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

Métodos de colapsagem de variantes raras para mapeamento de QTLs

Lara Midena João

Trabalho de Conclusão de Curso

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

Métodos de colapsagem de variantes raras para mapeamento de QTLs

Lara Midena João
Orientadora: Daiane Aparecida Zuanetti

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

São Carlos
Fevereiro 2024

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

Rare variant collapsing methods for QTLs mapping

Lara Midena João
Advisor: Daiane Aparecida Zuanetti

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

São Carlos
February 2024

Resumo

O mapeamento de regiões no genoma associadas a traços quantitativos (QTLs) através de marcadores genéticos do tipo SNP tem sido um dos problemas centrais em Genética e Biologia Molecular e vários métodos de detecção e identificação de QTLs têm sido propostos na literatura. Neste trabalho, pesquisamos, estudamos e descrevemos os principais métodos que têm sido utilizados para esse fim em dados independentes na presença de variantes raras. Também destacamos suas vantagens e desvantagens e elencamos os algoritmos que já estão implementados e disponíveis para utilização. O desempenho das diferentes metodologias estudadas foi comparado via análise dos dados GAW 17.

Palavras-chave: *dados independentes, GAW 17, variantes raras.*

Abstract

The mapping of regions in the genome associated with quantitative traits (QTLs) through SNP-type genetic markers has been one of the central problems in Genetics and Molecular Biology and several methods of detection and identification of QTLs have been proposed in the literature. In this work, we research, study and describe the main methods that have been used for this purpose in independent data with the presence of rare variants. We also highlight its advantages and disadvantages and list the algorithms that are already implemented and available for use. The performance of the different methodologies studied was compared via GAW 17 data analysis.

Keywords: *GAW 17, independent data, rare variants.*

Sumário

Capítulo 1

Introdução

Um dos problemas centrais na Genética e Biologia Molecular é a detecção e identificação de regiões no genoma associadas a traços quantitativos (fenótipos) de seres vivos. Essas regiões genômicas são geralmente conhecidas como regiões de traços quantitativos (do inglês *quantitative trait loci*, QTLs) e suas posições e efeitos sobre o fenótipo de interesse são estimados através de marcadores genéticos, mais comumente do tipo SNP (do inglês *single nucleotide polymorphism*), dos indivíduos.

Para identificar a(s) região(ões) genômica(s) causadora(s) ou promotora(s) (os QTLs) do fenótipo de interesse, milhares ou milhões de SNPs são genotipados em amostras compostas de centenas ou milhares de indivíduos. Os genótipos dos SNPs são, então, vistos como covariáveis que podem afetar o fenótipo, considerado como a variável resposta. O fenótipo, quando se trata de uma variável contínua, é geralmente modelado com uma função linear dos efeitos aditivos e de dominância do genótipo dos SNPs e/ou suas interações de segunda, terceira ou maior ordem.

Muitos métodos têm sido propostos e estudados para identificar e selecionar os SNPs mais associados ao fenótipo de interesse, entre eles destacam-se o modelo de regressão linear simples entre o genótipo de cada SNP e o fenótipo em estudo e metodologias como LASSO (do inglês *Least Absolute Shrinkage and Selection Operator*, Tibshirani, 1996) e SPLS (do inglês *Sparse Partial Least Squares*, Chun, 2010), que permitem analisar os SNPs conjuntamente. No entanto, esses métodos geralmente não apresentam tão boa seleção de variáveis quando lidamos com variantes raras cuja frequência do alelo menor é baixa, caso bastante comum, uma vez que a maioria das variantes humanas são raras e as que apresentam impacto funcional, em geral, também tendem a apresentar menor tamanho de efeito.

Desse modo, algumas técnicas que realizam colapsagem de variantes raras estão sendo bastante utilizadas nesse contexto de identificação de SNPs que afetam o fenótipo dos indivíduos. Entre os métodos mais tradicionais, destacam-se, segundo Lee *et al.* (2014), os testes de Burden e SKAT (do inglês, *Sequence Kernel Association Test*).

Apesar do SKAT utilizar uma estrutura de modelo de regressão linear múltiplo, nessa metodologia são atribuídos diferentes pesos para o genótipo de diferentes SNPs, de acordo com a sua frequência alélica, e os coeficientes de regressão associados aos genótipos são considerados como efeitos aleatórios. Desse modo, esse método se trata de um modelo misto que possibilita uma maior identificação de variáveis raras, uma vez que analisa o efeito conjunto das variantes. Outra vantagem dessa metodologia é que ela é computacionalmente eficiente, flexível e válida quando o fenótipo do indivíduo é contínuo e também quando é dicotômico, com pequenas adaptações (Wu *et al.*, 2011).

O teste de Burden, por sua vez, refere-se a um modelo de regressão linear múltiplo em que todas as variáveis de uma região são consideradas associadas ao fenótipo e apresentam o mesmo tamanho de efeito. Desse modo, é testado se todas as variantes raras colapsadas de uma região afetam significativamente, com mesmo peso (na versão tradicional do teste) e direção, ou não o fenótipo de interesse (Lin, 2022; Morgenthaler e Thilly, 2007; Li e Leal, 2008; Aldisi *et al.*, 2023).

Como uma combinação dessas duas metodologias, surgiu o SKAT-O (do inglês, *Optimal Unified Test*), que pondera de forma ótima os testes de Burden e SKAT a fim de maximizar o poder, ou seja, minimizar a probabilidade de considerar erroneamente que os SNPs são não significativos para o fenótipo de interesse.

O objetivo desse trabalho é, portanto, analisar e comparar a performance dos três métodos: teste de Burden, SKAT e SKAT-O em selecionar variantes raras relevantes para um fenótipo em estudo. Os dados aqui analisados se tratam do conjunto de dados independentes GAW 17 (Almasy *et al.*, 2011).

O relatório está organizado como segue. No Capítulo 2, apresentamos conceitos genéticos importantes para o entendimento e contextualização do problema em análise. No Capítulo 3, descrevemos as metodologias estudadas e aplicadas para a identificação de SNPs raros relevantes e como o desempenho delas é analisado e comparado. O Capítulo 4 apresenta o conjunto de dados GAW 17 estudado nesse trabalho e sua respectiva análise descritiva. O Capítulo 5 apresenta os resultados obtidos pelas metodologias para cada uma das abordagens realizadas e no Capítulo 6 apresentamos as considerações finais e os

estudos futuros que podem ser realizados.

Capítulo 2

Fundamentos Genéticos

Antes de comentarmos sobre as metodologias estatísticas que serão consideradas nesse trabalho para seleção de variantes raras, a descrição de alguns conceitos genéticos é importante para melhor compreensão das análises e estudos. Nesse capítulo, portanto, apresentamos alguns conceitos que são baseados em Midena e Zuanetti (2023).

Todos os seres vivos, sejam eles animais ou plantas, são caracterizados pelo seu genoma (ou DNA) que é o conjunto de todos os genes presentes nos cromossomos dos seres vivos e que contém as características hereditárias responsáveis por determinar como será o desenvolvimento biológico de cada ser, uma vez que o genoma é a sequência completa do DNA.

O DNA é formado por uma dupla hélice composta por nucleotídeos, sendo que cada nucleotídeo apresenta uma base nitrogenada (adenina, timina, citosina ou guanina), uma pentose (açúcar) e um fosfato. As bases nitrogenadas que compõem o DNA são divididas em dois grupos - bases purinas ou púricas (adenina e guanina) e bases pirimídicas ou pirimidinas (citosina e timina) - e são ligadas por pontes de hidrogênio do seguinte modo: adenina (A) é ligada à timina (T) e citosina (C) é ligada à guanina (G). A Figura 2.1 ilustra essa estrutura.

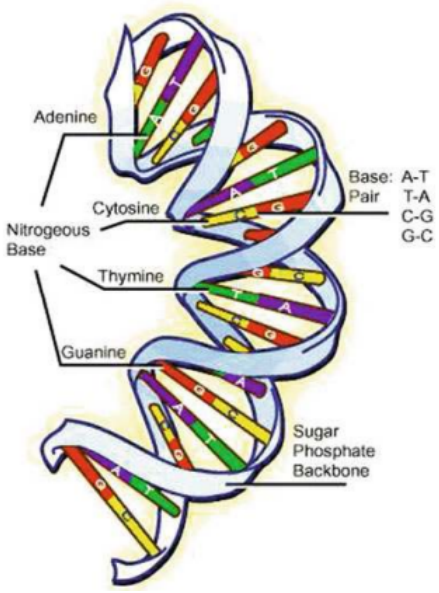

Figura 2.1: Fita de DNA ilustrando o emparelhamento de bases nitrogenadas complementares. Fonte: Laird e Lange (2011).

Os cromossomos, por sua vez, são longas sequências do DNA que contêm diversos genes. Em alguns organismos, denominados diploides, os cromossomos se manifestam aos pares e são chamados de homólogos, por apresentarem, em geral, sequências de DNA iguais. Os seres diploides são caracterizados por possuírem dois conjuntos de cromossomos completos, cada um proveniente de um progenitor.

Os seres humanos contém 23 pares de cromossomos, sendo 22 de cromossomos homólogos autossômicos e 1 de cromossomos sexuais. Os cromossomos sexuais são representados por XX em mulheres e por XY em homens, visto que o alelo materno sempre será X e o paterno pode ser X ou Y.

O gene, todavia, é uma sequência do DNA que é responsável pela transmissão e expressão de uma característica herdada geneticamente. Cada posição do gene é formada por dois alelos, um proveniente da mãe e outro do pai, que determinam como essa característica será expressa no indivíduo, ou seja, o alelo é a variação de cada base nitrogenada dentro de um gene.

A diversidade genética, também chamada de variabilidade genética, conforme descrito em Zuanetti (2023), refere-se à diversidade de alelos presentes em cada região do genoma dos indivíduos, de modo que duas medidas simples de variação genética são as frequências de alelos ou genótipos em diferentes partes do DNA.

O alelo é classificado de duas formas: dominante (quando ele determina uma carac-

terística mesmo em dose única) ou recessivo (só altera o fenótipo em dose dupla). Além dessa classificação, ele também pode ser caracterizado de acordo com a sua frequência. Segundo Zuanetti (2023), a frequência alélica de um específico tipo de alelo, em um grupo de indivíduos, é dada pela proporção de alelos de uma região que são desse específico tipo em relação ao total de alelos dessa região. Dessa forma, o alelo que aparece com menor frequência na população em uma específica região é denominado alelo menor, e o que aparece com maior frequência é denominado alelo maior. Sendo assim, um alelo (variante) é considerado raro quando sua frequência (MAF, do inglês *Minor Allele Frequency*) é muito baixa e menor que um limite pré-fixado. Ainda que não haja um consenso na literatura sobre o valor de MAF ideal para considerarmos um alelo como raro (variante rara), os dois mais comuns entre os autores são 5% e 1%, segundo Wu *et al.* (2011). Nesse trabalho, conforme proposto por Ionita-Laza *et al.* (2013), consideramos como rara a variante cujo MAF é inferior a 5%.

Marcadores genéticos, por sua vez, são pequenas sequências do DNA usadas a fim de diferenciar indivíduos e espécies. Os marcadores do tipo SNP geralmente representam o DNA no nível de uma base nitrogenada e identificam se mutações nessa única base (A, T, C ou G) podem ou não afetar o fenótipo. As mutações mais comuns são as transições de uma purina por outra purina (A por G ou G por A) ou de uma pirimidina por outra do mesmo grupo (C por T ou T por C). As transições entre grupos diferentes de bases também podem ocorrer, porém elas são mais incomuns. Além disso, sabe-se que em espécies não endogâmicas (espécies em que não ocorre união entre indivíduos geneticamente semelhantes e parentes) esse tipo de mutação é bastante frequente (Caetano, 2009). A Figura 2.2 ilustra como esses marcadores ocorrem em um determinado gene.

