

Werme, Josefin et al. (mar. de 2022). "An integrated framework for local genetic correlation analysis". Em: *Nature Genetics* 54.3, pp. 274–282. DOI: 10.1038/s41588-022-01017-y.

Ni, Guiyan et al. (mai. de 2021). "A Comparison of Ten Polygenic Score Methods for Psychiatric Disorders Applied Across Multiple Cohorts". en. Em: *Biol Psychiatry* 90.9, pp. 611–620. DOI: 10.1016/j.biopsych.2021.04.018.

Uffelmann, Emil et al. (ago. de 2021). "Genome-wide association studies". Em: *Nature Reviews Methods Primers* 1.1, p. 59. ISSN: 2662-8449. DOI: 10.1038/s43586-021-00056-9.

Yoon, Sora et al. (mar. de 2021). "Powerful p-value combination methods to detect incomplete association". Em: *Scientific Reports* 11.1, p. 6980.

Choi, Shing Wan, Timothy Shin-Heng Mak e Paul F O'Reilly (set. de 2020). "Tutorial: a guide to performing polygenic risk score analyses". Em: *Nature Protocols* 15.9, pp. 2759–2772. DOI: 10.1038/s41596-020-0353-1.

Milet, Jacqueline et al. (nov. de 2020). "Mixed logistic regression in genome-wide association studies". Em: *BMC Bioinformatics* 21.1, p. 536.

Shah, Sonia et al. (jan. de 2020). "Genome-wide association and Mendelian randomisation analysis provide insights into the pathogenesis of heart failure". en. Em: *Nat Commun* 11.1, p. 163. DOI: 10.1038/s41467-019-13690-5.

Privé, Florian et al. (nov. de 2019). "Making the Most of Clumping and Thresholding for Polygenic Scores". en. Em: *Am J Hum Genet* 105.6, pp. 1213–1221.

Gumpinger, Anja C et al. (2018). "Methods and Tools in Genome-wide Association Studies". en. Em: *Methods Mol Biol* 1819, pp. 93–136.

Marees, Andries T. et al. (2018). "A tutorial on conducting genome-wide association studies: Quality control and statistical analysis". Em: *International Journal of Methods in Psychiatric Research* 27.2. DOI: 10.1002/mpr.1608.

Morettin, P.A. e W. de O. Bussab (2017). *Estatística básica 9ed.* Saraiva Uni. ISBN: 9788547220228. URL: https://books.google.com.br/books?id=dxf8zwEACAAJ.

Presumey, Jessy, Allison R Bialas e Michael C Carroll (jul. de 2017). "Complement System in Neural Synapse Elimination in Development and Disease". en. Em: *Adv Immunol* 135, pp. 53–79. DOI: 10.1016/bs.ai.2017.06.004.

Auton, Adam et al. (out. de 2015). "A global reference for human genetic variation". Em: *Nature* 526.7571, pp. 68–74. DOI: 10.1038/nature15393.

Bulik-Sullivan, Brendan et al. (set. de 2015). "An atlas of genetic correlations across human diseases and traits". en. Em: *Nat Genet* 47.11, pp. 1236–1241. DOI: 10.1038/ng.3406.

Bulik-Sullivan, Brendan K et al. (fev. de 2015). "LD Score regression distinguishes confounding from polygenicity in genome-wide association studies". en. Em: *Nat Genet* 47.3, pp. 291–295. DOI: 10.1038/ng.3211.

Corvol, Harriet et al. (set. de 2015). "Genome-wide association meta-analysis identifies five modifier loci of lung disease severity in cystic fibrosis". en. Em: *Nat Commun* 6, p. 8382. DOI: 10.1038/ncomms9382.

Tak, Yu Gyoung e Peggy J Farnham (dez. de 2015). "Making sense of GWAS: using epigenomics and genome engineering to understand the functional relevance of SNPs in non-coding regions of the human genome". Em: *Epigenetics & Chromatin* 8.1, p. 57. DOI: 10.1186/s13072-015-0050-4.

Sham, Pak C. e Shaun M. Purcell (mai. de 2014). "Statistical power and significance testing in large-scale genetic studies". Em: *Nature Reviews Genetics* 15.5, pp. 335–346. ISSN: 1471-0064. DOI: 10.1038/nrg3706. URL: https://doi.org/10.1038/nrg3706.

Ball, Roderick D (2013). "Designing a GWAS: power, sample size, and data structure". en. Em: *Methods Mol Biol* 1019, pp. 37–98.

Cantor, Rita M, Kenneth Lange e Janet S Sinsheimer (jan. de 2010). "Prioritizing GWAS results: A review of statistical methods and recommendations for their application". en. Em: *Am J Hum Genet* 86.1, pp. 6–22.

Mayo, Oliver (jun. de 2008). "A century of Hardy-Weinberg equilibrium". en. Em: *Twin Res Hum Genet* 11.3, pp. 249–256.

Hogg, R.V., J.W. McKean e A.T. Craig (2005). *Introduction to Mathematical Statistics*. Pearson education international. Pearson Education. ISBN: 9780130085078. URL: https://books.google.com.br/books?id=dX4pAQAAMAAJ.

Kutner, M. et al. (2004). *Applied Linear Statistical Models*. McGraw-Hill Companies, Incorporated. ISBN: 9780073108742. URL: https://books.google.com.br/books?id=BU7rjwEACAAJ.

Casella, G. e R.L. Berger (2002). *Statistical Inference*. Duxbury advanced series in statistics and decision sciences. Thomson Learning. ISBN: 9780534243128. URL: https://books.google.com.br/books?id=0x_vAAAAMAAJ.

Evans, David M., N.A. Gillespie e N.G. Martin (2002). "Biometrical genetics". Em: *Biological Psychology* 61.1, pp. 33–51. ISSN: 0301-0511. DOI: https://doi.org/10.1016/S0301-0511(02)00051-0. URL: https://www.sciencedirect.com/science/article/pii/S0301051102000510.

Bolfarine, H. e M.C. Sandoval (2001). *Introdução à inferência estatística*. Coleção Matematica Aplicada. SBM. ISBN: 9788585818135. URL: https://books.google.com.br/books?id=W7lAQwAACAAJ.

Makałowski, W (2001). "The human genome structure and organization". en. Em: *Acta Biochim Pol* 48.3, pp. 587–598.

Falconer, Douglas S. e Trudy F.C. Mackay (1995). *Introduction to Quantitative Genetics (3rd Edition)*. Pearson Education. ISBN: 9780758146656.