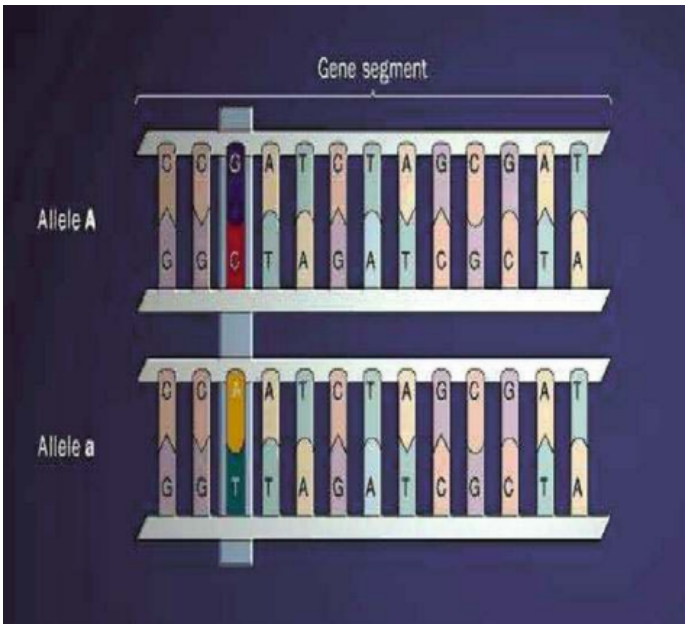

Figura 2.2: Ilustração de um marcador SNP em um par de cromossomos autossômicos. O terceiro par de bases dentro de um específico gene em um específico cromossomo apresenta variação, de modo que ela pode ser G-C ou A-T, e os rótulos A e a denotam os alelos. Fonte: Laird e Lange (2011).

O genótipo de uma posição no DNA, de um marcador genético tipo SNP por exemplo, é definido como sendo a constituição genética de um indivíduo nessa específica região e pode ser homozigoto (representado por letras iguais, por exemplo: AA e aa) ou heterozigoto (representado por letras diferentes, por exemplo: Aa). O fenótipo, por sua vez, é a manifestação observável de uma característica em um indivíduo, geralmente afetada e controlada pelo genótipo de regiões específicas no DNA. Para ilustrar, podemos utilizar o caso de polidactilia, em que o fenótipo é a presença de mais de cinco dedos no indivíduo e o genótipo que propicia essa característica é dominante (AA ou Aa) numa específica região do DNA. Doenças que, por sua vez, estão associadas aos efeitos de diversos genes são chamadas de doenças complexas. Elas ocorrem também em combinação com o estilo de vida e com os fatores ambientais.

Desse modo, um dos maiores desafios genéticos é identificar, dentro de uma quantidade enorme de marcadores genéticos tipo SNP, quais são os que realmente regulam e impactam um específico fenótipo em estudo, uma vez que é de nosso conhecimento que esses apresentam, muitas vezes, baixa diversidade genética e baixa frequência do alelo menor.

Capítulo 3

Métodos de colapsagem para variantes raras

Métodos de seleção tradicionais como os aplicados em modelos de regressão linear simples, regressão múltipla e o LASSO geralmente apresentam bom desempenho no caso de marcadores genéticos que possuem uma alta diversidade genética. Porém quando há variantes raras, esses métodos não apresentam bons resultados de seleção, de modo que outras metodologias foram propostas para identificá-las.

Nessa seção, apresentamos e descrevemos alguns métodos frequentemente utilizados para colapsagem de variantes raras para mapeamento de QTLs em dados independentes. São eles: os testes de Burden, SKAT e SKAT-O.

3.1 Teste de Burden

Com o objetivo de identificar e selecionar os SNPs que influenciam o fenótipo de interesse na presença de variantes raras, diversas técnicas de colapsagem dessas variantes estão sendo utilizadas, dentre elas destaca-se o teste de Burden como uma das mais tradicionais.

Nesse modelo, o genótipo de cada SNP é tido como variável preditora (covariável) e o fenótipo é considerado como variável resposta. Seja $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$ um traço quantitativo a ser observado em n indivíduos e $\mathbf{G}_i = (G_{i1}, \ldots, G_{ip})$ o genótipo de p SNPs com variante rara para o i-ésimo indivíduo. O valor esperado do fenótipo Y_i do i-ésimo

indivíduo pode ser modelado pelo seguinte modelo de regressão múltipla:

$$\mathbb{E}[Y_i] = \alpha_0 + \mathbf{X}_i\boldsymbol{\alpha} + \sum_{k=1}^{p} \beta_k w_k G_{ik}, \tag{3.1}$$

em que α_0 é o intercepto, $\boldsymbol{\alpha}$ é o vetor de coeficientes de dimensão $(j \times 1)$ associado às j variáveis ambientais, genéticas, de comportamento ou ancestralidade presentes no vetor $\mathbf{X}$, β_k é o efeito aditivo do k-ésimo SNP raro no valor esperado do fenótipo, w_k é o peso associado ao genótipo do k-ésimo SNP raro do i-ésimo indivíduo, representado por G_{ik} e codificado como $0, 1$ ou 2 para AA, Aa ou aa, respectivamente, $k = 1, \ldots, p$ e $i = 1, \ldots, n$.

Segundo Training (2022), o teste de Burden assume que todas as p variáveis são associadas ao fenótipo e têm os mesmos efeitos, ou seja, apresentam $\beta_k = \beta$ na Equação (3.1) e, por conveniência, em sua versão tradicional, assumimos pesos $w_k = 1$ para todas elas, $k = 1, \ldots, p$. Logo, testamos se todas as variantes raras colapsadas de uma região afetam significativamente, com mesmo peso e direção, ou não o fenótipo de interesse (Lin, 2022; Morgenthaler e Thilly, 2007; Li e Leal, 2008; Aldisi *et al.*, 2023). Sendo assim, podemos reescrever a Equação (3.1) da seguinte forma:

$$\mathbb{E}[Y_i] = \alpha_0 + \mathbf{X}_i\boldsymbol{\alpha} + \beta D_i, \tag{3.2}$$

em que α_0 é o intercepto, $\boldsymbol{\alpha}$ é o vetor de coeficientes de dimensão $(j \times 1)$ associado às j variáveis ambientais, genéticas, de comportamento ou ancestralidade presentes no vetor $\mathbf{X}$, $D_i = \sum_{k=1}^{p} G_{ik}$ representa o número total de variantes raras (assumindo o a como alelo menor) de modo que resume toda a informação genética do i-ésimo indivíduo presente nos SNPs raros (com variante rara), $i = 1, \ldots, n$.

Nesse modelo, a significância conjunta dos SNPs em estudo pode ser analisada via um teste com as seguintes hipóteses:

H_0: $\beta = 0$ (o efeito dos SNPs não é significativo) contra

H_1: $\beta \neq 0$ (o efeito dos SNPs é significativo).

De acordo com Lee *et al.* (2014), temos que a estatística teste para verificar essas

hipóteses é dada por:

$$Q_{Burden} = \left(\sum_{k=1}^{p} w_k S_k \right)^2, \tag{3.3}$$

em que $S_k = \sum_{i=1}^{n} G_{ik}(Y_i - \hat{\mu}_i)$ indica o efeito do k-ésimo SNP raro nos n indivíduos e y_i e $\hat{\mu}_i$ representam, respectivamente, o valor observado e o valor predito sob H_0 do fenótipo do i-ésimo indivíduo, $i = 1, \ldots, n$. Novamente, podemos assumir peso $w_k = 1$ para todos os SNPs, mas também é comum na literatura assumirmos $w_k = \frac{1}{[MAF_k(1-MAF_k)]^{1/2}}$. Outra abordagem bastante comum e implementada no software R determina que os pesos w_k assumem o valor da densidade da distribuição $Beta(1, 25)$ calculada no MAF do SNP em análise. Observe que a estatística S_k é muito semelhante à covariância entre o genótipo do k-ésimo SNP raro ($\mathbf{G}_k$) e o resíduo sob H_0 ($\mathbf{Y} - \hat{\boldsymbol{\mu}}$). Logo, valores distantes de zero de S_k indicam que esse SNP pode ser relevante e estar associado ao fenótipo Y. Na prática, como se tratam de SNPs com pouca variabilidade genética, geralmente não observamos valores individuais de S_k muito distantes de zero e, por isso, a necessidade de juntar seus efeitos ponderados em um único efeito comum ao quadrado.

Sob H_0, essa estatística segue, aproximadamente, uma mistura de distribuições Qui-Quadrado, cada uma com 1 grau de liberdade. Apesar disso, no caso de variantes raras essa aproximação pode não ser adequada e, portanto, é possível calcular a distribuição da estatística teste Q_{Burden} sob H_0 de forma exata via permutação.

O conjunto de SNPs testado é considerado significativo, ou seja, ele afeta o fenótipo, quando $\beta \neq 0$. Fixado um nível de significância α, rejeitamos H_0 se o valor-p é menor que α e não rejeitamos H_0, caso contrário. Na situação de termos de conduzir esse teste para inúmeros conjuntos de SNPs, o nível de significância considerado em cada teste geralmente é corrigido por Bonferroni ou outro método parecido.

3.2 SKAT

A fim de identificar e selecionar os SNPs que influenciam o fenótipo de interesse, o SKAT (do inglês, *Sequence Kernel Association Test*) também tem sido uma das metodologias mais utilizadas quando estudamos métodos de colapsagem de variantes raras. Ele se trata de um modelo linear misto, ou seja, de uma regressão linear múltipla que apresenta tanto efeitos fixos, atribuídos às variáveis comuns (sejam elas ambientais, genéticas,

de comportamento ou ancestralidade), quanto aleatórios, os quais são atribuídos, por sua vez, aos SNPs raros (com MAF menor que 5%).

Seja $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$ um traço quantitativo a ser observado em n indivíduos. O fenótipo Y_i do i-ésimo indivíduo pode ser modelado pelo seguinte modelo linear misto:

$$Y_i = \alpha_0 + \mathbf{X}_i\boldsymbol{\alpha} + \mathbf{G}_i\boldsymbol{\beta} + \epsilon_i, \tag{3.4}$$

em que α_0 é o intercepto, $\boldsymbol{\alpha}$ é o vetor de coeficientes de dimensão $(j \times 1)$ associado às j variáveis ambientais, genéticas, de comportamento ou ancestralidade presentes no vetor $\mathbf{X}$, $\boldsymbol{\beta}^t = (\beta_1, \ldots, \beta_p)$ é o vetor de coeficientes de regressão para as p variáveis genéticas raras observadas na região, o vetor $\mathbf{G}_i = (G_{i1}, \ldots, G_{ip})$ representa o genótipo de p SNPs raros para o i-ésimo indivíduo, ϵ_i é o erro aleatório com média zero e variância σ^2, de modo que $k = 1, \ldots, p$ e $i = 1, \ldots, n$.

Conforme descrito em Wu *et al.* (2011), com o intuito de aumentar o poder do teste, ao invés de testar $H_0 : \boldsymbol{\beta} = \mathbf{0}$, assumimos que cada efeito aleatório β_k segue uma distribuição arbitrária com média 0 e variância $w_k^2\tau$, em que τ é um componente de variância, w_k é um peso pré-especificado para a k-ésima variável e $k = 1, \ldots, p$.

A significância dos SNPs é verificada via teste de hipóteses baseado no modelo estimado na Equação (3.4). Como o efeito dos SNPs raros é assumido ser aleatório, verificar se os coeficientes β_k são significativamente diferentes de zero ou não, corresponde a testar as hipóteses:

H_0: $\tau = 0$ (não há efeito dos SNPs raros no fenótipo de interesse) contra

H_1: $\tau > 0$ (há efeito dos SNPs raros no fenótipo de interesse).

De acordo com Wu *et al.* (2011), a estatística teste para verificar essas hipóteses é descrita por:

$$Q_{SKAT} = (\mathbf{Y} - \hat{\boldsymbol{\mu}})^t\mathbf{K}(\mathbf{Y} - \hat{\boldsymbol{\mu}}), \tag{3.5}$$

em que $\mathbf{K} = \mathbf{GWG}^t$, $\hat{\boldsymbol{\mu}}$ corresponde ao vetor de valores preditos do fenótipo sob H_0, $\mathbf{G}$ representa uma matriz $n \times p$, de modo que o elemento G_{ik} representa o genótipo do k-ésimo SNP do i-ésimo indivíduo codificado como $0, 1$ ou 2 para AA, Aa ou aa, respectivamente, e $\mathbf{W}$ é uma matriz diagonal que contém o peso das p variáveis genéticas raras. A matriz

$\mathbf{K}$, por sua vez, apresenta dimensão $n \times n$ e contém, na i-ésima linha e $i^{'}$-ésima coluna, o valor $K(\mathbf{G}_i, \mathbf{G}_{i'}) = \sum_{k=1}^{p} w_k G_{ik} G_{i'k}$, o qual, através dessa função de kernel K, mede a associação entre os indivíduos i e i' em relação às p variáveis genéticas. A penalização dos efeitos dos SNPs proposta por Wu *et al.* (2011) é que a raiz quadrada dos pesos w_k tivessem valores iguais aos da densidade da distribuição $Beta(1, 25)$ calculada no MAF do SNP em análise. Com essa definição, há simultaneamente um aumento do peso das variantes mais raras e uma ponderação das variantes que possuem MAF entre 1 e 5% de forma não nula.

A Equação (3.5) pode ser algebricamente simplificada como:

$$Q_{SKAT} = \sum_{k=1}^{p} w_k \left(\sum_{i=1}^{n} (Y_i - \hat{\mu}_i) G_{ik} \right)^2 . \tag{3.6}$$

De acordo com Wu *et al.* (2011), sob H_0, essa estatística segue, aproximadamente, uma mistura de distribuições Qui-Quadrado com 1 grau de liberdade, a qual pode ser aproximada através do método computacionalmente eficiente de Davies.

Conjuntos de SNPs raros que apresentam efeito significativo para o fenótipo são os que possuem valor-p menor que o nível de significância especificado ou esse nível corrigido por Bonferroni (ou outro método parecido) em caso de teste para inúmeros grupos de SNPs raros.

Segundo Training (2022), testes de componentes de variância como o SKAT são adequados e apresentam maior poder quando temos efeitos em direções opostas, ou seja, quando há tanto variantes de risco (que aumentam o efeito indesejável no fenótipo) quanto de proteção (que atenuam ou reduzem o efeito indesejável, Sapienza e Pedromônico, 2005).

Ademais, conforme descrito em Lin (2022) e Aldisi *et al.* (2023), enquanto o SKAT apresenta maior poder quando uma pequena proporção das variantes são associadas ao fenótipo, o teste de Burden apresenta maior poder quando uma grande parte das variantes são realmente significativas e seus efeitos possuem a mesma direção. Sendo assim, foi proposto o método SKAT-O (do inglês, *Optimal Unified Test*) que combina ambos os testes e determina quanto cada um deles contribui no resultado final de forma adaptativa.

3.3 SKAT-O

Essa metodologia se trata de uma combinação ótima dos testes de Burden e SKAT, de modo a ponderá-los a fim de maximizar o poder, ou seja, minimizar a probabilidade de considerar erroneamente que os SNPs são não significativos para o fenótipo de interesse.

Proposta por Lee *et al.* (2012), essa abordagem combina as estatísticas teste descritas nas Equações (3.3) e (3.5) do seguinte modo:

$$Q_\rho = (1-\rho)Q_{SKAT} + \rho Q_{Burden},$$

em que Q_{Burden} e Q_{SKAT} são, respectivamente, os escores estatísticos utilizados no teste de Burden e no SKAT, $0 \leq \rho \leq 1$ e ρ representa a associação entre os coeficientes β_k na Equação (3.1). Logo, se $\rho = 0$, o SKAT-O resulta exatamente na metodologia SKAT, uma vez que os efeitos dos SNPs raros são considerados independentes nessa metodologia, e se $\rho = 1$, estaremos utilizando o teste de Burden, visto que nessa abordagem todos os SNPs são considerados com efeito igual a β e, portanto, são completamente correlacionados.

Segundo Lee *et al.* (2014), o SKAT-O é um procedimento adaptativo que aproxima o teste ao utilizar um valor ótimo de ρ. O valor ótimo de ρ é aquele cujo teste associado apresenta o menor valor-p em uma grade de valores para ρ. O valor-p assintótico dessa metodologia também pode ser calculado via integração numérica unidimensional computacionalmente eficiente.

Ao conduzirmos esse teste para inúmeros conjuntos de SNPs, o nível de significância considerado em cada teste geralmente é corrigido por Bonferroni ou outro método parecido.

3.4 Aplicação das metodologias estudadas

As metodologias aqui apresentadas estão disponíveis na biblioteca "SKAT" do software estatístico R (Lee e Zhao, 2023). Para aplicá-las no banco de dados em análise, definimos o modelo sob H_0 através da função "SKAT_Null_Model" e, para as metodologias teste de Burden e SKAT, adequamos o valor do atributo "r.corr" da função "SKAT" da seguinte forma:

- Para o teste de Burden, utilizamos o valor 1 no atributo "r.corr"; e

- Para o SKAT, utilizamos o valor 0 no atributo “r.corr”.

Já para aplicarmos a metodologia SKAT-O, após a definição do modelo sob H_0, especificamos o valor “optimal.adj” no atributo “method” da função “SKAT”, a mesma utilizada nas demais metodologias.

3.5 Medidas de desempenho

A fim de medir e comparar o desempenho das diferentes metodologias testadas na seleção de SNPs influentes, adotamos dois indicadores denominados especificidade e sensibilidade (Lee *et al.*, 2011).

A sensibilidade é calculada como a proporção de estimativas de β diferentes de 0 dentre os verdadeiros elementos β que são não nulos. Já a especificidade é a razão da quantidade de estimativas de β iguais a 0 dentre os verdadeiros elementos β que são nulos. De maneira mais direta, o numerador da especificidade pode ser calculado como a diferença entre a quantidade de SNPs que não são verdadeiramente significativos e a quantidade de SNPs que foram erroneamente identificados como significativos e o denominador dessa fração é a quantidade de SNPs que realmente não são significativos. O cálculo desses indicadores é possível apenas em dados simulados (como os dados GAW 17 que aqui foram analisados) quando conhecemos de fato quais coeficientes são zero e quais não são.

Uma seleção de variáveis ideal ocorre quando tanto a sensibilidade, quanto a especificidade são iguais a um. Todavia, geralmente há uma compensação entre essas medidas, ou seja, quando a especificidade tende a 1, a sensibilidade tende a 0, e vice-versa.

No caso de SNPs comuns, essa verificação do desempenho ocorre SNP a SNP, enquanto que, para SNPs raros, verificamos se um conjunto de SNPs (agrupados, por exemplo, pelo gene ao qual fazem parte) foi identificado de maneira correta ou não.

Capítulo 4

Banco de dados GAW 17

O banco de dados utilizado nesse estudo é denominado *Genetic Analysis Workshop 17* (GAW 17) e é formado por dados simulados, a partir de características reais da população, de 697 indivíduos sem parentesco, sendo eles 327 homens e 370 mulheres.

Como base para a simulação de uma doença comum e complexa, de fenótipos quantitativos e dos fatores de risco (os marcadores SNPs) foram utilizados os dados reais contidos no *1000 Genomes Project,* que consideram variações genéticas de vários grupos de populações humanas, sendo eles: Europa, Leste Asiático, do sul da Ásia, África Ocidental e de índios americanos.

Um fenótipo binário que indica presença ou não de uma doença e três fenótipos quantitativos contínuos: Q_1, Q_2 e Q_3 foram simulados. Também foram simuladas informações do indivíduo como idade, sexo e condição de fumante (com uma prevalência de 25%). Para mais detalhes, ver Ióca e Zuanetti (2021) e Almasy *et al.* (2011).

Originalmente, o banco de dados nos fornecia as informações dos genótipos dos marcadores SNPs bialélicos através de 16 possíveis pares de bases nitrogenadas, sendo eles: A/A, T/T, C/C, G/G, A/T, A/C, A/G, T/A, T/C, T/G, C/A, C/T, C/G, G/A, G/T, G/C. Todavia, a fim de possibilitarmos esse estudo, analisamos quais duas das quatro bases nitrogenadas ocorriam em cada marcador. A base menos frequente foi denominada de a e a mais frequente de A em cada marcador e, com base nisso, marcamos como 0 os genótipos AA, como 1 os genótipos Aa ou aA e como 2 os genótipos aa.

Nesse estudo, analisamos o fenótipo Q_1 que originalmente é impactado por 39 SNPs em 9 genes e pelas variáveis ambientais e de comportamento Idade e Fuma, respectivamente. Nosso objetivo é verificar o desempenho das metodologias estudadas em identificar e selecionar os SNPs e genes relevantes dentre os 24487 SNPs e 3205 genes disponíveis para

análise. Ademais, ao considerarmos o limite de 5% para o MAF, para definir se um SNP é raro ou não, temos que, dentre os SNPs disponíveis, 21383 são raros e 3104 são comuns. O boxplot do fenótipo estudado Q_1 está disponível na Figura 4.1.

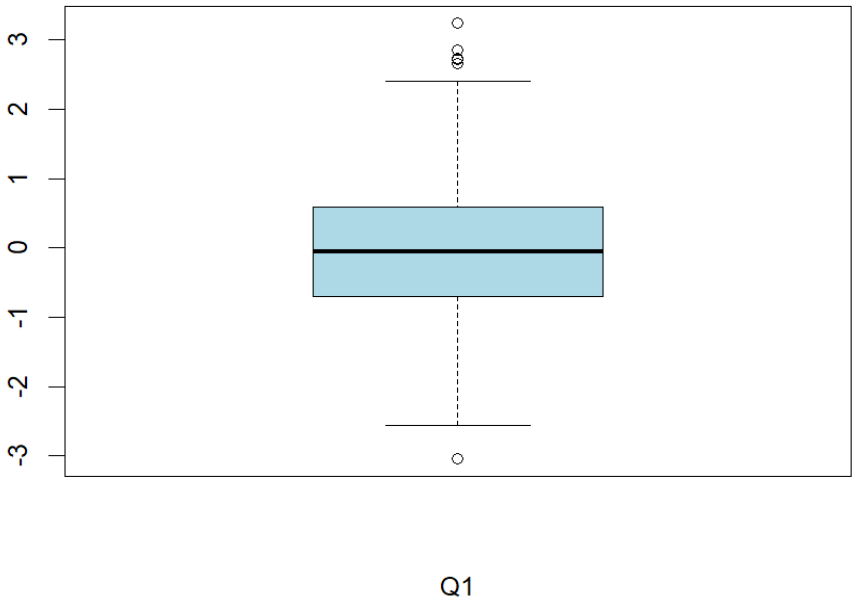

Figura 4.1: Boxplot do fenótipo Q_1.

Notamos, através da Figura 4.1 que o fenótipo Q_1 é uma variável contínua que assume valores no intervalo $(-3.0, 3.5)$ e possui 6 outliers, sendo que 5 deles ultrapassam o limite superior e 1 ultrapassa o limite inferior. Além disso, Q_1 possui comportamento simétrico e 50% dos valores observados estão entre -1 e 1.

Ademais, de acordo com a simulação realizada, os cromossomos que possuem SNPs significativos são os cromossomos 1, 4, 5, 6, 13, 14 e 19. Esses SNPs que realmente influenciam o fenótipo Q_1 são listados a seguir de acordo com o cromossomo e com o gene em que eles se encontram:

- Cromossomo 1:
    - gene ARNT: C1S6533, C1S6537, C1S6540, C1S6542 e C1S6561;
    - gene ELAVL4: C1S3181 e C1S3182;
- Cromossomo 4:
    - gene KDR: C4S1861, C4S1873, C4S1874, C4S1877, C4S1878, C4S1879, C4S1884, C4S1887, C4S1889 e C4S1890;

    - gene VEGFC: C4S4935;

- Cromossomo 5:

    - gene FLT4: C5S5133 e C5S5156;

- Cromossomo 6:

    - gene VEGFA: C6S2981;

- Cromossomo 13:

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

- Cromossomo 14:

    - gene HIF1A: C14S1718, C14S1729, C14S1734 e C14S1736; e

- Cromossomo 19:

    - gene HIF3A: C19S4799, C19S4815 e C19S4831.

Vale ressaltar que apenas 7 entre os 39 SNPs verdadeiramente relevantes possuem genótipos diferentes em mais que 1% do total de indivíduos, ou seja, grande parte dos SNPs são variantes raras e dificilmente identificados por metodologias de seleção de variáveis tradicionais.

Desse modo, analisamos os SNPs presentes em cada um dos cromossomos de acordo com seus MAFs, com o intuito de verificarmos a distribuição dessas frequências e a proporção de SNPs comuns e raros em cada um dos cromossomos do conjunto de dados GAW 17. Os boxplots das frequências do alelo menor e os gráficos de barras das frequências de SNPs raros e comuns, ambos por cromossomo, estão disponíveis, respectivamente, nas Figuras 4.2, 4.3, 4.4 e 4.5.

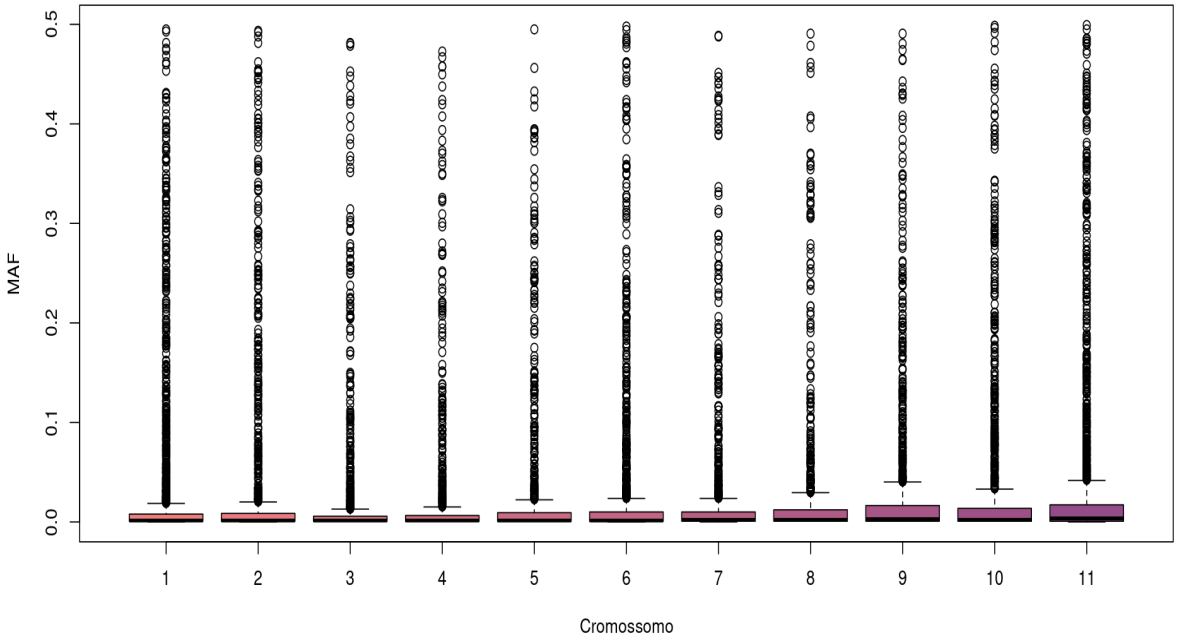

Figura 4.2: Boxplot da frequência do alelo menor (MAF) para os cromossomos de 1 a 11.

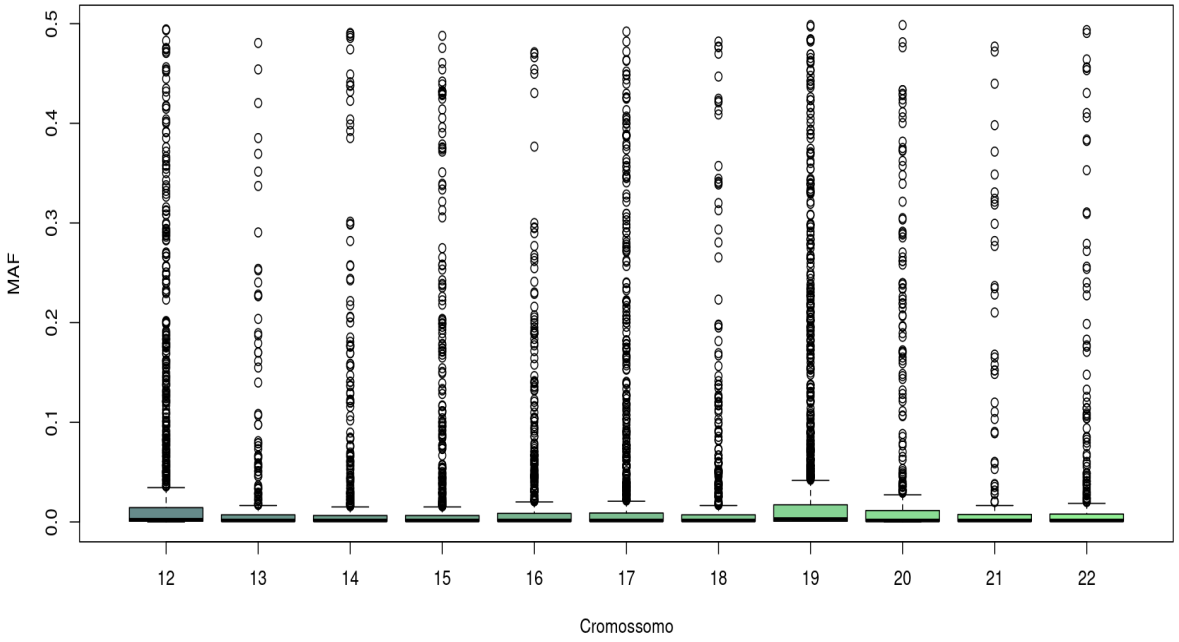

Figura 4.3: Boxplot da frequência do alelo menor (MAF) para os cromossomos de 12 a 22.

Ao analisarmos as Figuras 4.2 e 4.3, notamos que, para cada um dos 22 cromossomos autossômicos, a frequência do alelo menor assume valores extremos, porém todos abaixo

de 50%. Ademais, grande parte dos marcadores SNPs possuem MAF bem abaixo de 5% e, por esse motivo, são considerados raros.

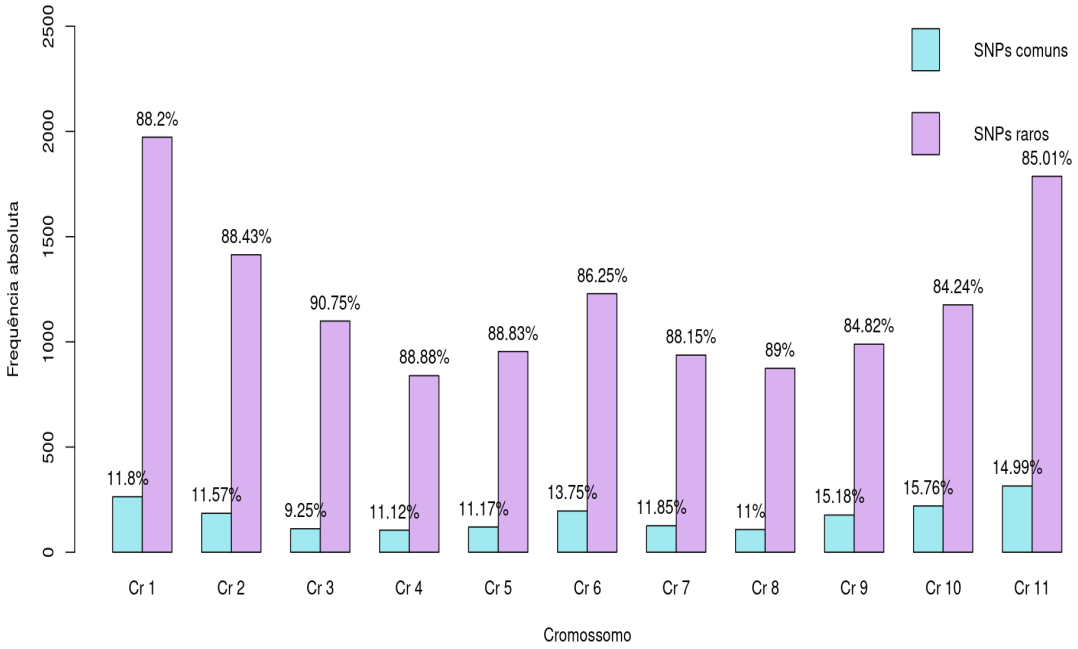

Figura 4.4: Gráfico de barras da frequência dos SNPs comuns e raros presentes nos cromossomos de 1 a 11.

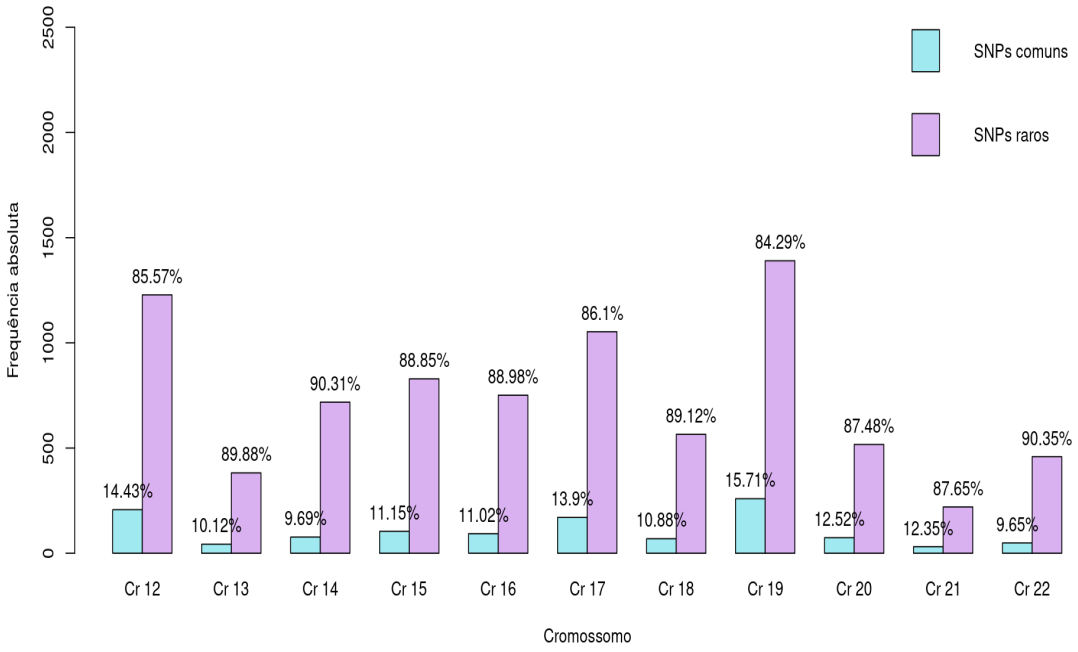

Figura 4.5: Gráfico de barras da frequência dos SNPs comuns e raros presentes nos cromossomos de 12 a 22.

De acordo com as Figuras 4.4 e 4.5, mais de 84% dos marcadores de cada um dos cromossomos é raro, ou seja, apresenta frequência do alelo menor inferior a 5%. Portanto, metodologias tradicionais de seleção de variáveis, de acordo com Midena e Zuanetti (2023), não apresentam, em geral, bons resultados. Desse modo, métodos de colapsagem de variantes raras foram utilizadas nesse conjunto de dados e seus desempenhos foram comparados.

Capítulo 5

Resultados

Nessa seção apresentamos os resultados obtidos ao aplicarmos as diferentes metodologias estudadas na análise dos dados independentes GAW 17. Como o foco de nosso estudo é a seleção de variáveis raras, conduzimos duas diferentes análises:

- Na primeira análise, consideramos um modelo basal, sob H_0, composto apenas pelo intercepto, de modo que variáveis ambientais, de comportamento ou genótipo de SNPs comuns não foram considerados;

- Em uma segunda análise, consideramos como modelo basal o modelo composto pelos SNPs comuns e pelas variáveis ambientais e de comportamento relevantes presentes no GAW 17. Os SNPs comuns, assim como as variáveis ambientais e de comportamento relevantes, foram selecionados previamente pela metodologia LASSO. Escolhemos a metodologia LASSO, pois ela mostrou o melhor desempenho na seleção em Midena e Zuanetti (2023).

5.1 Considerando apenas os SNPs raros

Os resultados obtidos pelas metodologias teste de Burden, SKAT e SKAT-O para a seleção de SNPs raros (agrupados por gene) ao considerarmos como modelo basal, sob H_0, o modelo sem covariáveis, composto apenas pelo intercepto, são apresentados nessa seção.

5.1.1 Teste de Burden

Nessa metodologia, foram realizados testes de hipóteses para verificar a significância do coeficiente associado ao genótipo dos SNPs raros em cada gene e em cada cromossomo.

Apesar do teste de Burden tradicional considerar que todos os SNPs testados dentro do mesmo gene apresentam o mesmo efeito e pesos iguais a 1, com o avanço da tecnologia e do estudo desse método, outras abordagens foram propostas. Desse modo, na versão mais recente da função descrita por Lee e Zhao (2023) que utilizamos para implementar tanto essa metodologia quanto o método SKAT no software R, o argumento que refere-se aos pesos w_ks apresenta como *default* o valor da densidade da distribuição $Beta(1, 25)$ calculada no MAF do SNP em análise.

Em nosso estudo, como fazemos um teste para os SNPs raros de cada gene, consideramos o nível de significância α de 0.05 corrigido por Bonferroni para decidir quais são significativos. A correção foi realizada da seguinte forma para cada um dos 22 cromossomos autossômicos:

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

Essa correção no nível de significância foi estendida para as demais metodologias que utilizavam como modelo basal o modelo que continha apenas o intercepto.

Os resultados obtidos nesses testes para cada cromossomo autossômico são expostos a seguir e comparados com os 9 genes que são realmente significativos para Q_1 de acordo com Almasy *et al.* (2011).

De acordo com a Tabela 5.1, a metodologia identificou que apenas os cromossomos autossômicos 2, 3, 4, 5, 6, 12, 13, 16, 17, 18, 19 e 22 possuem genes significativos, detectando, no total, 24 genes como significativos. Além disso, dos 9 genes que realmente possuem efeito significativo sobre Q_1, apenas 2 foram identificados pelo modelo, são eles: KDR e FLT1. Desse modo, 22 genes foram identificados incorretamente como significativos pela metodologia utilizada. Para o nível de significância de 5% corrigido por Bonferroni, a sensibilidade (S) e a especificidade (E) valem, respectivamente:

$$S = \frac{2}{9} = 0.222 \; e \; E = \frac{3174}{3196} = 0.993.$$

Os valores acima possuem a seguinte interpretação: 22.2% dos genes que são realmente

significativos são identificados pelo método e 99.3% dos genes que não possuem efeito importante sobre o fenótipo não são identificados pelo método.

Tabela 5.1: Genes identificados como significativos pelo teste de Burden ao considerarmos como modelo basal o modelo ajustado apenas com o intercepto.

Genes identificados	**Valor-p Burden**	**Cromossomo**
RGPD4	0.000336165968942777	2
RGPD8	1.07505650667363e-06	2
UGT1A10	1.79155535508175e-05	2
WBP1	4.26735629985795e-05	2
ATR	0.000226290568633883	3
KDR	7.83012828613719e-08	4
ZNF595	3.73835105323207e-05	4
ZNF718	0.000419049760927416	4
ZNF454	0.00033625412266969	5
HLA-B	0.000369515078138339	6
PRR4	8.26090885488344e-07	12
SLC2A13	0.000261494602610332	12
TAS2R46	9.56257954067618e-05	12
TAS2R48	0.00011400121933097	12
FLT1	1.36528003524741e-09	13
TPSD1	5.07254503833344e-05	16
KRT13	0.000115892010123621	17
PIK3C3	0.00049524938058253	18
HSZFP36	4.69382603956401e-05	19
LOC100128853	5.39280329803136e-05	19
LOC645118	0.000175891767762992	19
ZNF77	2.29723772811781e-05	19
ZNF91	8.77048824619502e-06	19
PPP1R14BP1	1.70997107905415e-05	22

5.1.2 SKAT

Os resultados obtidos através do SKAT para cada cromossomo são apresentados a seguir e comparados com os genes que realmente influenciam Q_1 segundo Almasy *et al.* (2011). A sensibilidade e especificidade atingidas por esse método também são expostas a seguir.

Tabela 5.2: Genes identificados como significativos pelo SKAT ao considerarmos como modelo basal o modelo ajustado apenas com o intercepto.

Genes identificados	**Valor-p SKAT**	**Cromossomo**
R3HDM1	0.000158032910756223	2
RGPD8	2.95571533210914e-05	2
UGT1A10	0.000296118521504574	2
KDR	9.18808487244682e-05	4
ZNF595	0.000140953059886373	4
ZNF718	0.000419049760927416	4
ZNF454	0.00033625412266969	5
RUNX2	0.000296407057297277	6
ASL	0.000236612028353456	7
OR13C2	0.000143639720951771	9
POMT1	0.000217046518384056	9
GRIA4	5.74986625294649e-05	11
PRR4	9.94423394928834e-06	12
TAS2R48	1.58598556798673e-06	12
FLT1	2.40477005792767e-15	13
KRT13	0.000262317179338889	17
PIK3C3	0.000147058160454572	18
HSZFP36	4.69382603956401e-05	19
ZNF77	0.000189019340468999	19
ZNF91	0.000221176700293313	19
PPP1R14BP1	1.80701390420923e-05	22

Ao analisarmos a Tabela 5.2, notamos que, ao nível de significância de 5% corrigido por Bonferroni, a metodologia identificou que os cromossomos autossômicos 2, 4, 5, 6, 7, 9, 11, 12, 13, 17, 18, 19 e 22 possuem genes significativos, detectando, no total, 21 genes como significativos. Além disso, dos 9 genes que realmente apresentam efeito significativo sobre Q_1, apenas 2 foram identificados pelo modelo, são eles: KDR e FLT1. Logo, 19 genes foram identificados incorretamente como significativos pela metodologia utilizada. Para o nível de significância fixado, a sensibilidade (S) e a especificidade (E) valem, respectivamente:

$$S = \frac{2}{9} = 0.222 \; e \; E = \frac{3177}{3196} = 0.994.$$

Os valores acima possuem a seguinte interpretação: 22.2% dos genes que são realmente significativos são identificados pelo método e 99.4% dos genes que não possuem efeito importante sobre o fenótipo não são identificados pelo método. Notamos que todos os genes verdadeiramente significativos identificados pelo SKAT também foram identificados pelo teste de Burden ao nível de significância de 5% corrigido por Bonferroni. Os dois genes

corretamente selecionados se tratam dos genes que mais possuem SNPs raros realmente relevantes para o fenótipo.

5.1.3 SKAT-O

Como essa abordagem é uma combinação ótima das metodologias teste de Burden e SKAT, de modo que o valor de ρ determina a associação entre os efeitos β_ks, além de apresentarmos na Tabela 5.3 quais genes foram identificados como significativos, seu respetivo valor-p obtido e a qual cromossomo ele pertence, também disponibilizamos a informação do valor de ρ escolhido como ótimo.

Tabela 5.3: Genes identificados como significativos pelo SKAT-O ao considerarmos como modelo basal o modelo ajustado apenas com o intercepto.

Genes identificados	**Valor-p SKAT-O**	ρ	**Cromossomo**
R3HDM1	0.000158032910755779	0	2
RGPD4	0.000300217722252971	0.09	2
RGPD8	1.06839764435485e-06	1	2
SPHKAP	0.000236338190098273	0.09	2
UGT1A10	1.7932187939107e-05	1	2
WBP1	4.27019421656016e-05	1	2
ATR	0.000226305694323692	1	3
KDR	6.52392959654691e-08	1	4
ZNF595	3.73934364619188e-05	1	4
HLA-B	0.000369538030987271	1	6
RUNX2	0.000296407057297499	0	6
ASL	0.000212638793239517	0.04	7
OR13C2	0.00014363972095166	0	9
POMT1	0.000166590794170252	0.09	9
GRIA4	3.17642568944398e-05	0.09	11
PRR4	8.42827311720029e-07	1	12
SLC2A13	0.000194055809679017	0.25	12
TAS2R46	9.56396024950346e-05	1	12
TAS2R48	1.43658145501657e-06	0.04	12
FLT1	8.64801945831826e-17	0.04	13
TPSD1	5.07291103644114e-05	1	16
KRT13	0.000115906975208135	1	17
PIK3C3	0.000128115865583256	0.09	18
CYP4F3	0.00014005215381907	0.09	19
LOC100128853	5.39423704712938e-05	1	19
LOC645118	0.000175916279413713	1	19
PSG5	0.00020200634859513	0.09	19
ZNF77	2.30034903874188e-05	1	19
ZNF91	8.72472336610475e-06	1	19
PPP1R14BP1	1.70843979263591e-05	1	22

Através da Tabela 5.3, notamos que ao nível de significância de 5% corrigido por Bonferroni, a metodologia identificou que os cromossomos autossômicos 2, 4, 6, 7, 9, 11, 12, 13, 16, 17, 18, 19 e 22 possuem genes significativos, detectando, ao todo, 30 genes como significativos. Além disso, dos 9 genes que realmente têm efeito significativo sobre Q_1, apenas 2 foram identificados pelo modelo, são eles: KDR e FLT1. Portanto, 28 genes foram identificados incorretamente como significativos pela metodologia utilizada. Sendo assim, ao nível de significância de 5% corrigido por Bonferroni, temos que a sensibilidade (S) e a especificidade (E) valem, respectivamente:

$$S = \frac{2}{9} = 0.222 \; e \; E = \frac{3168}{3196} = 0.991.$$

Notamos que todos os genes que realmente impactam Q_1 e que foram identificados pelo teste de Burden e pelo SKAT também foram identificados pelo SKAT-O ao nível de significância fixado. Ademais, dentre os 30 genes identificados como significativos, mais de 50% deles apresentaram resultado exatamente igual ao teste de Burden, uma vez que apresentaram valor ótimo de ρ igual a 1, enquanto que 3 deles apresentaram resultado igual ao SKAT.

5.2 Considerando modelo com variáveis ambientais e de comportamento e SNPs comuns relevantes

Nessa seção, apresentamos os resultados obtidos ao aplicarmos as metodologias teste de Burden, SKAT e SKAT-O em um modelo ajustado que considera as variáveis ambientais e de comportamento e os SNPs comuns do GAW 17 selecionados pelo LASSO como covariáveis.

Para a execução das metodologias descritas nessa abordagem, primeiro selecionamos através do LASSO, para cada um dos cromossomos, quais variáveis ambientais e de comportamento dentre as variáveis Sexo, Idade e Fuma e quais SNPs comuns eram relevantes para o fenótipo Q_1.

Nessa primeira seleção de variáveis, foram identificadas 418 variáveis comuns como relevantes pelo LASSO, de modo que 2 delas eram variáveis ambientais e de comportamento (Idade e Fuma) e realmente impactavam o fenótipo de interesse e o restante eram SNPs com MAF maior que 5%. Dentre os 3104 SNPs comuns disponíveis, 416 foram se-

lecionados, sendo que apenas 1 de fato impactava o fenótipo Q_1. Esse SNP é identificado como C13S523 e apresenta elevado valor verdadeiro de coeficiente β (0.653351). Essa baixa seleção de SNPs comuns corretamente identificados ocorre devido ao fato de, dentre os 39 marcadores que realmente impactam o fenótipo, apenas 2 serem comuns (são eles C13S523 e C14S1878) e o MAF desses marcadores e seus coeficientes associados valem, respectivamente, 0.066714 e 0.653351 para o marcador C13S523 e 0.164993 e 0.149975 para o marcador C14S1878. Desse modo, temos que a sensibilidade (S) e a especificidade (E) do LASSO na seleção de SNPs comuns valem, respectivamente:

$$S = \frac{1}{2} = 0.50 \; e \; E = \frac{2787}{3102} = 0.870.$$

Os valores acima possuem a seguinte interpretação: 50% dos marcadores comuns que são realmente significativos são identificados pelo LASSO e 87% dos marcadores comuns que não possuem efeito importante sobre o fenótipo não são identificados pelo modelo.

Selecionadas então as variáveis comuns através do LASSO, ajustamos um modelo basal considerando essas variáveis relevantes como covariáveis do modelo e aplicamos cada uma das metodologias teste de Burden, SKAT e SKAT-O para selecionar, dentre os genes que contém SNPs raros, quais realmente impactam o fenótipo.

5.2.1 Teste de Burden

Nessa metodologia, foram realizados testes de hipóteses para verificar a significância do coeficiente associado aos genótipos dos marcadores raros (agrupados por gene).

Assim como na abordagem que utilizava como modelo basal o modelo apenas com intercepto, nessa segunda abordagem também utilizamos a versão mais recente da função descrita por Lee e Zhao (2023), de modo que o argumento que refere-se aos pesos w_ks das metodologias teste de Burden e SKAT apresenta como *default* o valor da densidade da distribuição $Beta(1, 25)$ calculada no MAF do SNP em análise. Ademais, também consideramos o nível de significância corrigido por Bonferroni descrito na Equação (5.1) para todas as metodologias utilizadas.

Os resultados obtidos nesses testes para cada cromossomo autossômico são expostos a seguir e comparados com os 9 genes que são realmente significativos para Q_1 de acordo com Almasy *et al.* (2011).

Tabela 5.4: Genes identificados como significativos pelo teste de Burden ao considerarmos como modelo basal o modelo ajustado com covariáveis selecionadas pelo LASSO.

Genes identificados	**Valor-p Burden com covariáveis**	**Cromossomo**
UGT1A10	1.28887710670831e-06	2
UGT1A7	0.0002652940327598	2
UGT1A8	0.000155598056136802	2
KDR	2.12515638110853e-08	4
VEGFC	0.000406001679898976	4
ZNF595	7.76941856516854e-05	4
TRPV4	3.92682333732357e-05	12
FLT1	1.27177355035177e-07	13
KRT13	2.48664215434705e-05	17
PIK3C3	0.000836594076792732	18
C20orf26	0.000554536581715962	20
APOBEC3B	0.000610754925170087	22
PPP1R14BP1	0.000698061123652481	22

De acordo com a Tabela 5.4, o método identificou que apenas os cromossomos autossômicos 2, 4, 12, 13, 17, 18, 20 e 22 possuem genes significativos, detectando, no total, 13 genes como significativos. Além disso, dos 9 genes que realmente têm efeito significativo sobre Q_1, apenas 3 foram identificados pelo modelo, são eles: KDR, VEGFC e FLT1. Desse modo, 10 genes foram identificados incorretamente como significativos pela metodologia utilizada. Para o nível de significância de 5% corrigido por Bonferroni, a sensibilidade (S) e a especificidade (E) valem, respectivamente:

$$S = \frac{3}{9} = 0.333 \; e \; E = \frac{3186}{3196} = 0.997.$$

Os valores acima possuem a seguinte interpretação: 33.3% dos genes que são realmente significativos são identificados pelo método e 99.7% dos genes que não possuem efeito importante sobre o fenótipo não são identificados.

5.2.2 SKAT

Os resultados obtidos através do SKAT para cada cromossomo são apresentados a seguir e comparados com os genes que realmente influenciam Q_1 segundo Almasy *et al.* (2011). A sensibilidade e especificidade atingidas por esse método também são expostas a seguir.

Tabela 5.5: Genes identificados como significativos pelo SKAT ao considerarmos como modelo basal o modelo ajustado com covariáveis selecionadas pelo LASSO.

Genes identificados	**Valor-p SKAT com covariáveis**	**Cromossomo**
UGT1A10	3.06326815561198e-05	2
KDR	3.67190752598168e-05	4
VEGFC	0.000406001679898976	4
ZNF595	0.000354402365636219	4
LOC100129248	0.000136645395670154	9
AKAP3	0.000107055846092163	12
FLT1	6.19533542578488e-07	13
KRT13	7.90730047615096e-05	17
PIK3C3	0.000566659188032115	18
APOBEC3B	0.000540370342978491	22
PPP1R14BP1	0.000744219164377058	22

Ao analisarmos a Tabela 5.5, notamos que, ao nível de significância de 5% corrigido por Bonferroni, a metodologia identificou que os cromossomos autossômicos 2, 4, 9, 12, 13, 17, 18 e 22 possuem genes significativos, detectando, no total, 11 genes como significativos. Além disso, dos 9 genes que realmente possuem efeito significativo sobre Q_1, apenas 3 foram identificados pelo modelo, são eles: KDR, VEGFC e FLT1. Sendo assim, 8 genes foram identificados incorretamente como significativos pela metodologia utilizada. Para o nível de significância fixado, a sensibilidade (S) e a especificidade (E) valem, respectivamente:

$$S = \frac{3}{9} = 0.333 \; e \; E = \frac{3188}{3196} = 0.997.$$

Notamos que todos os genes verdadeiros identificados pelo SKAT também foram identificados pelo teste de Burden ao nível de significância de 5% corrigido por Bonferroni. Ademais, dois dos genes corretamente selecionados se tratam dos genes que mais possuem SNPs raros realmente relevantes para o fenótipo.

5.2.3 SKAT-O

Como essa abordagem é uma combinação ótima das metodologias teste de Burden e SKAT, de modo que o valor de ρ determina a associação entre os efeitos β_ks, assim como fizemos na Seção 5.1.3, além de apresentarmos na Tabela 5.6 quais genes foram identificados como significativos, seu respetivo valor-p obtido e a qual cromossomo ele pertence, também disponibilizamos a informação do valor de ρ.

Tabela 5.6: Genes identificados como significativos pelo SKAT-O ao considerarmos como modelo basal o modelo ajustado com covariáveis selecionadas pelo LASSO.

Genes identificados	**Valor-p SKAT-O com covariáveis**	ρ	**Cromossomo**
UGT1A10	1.24364748421701e-06	1	2
UGT1A7	0.000260113587994937	0.5	2
UGT1A8	0.000155625029258255	1	2
KDR	8.61624382952897e-09	0.5	4
ZNF595	7.77150145556682e-05	1	4
LOC100129248	0.000136645395670154	0	9
AKAP3	0.000107055846092274	0	12
TRPV4	3.92770136100618e-05	1	12
FLT1	1.44443709860517e-09	0.09	13
KRT13	2.48725023590524e-05	1	17
PIK3C3	0.000412980752194869	0.25	18
C20orf26	0.000532538292221107	0.25	20
APOBEC3B	0.000407751473574858	0.25	22
PPP1R14BP1	0.000698088823340215	1	22

Através da Tabela 5.6, notamos que ao nível de significância de 5% corrigido por Bonferroni, o modelo identificou que os cromossomos autossômicos 2, 4, 9, 12, 13, 17, 18, 20 e 22 possuem genes significativos, detectando, ao todo, 14 genes como significativos. Além disso, dos 9 genes que realmente afetam significativamente o fenótipo Q_1, apenas 2 foram identificados pelo modelo, são eles: KDR e FLT1. Desse modo, 12 genes foram identificados incorretamente como significativos pela metodologia utilizada. Logo, ao nível de significância de 5% corrigido por Bonferroni, temos que a sensibilidade (S) e a especificidade (E) valem, respectivamente:

$$S = \frac{2}{9} = 0.222 \; e \; E = \frac{3184}{3196} = 0.996.$$

Notamos que, exceto pelo gene VEFGC, todos os genes verdadeiros identificados pelo teste de Burden e pelo SKAT também foram identificados pelo SKAT-O ao nível de significância fixado. Essa não identificação do gene VEFGC como significativo pelo SKAT-O decorre do fato dessa metodologia não permitir a análise de um gene quando essa região possui um único SNP raro. As demais metodologias não apresentam esse problema de execução da análise de algum agrupamento, de modo que conseguem selecionar esse gene como significativo. Além disso, dentre os 14 genes identificados como significativos pelo SKAT-O, cerca de 42% deles apresentaram resultado exatamente igual ao teste de Burden, uma vez que apresentaram valor ótimo de ρ igual a 1, enquanto que 2 deles apresentaram

resultado igual ao SKAT.

5.2.4 Comparações dos resultados obtidos

Com o propósito de comparar o desempenho das metodologias para selecionar os genes que influenciam o fenótipo Q_1, ressaltamos o número de genes identificados como significativos e os valores de sensibilidade e especificidade advindos de cada um dos métodos estudados anteriormente. Os valores são mostrados na Tabela 5.7.

Tabela 5.7: Valores de sensibilidade, especificidade e número de genes identificados como significativos para cada uma das abordagens de modelos e em cada um dos métodos estudados ao nível de significância de 5% corrigido por Bonferroni.

	Modelo basal apenas com intercepto			**Modelo basal com covariáveis selecionadas pelo LASSO**		
	Teste de Burden	**SKAT**	**SKAT-O**	**Teste de Burden**	**SKAT**	**SKAT-O**
Sensibilidade	0.222	0.222	0.222	0.333	0.333	0.222
Especificidade	0.993	0.994	0.991	0.997	0.997	0.996
N° de genes identificados como significativos	24	21	30	13	11	14

Através da Tabela 5.7, notamos que o teste de Burden e o SKAT, ambos utilizando o modelo com covariáveis selecionadas pelo LASSO como modelo basal e ao nível de significância de 5% corrigido por Bonferroni, possuem a melhor especificidade dentre os métodos comparados, pois seu valor é o mais próximo de 1. Além disso, independente do modelo basal, esses mesmos dois métodos apresentaram os melhores desempenhos em relação à sensibilidade, ainda que elas sejam relativamente baixas, indicando que esses modelos não consideram como significativos grande parte dos genes "raros" (ou seja, que contém SNPs raros) que realmente influenciam o fenótipo.

Notamos ainda que ao considerarmos como modelo basal o modelo apenas com intercepto, todas as metodologias estudadas apresentam o mesmo valor de sensibilidade, além de possuírem valores mais baixos de especificidade e identificarem mais genes como significativos quando comparamos com o desempenho de todas as metodologias estudadas ao utilizar como modelo basal o modelo com as covariáveis selecionadas pelo LASSO. Desse modo, temos que o modelo basal com covariáveis identifica menor quantidade de falsos positivos e negativos independentemente da metodologia utilizada, indicando um melhor desempenho. Ademais, dentre as três metodologias utilizando como modelo basal o modelo com as covariáveis selecionadas pelo LASSO, o SKAT-O apresentou os piores

resultados, de modo que tanto a sensibilidade quanto especificidade e número de genes identificados como significativos ainda apresentam performance igual ou melhor que o modelo basal ajustado apenas com o intercepto para cada uma das metodologias estudadas. Essa maior identificação de SNPs raros como significativos para o SKAT-O já era esperada devido ao fato desse método escolher o valor de ρ de modo que o teste associado apresente o menor valor-p em uma grade de valores para ρ.

Ao analisarmos os genes que são realmente significativos para o fenótipo Q_1, percebemos que os genes KDR e FLT1 foram identificados por todos os métodos utilizados, independentemente do modelo basal utilizado, e o gene VEGFC foi identificado tanto pelo teste de Burden quanto pelo SKAT ao utilizar o modelo com covariáveis selecionadas pelo LASSO como modelo basal e ao nível de significância de 5% corrigido. A identificação desse gene por mais de um método pode ocorrer devido ao elevado valor do efeito dessa região, o qual vale 1.40529.

Comparado com todas as metodologias e abordagens realizadas, os métodos que apresentam simultaneamente bons valores de sensibilidade e especificidade, além de terem identificado (mesmo sem aumentar consideravelmente o número de falso positivos) genes influentes foram o teste de Burden e o SKAT, ambos utilizando como modelo basal o modelo com covariáveis selecionadas pelo LASSO. Ademais, ainda que o SKAT-O seja uma otimização das metodologias teste de Burden e SKAT e que tenha apresentado, para o modelo ajustado apenas com o intercepto, elevado valor de especificidade, ao compararmos seu desempenho com os demais métodos e abordagens utilizados, ele foi o que apresentou o pior desempenho no geral, uma vez que obteve, simultaneamente, os menores valores de sensibilidade e especificidade e a maior quantidade de genes falso positivos. Um comportamento análogo a esse é visto quando analisamos o desempenho dessa mesma metodologia no modelo ajustado com as covariáveis selecionadas pelo LASSO.

Nessa análise, os métodos Burden e SKAT apresentaram um desempenho muito parecido, sendo o SKAT levemente melhor na especificidade. Isso se deve, talvez, ao fato de que todos os SNPs realmente significativos apresentam efeito positivo (na mesma direção apesar de serem diferentes) no fenótipo nos dados GAW 17 e isso favorece uma boa performance do teste de Burden.

Ademais, ao fixarmos uma metodologia e compararmos os resultados obtidos quando o modelo utilizado continha apenas o intercepto e quando o modelo ajustado considerava as covariáveis selecionadas como relevantes pelo LASSO, notamos que o valor-p apresentou

uma mudança expressiva, de modo que no modelo mais complexo, esse valor era maior, ou seja, os SNPs raros apresentavam menor significância, uma vez que as covariáveis (SNPs comuns e variáveis ambientais e de comportamento selecionadas) agregam mais informações para o modelo ajustado que apenas o intercepto.

Capítulo 6

Conclusão e próximos passos

Nesse trabalho, estudamos, aplicamos e avaliamos o desempenho de três diferentes metodologias em duas diferentes abordagens para identificação de SNPs raros relevantes no fenótipo Q_1 dos dados independentes GAW 17. Para isso, agrupamos a informação desses marcadores raros através do gene em que se localizavam.

Ao analisarmos o desempenho da metodologia SKAT-O tanto utilizando como modelo basal, sob H_0, o modelo ajustado apenas com o intercepto quanto utilizando o ajustado com as covariáveis selecionadas pelo LASSO, notamos que esse método foi o que apresentou os menores valores de sensibilidade e especificidade e a maior quantidade de genes falso positivos.

O modelo ajustado com as covariáveis selecionadas pelo LASSO, apesar de ser mais complexo por considerar tanto variáveis ambientais quanto de comportamento e SNPs comuns, apresentou melhor equilíbrio entre sensibilidade, especificidade e taxa de falso positivos para todas as metodologias estudadas, quando comparamos com o modelo ajustado apenas com o intercepto. Em relação à seleção de variáveis ambientais e de comportamento através do LASSO, essa metodologia identificou corretamente as variáveis Idade e Fuma como significativas para o fenótipo Q_1, além de ter apresentado um equilíbrio entre a sensibilidade e especificidade e ter identificado uma quantidade relativamente baixa de SNPs comuns como significativos.

Dentre todas as três metodologias estudadas em cada uma das abordagens utilizadas (modelo ajustado apenas com o intercepto e modelo ajustado com as covariáveis selecionadas pelo LASSO), notamos que o teste de Burden e o SKAT, ambos considerando como modelo basal o modelo com covariáveis selecionadas previamente pelo LASSO e ao nível de significância de 5% corrigido, além de apresentarem um equilíbrio entre a sensibilidade

e a especificidade e possuírem baixa taxa de genes falso positivos, também identificaram o gene VEGFC, o qual afeta Q_1 e possui apenas um SNP muito raro significativo (com $MAF < 1\%$). Como esse gene apresenta um único SNP, a metodologia SKAT-O, por sua vez, não consegue analisá-lo e identificá-lo como significativo.

Nessa análise dos dados GAW 17, observamos um desempenho muito semelhante entre o teste de Burden e o SKAT, talvez porque o efeito verdadeiro de todos os SNPs significativos é positivo (na mesma direção), característica que beneficia o teste de Burden. Em outros conjuntos de dados que apresentem efeitos com direções opostas, o desempenho das duas metodologias pode não ser tão parecido.

Por fim, observamos que, apesar de todas as metodologias testadas apresentarem elevados valores de especificidade, ou seja, não identificarem como significativos genes que realmente não impactam o fenótipo, nenhuma delas apresenta ótima performance em identificar genes relevantes no fenótipo, uma vez que o maior valor de sensibilidade observado dentre todos os resultados obtidos ainda foi relativamente baixo (0.333). Desse modo, para alcançarmos essa ótima performance de identificação, é necessário o estudo e a proposição de novas metodologias, principalmente para seleção de SNPs raros.

Uma abordagem diferente proposta por Jia (2015) para dados familiares que poderia ser estendida para dados independentes e trazer melhores resultados para o conjunto de dados estudado nesse trabalho seria agrupar os SNPs raros por gene e, dentro de cada gene, aplicar componentes principais nos genótipos dos seus SNPs. Desse modo, as informações dos genótipos dos SNPs raros seriam colapsadas em poucos componentes principais dentro de cada gene e testaríamos a significância desses componentes, com efeitos fixos, por exemplo, e não aleatórios. Uma segunda abordagem seria realizar análise de componentes principais para os SNPs comuns e ajustar os testes para os componentes principais mais importantes (que, em geral, são utilizados para representar a estrutura populacional/de ancestralidade). Sendo assim, futuramente estudaremos essas formas de seleção de variáveis e aplicaremos no conjunto de dados independentes GAW 17, utilizado nesse trabalho, de modo a verificar se a identificação correta de SNPs raros como significativos e não significativos apresenta melhores resultados.

Referências Bibliográficas

Aldisi, R., Hassanin, E., Sivalingam, S., Buness, A., Klinkhammer, H., Mayr, A., Fröhlich, H., Krawitz, P. e Maj, C. (2023). Gene-based burden scores identify rare variant associations for 28 blood biomarkers. *BMC Genomic Data*, **24**, 50.

Almasy, L., Dyer, T. D., Peralta, J. M., Kent, J. W., Charlesworth, J. C., Curran, J. E. e Blangero, J. (2011). Genetic Analysis Workshop 17 mini-exome simulation. Em *BMC Proceedings*, volume 5, página S2. BioMed Central.

Caetano, A. R. (2009). Marcadores SNP: conceitos básicos, aplicações no manejo e no melhoramento animal e perspectivas para o futuro. *Revista Brasileira de Zootecnia*, **38**(SPE), 64–71.

Chun, H. e Keleş, S. (2010). Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **72**(1), 3–25.

Ióca, M. P. e Zuanetti, D. A. (2021). Selection of SNP markers: Analyzing GAW17 data using different methodologies. *Brazilian Journal of Biometrics*, **39**(1), 71–88.

Ionita-Laza, I., Lee, S., Makarov, V., Buxbaum, J. D. e Lin, X. (2013). Sequence kernel association tests for the combined effect of rare and common variants. *The American Journal of Human Genetics*, **92**(6), 841–853.

Jia, J. (2015). *Association analysis between binary traits and common or rare genetic variants on family-based data*. Tese de doutorado, University of Pittsburgh.

Laird, N. M. e Lange, C. (2011). *The fundamentals of modern statistical genetics*. Springer.

Lee, D., Lee, W., Lee, Y. e Pawitan, Y. (2011). Sparse partial least-squares regression and its applications to high-throughput data analysis. *Chemometrics and Intelligent Laboratory Systems*, **109**(1), 1–8.

Lee, S. e Zhao, Z. (2023). SKAT Package. Disponível em: https://cran.r-project.org/web/packages/SKAT/vignettes/SKAT.pdf. Acessado: 2023-08-06.

Lee, S., Wu, M. C. e Lin, X. (2012). Optimal tests for rare variant effects in sequencing association studies. *Biostatistics*, **13**(4), 762–775.

Lee, S., Abecasis, G. R., Boehnke, M. e Lin, X. (2014). Rare-variant association analysis: study designs and statistical tests. *The American Journal of Human Genetics*, **95**(1), 5–23.

Li, B. e Leal, S. M. (2008). Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *The American Journal of Human Genetics*, **83**(3), 311–321.

Lin, X. (2022). Burden, SKAT and Optimal Unified Tests (SKAT-O) for WES Association Studies. Disponível em: https://esp.gs.washington.edu/drupal/system/files/7B_Lin_SKATO.pdf. Acessado: 2023-05-26.

Midena, L. e Zuanetti, D. (2023). Métodos para mapeamento de QTL através de marcadores tipo SNP: uma comparação. Em A. D. P. Silva, editor, *Trajetórias e perspectivas para a pesquisa em matemática 2*, páginas 125–140. Atena Editora.

Morgenthaler, S. e Thilly, W. G. (2007). A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, **615**(1-2), 28–56.

Sapienza, G. e Pedromônico, M. R. M. (2005). Risco, proteção e resiliência no desenvolvimento da criança e do adolescente. *Psicologia em Estudo*, **10**, 209–216.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, **58**(1), 267–288.

Training, E.-E. (2022). Methods for rare-variant association analysis. Disponível em: https://www.ebi.ac.uk/training/events/

methods-rare-variant-association-analysis/#vf-tabs__section--tab1. Acessado: 2023-05-25.

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M. e Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics*, **89**(1), 82–93.

Zuanetti, D. A. (2023). Notas de aula da disciplina Tópicos em Estatística Genética do Bacharelado em Estatística da UFSCar.

Apêndice A

Códigos

```
library(dplyr)
library(psych)
library(readr)
library(ggplot2)
library(grDevices)
library(SKAT)
library(glmnet)

######## LEITURA DA BASE DE DADOS DO CROMOSSOMO 1 E DO FENÓTIPO
unr_phen <- read_csv("Lara/Dados_independentes/unr_phen.1")
snp_info <- read_csv("Lara/Dados_independentes/snp_info")
c1_snps <- read_csv("Lara/Dados_independentes/c1_snps.unr")
fuma<- unr_phen$SMOKE
idade<- unr_phen$AGE
sexo<- unr_phen$SEX-1

boxplot(unr_phen$Q1, col = "lightblue", xlab = "Q1")
boxplot.stats(unr_phen$Q1)$out

############## RETIRANDO SNPS SEM VARIAÇÃO NO CROMOSSOMO 1
mono_1<-numeric(ncol(c1_snps)-1)
for (i in 1:(ncol(c1_snps)-1)) mono_1[i]<-length(table(c1_snps[,i+1]))
checa<-table(mono_1)
```

```
j<-1
excl<-numeric(checa[1])
for (i in 1:(ncol(c1_snps)-1)){
  if (mono_1[i]==1) {
    excl[j]<-i
    j<-j+1}}
excl<-excl+1
c1_snps<-c1_snps[,-excl]
ncol(c1_snps)


######## CLASSIFICACAO DAS BASE DE DADOS EM 0, 1 E 2
dados_1<-matrix(0,nrow=nrow(c1_snps),ncol = ncol(c1_snps))
for (i in 1:nrow(dados_1)){
  for (j in 1: ncol(dados_1)){
    if (c1_snps[i,j]=="A/A"|c1_snps[i,j]=="T/T") dados_1[i,j]=2
    if (c1_snps[i,j]=="G/G"|c1_snps[i,j]=="C/C") dados_1[i,j]=0
    if (c1_snps[i,j]!="A/A" & c1_snps[i,j]!="T/T" &
        c1_snps[i,j]!="G/G" & c1_snps[i,j]!="C/C") dados_1[i,j]=1
  }
}
View(dados_1)
ncol(dados_1)
table(dados_1[,7])



######## VENDO A FREQ ALÉLICA DE TODOS OS SNPS DO CR1
freq_alelica_SNPs_cr1 <-matrix(0,nrow=2,ncol = ncol(c1_snps))
row.names(freq_alelica_SNPs_cr1)<- c('freq alelo A','freq alelo a')
colnames(freq_alelica_SNPs_cr1)<- colnames(c1_snps)
View(freq_alelica_SNPs_cr1)
for (i in 1:ncol(dados_1)){
```

```
  freq_alelica_SNPs_cr1[1,i]<-
(0*length(dados_1[which(dados_1[,i]=='2'),i])/(2*nrow(dados_1))) +
    (1*length(dados_1[which(dados_1[,i]=='1'),i])/(2*nrow(dados_1))) +
    (2*length(dados_1[which(dados_1[,i]=='0'),i])/(2*nrow(dados_1)))

  freq_alelica_SNPs_cr1[2,i]<-
(2*length(dados_1[which(dados_1[,i]=='2'),i])/(2*nrow(dados_1))) +
    (1*length(dados_1[which(dados_1[,i]=='1'),i])/(2*nrow(dados_1))) +
    (0*length(dados_1[which(dados_1[,i]=='0'),i])/(2*nrow(dados_1)))

}
View(freq_alelica_SNPs_cr1)

######## SEPARANDO OS SNPS EM RAROS E COMUNS E IDENTIFICANDO-OS
indice_SNPs_raros_cr_1<-which(freq_alelica_SNPs_cr1[,1:ncol(dados_1)]<0.05)
for (i in 1:length(indice_SNPs_raros_cr_1)){
  if (indice_SNPs_raros_cr_1[i]%%2==0) indice_SNPs_raros_cr_1[i]
  = indice_SNPs_raros_cr_1[i]/2
  else indice_SNPs_raros_cr_1[i] = (indice_SNPs_raros_cr_1[i]+1)/2
}
indice_SNPs_raros_cr_1

SNPs_raros_cr1<- dados_1[,indice_SNPs_raros_cr_1]
View(SNPs_raros_cr1)
colnames(SNPs_raros_cr1)<- nomes_1[indice_SNPs_raros_cr_1]
SNPs_comuns_cr1<- dados_1[,-c(indice_SNPs_raros_cr_1)]
View(SNPs_comuns_cr1)


######## CALCULANDO O MAF DE CADA SNP DO CR 1 PRA FAZER BOXPLOT
DO MAF POR CROMOSSOMO
maf_1<- NULL
for (i in 1:ncol(freq_alelica_SNPs_cr1)){
```

```
  maf_1[i]<- min(freq_alelica_SNPs_cr1[1,i],freq_alelica_SNPs_cr1[2,i])
}
maf_1
boxplot(maf_1)
summary(maf_1)


######## FENÓTIPO
ytr_1 = NULL #cria uma lista vazia

for (i in 1:nrow(unr_phen)){
  ytr_1[i] = unr_phen$Q1[i]
}
ytr_1


######## SEPARANDO OS SNPS RAROS DE ACORDO COM O GENE
gene_1<- matrix(0,2,ncol(dados_1))
snp_info$gene[which(snp_info$chromosome==1)]
tab_nomes_1<-as.matrix(colnames(SNPs_raros_cr1))
colnames(tab_nomes_1)<-c("snp_name")
tab_nomes_1<-as.data.frame(tab_nomes_1)
SNPs_raros_cr1<-as.data.frame(SNPs_raros_cr1)
tabela_full_cr1<- tab_nomes_1 %>% left_join(snp_info,by='snp_name')
tabela_gene_cr1<- tabela_full_cr1[,-c(2,3,5,6,7)]


######## AGRUPANDO OS SNPS Q ESTÃO NO MESMO GENE
table(tabela_gene_cr1[,2]) #gene e quantidade de snps em cada um deles
genes_cr1<-as.character(data.frame(table(tabela_gene_cr1[,2]))[,1])
genes_cr1 #ta em ordem alfabetica
SNPs_raros_cr1
```

```
######### PRIMEIRA ABORDAGEM: APLICA AS METODOLOGIAS NO MODELO
AJUSTADO APENAS COM O INTERCEPTO


######## TESTANDO SIGNIFICANCIA DOS GENES DE SNPS RAROS PARA CADA
METODOLOGIA EM UM MODELO SÓ COM O INTERCEPTO
valorp_Burden<-NULL
valorp_SKAT<-NULL
valorp_SKAT_O<-NULL
SKAT_O<-NULL
rho_SKAT_O<-NULL
#valorp_gene<- matrix(0,nrow=length(genes),ncol=2)
for (i in 1:length(genes_cr1)){
  snps<-which(tabela_gene_cr1[,2]==genes_cr1[i]) # posição na lista genes
  (tá em ordem alfabética) dos snps que estão no gene pedido da tabela_gene_cr1
  obj<-SKAT_Null_Model(ytr_1 ~ 1, out_type="C")
  valorp_Burden[i]<-SKAT(as.matrix(SNPs_raros_cr1[,snps]), obj, r.corr=1)$p.value
  #ordem dos genes em ordem alfabética
  valorp_SKAT[i]<-SKAT(as.matrix(SNPs_raros_cr1[,snps]), obj, r.corr=0)$p.value
  SKAT_O<-SKAT(as.matrix(SNPs_raros_cr1[,snps]), obj, method="optimal.adj")
  if (is.null(SKAT_O$param$minp)==FALSE){
    valorp_SKAT_O[i]<- SKAT_O$param$minp
    rho_SKAT_O[i]<-SKAT_O$param$rho_est
  }
}
sign_gene_cr1<-cbind(genes_cr1, valorp_Burden, valorp_SKAT, valorp_SKAT_O,
rho_SKAT_O)
colnames(sign_gene_cr1)<- c("Genes", "Valor-p Burden", "Valor-p SKAT",
"Valor-p SKAT-O", "Rho SKAT-O")
```

```
######### SEGUNDA ABORDAGEM: SELECIONA AS COVARIÁVEIS PELO LASSO E APLICA
AS METODOLOGIAS NO MODELO AJUSTADO COM AS COVARIÁVEIS SELECIONADAS PELO LASSO



######## SELECIONANDO SNPS COMUNS PELO LASSO
set.seed(202310)
x_cr1<- as.matrix(SNPs_comuns_cr1)
X_cr1<- cbind(x_cr1,fuma,idade,sexo)
cv.out_cr1 <- cv.glmnet(as.matrix(X_cr1), ytr_1, alpha = 1,
family="gaussian")
plot(cv.out_cr1, ylab = "EQP", xlab = "log(lambda)")
bestlam_1_cr1 <- cv.out_cr1$lambda.min

### ARMAZENA O MELHOR LÂMBIDA E O GRAU DE LIBERDADE
lasso.mod_cr1 <- glmnet(as.matrix(X_cr1), ytr_1, alpha = 1, lambda = bestlam_1_cr1,
family="gaussian")
lasso.coef_cr1  <- predict(lasso.mod_cr1, type = 'coefficients', s = bestlam_1_cr1)
coef_lasso_cromo_1 = as.matrix(lasso.coef_cr1)

######## REMOVE O VALOR DO INTERCEPTO NO LASSO E DETERMINA QUAIS VARIÁVEIS SÃO
SIGNIFICATIVAS
betas_cr1<- matrix(0,nrow=ncol(X_cr1),ncol=2)
betas_cromo_1 = coef_lasso_cromo_1[c(-1)]
betas_cr1[,1]<-colnames(X_cr1)
betas_cr1[,2]<- betas_cromo_1
snps_comuns_sign_cr1 = betas_cr1[which(betas_cr1[,2]!=0),]
snps_comuns_sign_cr1<- as.data.frame(snps_comuns_sign_cr1)
colnames(snps_comuns_sign_cr1)<- c('Variável comum','Coeficiente beta')
indice_var_comum_sign_cr1 = which(betas_cr1[,2]!=0)
```

```
var_comum_sign_cr1<- X_cr1[,indice_var_comum_sign_cr1]



######## TESTANDO SIGNIFICANCIA DOS GENES DE SNPS RAROS USANDO COMO COVARIÁVEIS
OS SNPS COMUNS E VARIÁVEIS SELECIONADAS PELO LASSO
valorp_Burden_cov<-NULL
valorp_SKAT_cov<-NULL
valorp_SKAT_O_cov<-NULL
SKAT_O_cov<-NULL
rho_SKAT_O_cov<-NULL
#valorp_gene<- matrix(0,nrow=length(genes),ncol=2)
for (i in 1:length(genes_cr1)){
  snps<-which(tabela_gene_cr1[,2]==genes_cr1[i]) #posição na lista genes
  (está em ordem alfabética) dos snps que estão no gene pedido da tabela_gene_cr1
  obj<-SKAT_Null_Model(ytr_1 ~ var_comum_sign_cr1, out_type="C")
  valorp_Burden_cov[i]<-SKAT(as.matrix(SNPs_raros_cr1[,snps]),
  obj, r.corr=1)$p.value #ordem dos genes em ordem alfabética
  valorp_SKAT_cov[i]<-SKAT(as.matrix(SNPs_raros_cr1[,snps]), obj,
  r.corr=0)$p.value
  SKAT_O_cov<-SKAT(as.matrix(SNPs_raros_cr1[,snps]), obj,   method="optimal.adj")
  if (is.null(SKAT_O_cov$param$minp)==FALSE){
    valorp_SKAT_O_cov[i]<- SKAT_O_cov$param$minp
    rho_SKAT_O_cov[i]<-SKAT_O_cov$param$rho_est
  }
}
sign_gene_cr1_cov<-cbind(genes_cr1, valorp_Burden_cov, valorp_SKAT_cov,
valorp_SKAT_O_cov, rho_SKAT_O_cov)
colnames(sign_gene_cr1_cov)<- c("Genes", "Valor-p Burden com Cov",
"Valor-p SKAT com Cov", "Valor-p SKAT-O com Cov", "Rho SKAT-O com Cov")
```

```
######## UTILIZANDO A CORREÇÃO DE BONFERRONI E VENDO SE ALGUM GENE RARO
É SIGNIFICATIVO NO TESTE DE BURDEN
ind_gene_raro_sign_bf_sem_cov_cr1_burden<-
which(as.numeric(sign_gene_cr1[,2])<0.05/length(genes_cr1))

gene_raro_sign_bf_sem_cov_cr1_burden<-
as.matrix(sign_gene_cr1[ind_gene_raro_sign_bf_sem_cov_cr1_burden,1:2])


ind_gene_raro_sign_bf_com_cov_cr1_burden<-
which(as.numeric(sign_gene_cr1_cov[,2])<0.05/length(genes_cr1))
gene_raro_sign_bf_com_cov_cr1_burden<-
as.matrix(sign_gene_cr1_cov[ind_gene_raro_sign_bf_com_cov_cr1_burden,1:2])



######## UTILIZANDO A CORREÇÃO DE BONFERRONI E VENDO SE ALGUM
GENE RARO É SIGNIFICATIVO NO SKAT
ind_gene_raro_sign_bf_sem_cov_cr1_skat<-
which(as.numeric(sign_gene_cr1[,3])<0.05/length(genes_cr1))
gene_raro_sign_bf_sem_cov_cr1_skat<-
as.matrix(sign_gene_cr1[ind_gene_raro_sign_bf_sem_cov_cr1_skat,c(1,3)])

ind_gene_raro_sign_bf_com_cov_cr1_skat<-
which(as.numeric(sign_gene_cr1_cov[,3])<0.05/length(genes_cr1))
gene_raro_sign_bf_com_cov_cr1_skat<-
as.matrix(sign_gene_cr1_cov[ind_gene_raro_sign_bf_com_cov_cr1_skat,c(1,3)])


######## UTILIZANDO A CORREÇÃO DE BONFERRONI E VENDO SE ALGUM
GENE RARO É SIGNIFICATIVO NO SKAT-O
ind_gene_raro_sign_bf_sem_cov_cr1_skat_o<-
```

```
which(as.numeric(sign_gene_cr1[,4])<0.05/length(genes_cr1))
gene_raro_sign_bf_sem_cov_cr1_skat_o<-
as.matrix(sign_gene_cr1[ind_gene_raro_sign_bf_sem_cov_cr1_skat_o,c(1,4,5)])

ind_gene_raro_sign_bf_com_cov_cr1_skat_o<-
which(as.numeric(sign_gene_cr1_cov[,4])<0.05/length(genes_cr1))
gene_raro_sign_bf_com_cov_cr1_skat_o<-
as.matrix(sign_gene_cr1_cov[ind_gene_raro_sign_bf_com_cov_cr1_skat_o,c(1,4,5)])


######## JUNTANDO AS INFORMAÇÕES DE TODOS OS 22 CROMOSSOMOS


###### ANÁLISE DESCRITIVA

#### AGREGANDO A INFORMAÇÃO DE QUAL O CROSSOMO PARA CADA FREQUÊNCIA
DO ALELO MENOR

### PARTE 1
MAF_1<- cbind(maf_1, rep(1,length(maf_1)))
MAF_2<- cbind(maf_2, rep(2,length(maf_2)))
MAF_3<- cbind(maf_3, rep(3,length(maf_3)))
MAF_4<- cbind(maf_4, rep(4,length(maf_4)))
MAF_5<- cbind(maf_5, rep(5,length(maf_5)))
MAF_6<- cbind(maf_6, rep(6,length(maf_6)))
MAF_7<- cbind(maf_7, rep(7,length(maf_7)))
MAF_8<- cbind(maf_8, rep(8,length(maf_8)))
MAF_9<- cbind(maf_9, rep(9,length(maf_9)))
MAF_10<- cbind(maf_10, rep(10,length(maf_10)))
MAF_11<- cbind(maf_11, rep(11,length(maf_11)))
MAF_tds<- rbind(MAF_1,MAF_2,MAF_3,MAF_4,MAF_5,MAF_6,
MAF_7,MAF_8,MAF_9,MAF_10,MAF_11)
MAF_tds<- as.data.frame(MAF_tds)
```

```
colnames(MAF_tds)<- c('MAF','Cr')
red_to_lilac_palette_1 <- colorRampPalette(c("lightcoral","orchid4"))(12)

### CONSTRUINDO BOXPLOT PARA O MAF DOS CROMOSSOMOS DE 1 A 11
boxplot(MAF_tds$MAF~MAF_tds$Cr, xlab = "Cromossomo", ylab= "MAF",
col=red_to_lilac_palette_1)

### CONSTRUINDO GRÁFICO DE BARRAS PARA OS CROMOSSOMOS DE 1 A 11
SNPs_1<- rbind(ncol(SNPs_comuns_cr1),ncol(SNPs_raros_cr1))
SNPs_2<- rbind(ncol(SNPs_comuns_cr2),ncol(SNPs_raros_cr2))
SNPs_3<- rbind(ncol(SNPs_comuns_cr3),ncol(SNPs_raros_cr3))
SNPs_4<- rbind(ncol(SNPs_comuns_cr4),ncol(SNPs_raros_cr4))
SNPs_5<- rbind(ncol(SNPs_comuns_cr5),ncol(SNPs_raros_cr5))
SNPs_6<- rbind(ncol(SNPs_comuns_cr6),ncol(SNPs_raros_cr6))
SNPs_7<- rbind(ncol(SNPs_comuns_cr7),ncol(SNPs_raros_cr7))
SNPs_8<- rbind(ncol(SNPs_comuns_cr8),ncol(SNPs_raros_cr8))
SNPs_9<- rbind(ncol(SNPs_comuns_cr9),ncol(SNPs_raros_cr9))
SNPs_10<- rbind(ncol(SNPs_comuns_cr10),ncol(SNPs_raros_cr10))
SNPs_11<- rbind(ncol(SNPs_comuns_cr11),ncol(SNPs_raros_cr11))
SNPs_10<- rbind(ncol(SNPs_comuns_cr10),ncol(SNPs_raros_cr10))
SNPs_11<- rbind(ncol(SNPs_comuns_cr11),ncol(SNPs_raros_cr11))
SNPs<- cbind(SNPs_1,SNPs_2,SNPs_3,SNPs_4,SNPs_5,SNPs_6,
SNPs_7,SNPs_8,SNPs_9,SNPs_10,SNPs_11)
rownames(SNPs)<- c("SNPs comuns", "SNPs raros")
colnames(SNPs)<- c("Cr 1", "Cr 2","Cr 3", "Cr 4", "Cr 5","Cr 6",
"Cr 7","Cr 8","Cr 9","Cr 10","Cr 11")
bp<- barplot(SNPs,horiz=F,beside=T, col=c("#a1e9f0", "#d9b1f0"),ylim=c(0,2500),
xlab="Cromossomo",ylab="Frequência absoluta")
legend("topright",inset = c(-0.13, -0.115),"#a1e9f0", "#d9b1f0",
legend=c("SNPs comuns","SNPs raros"), fill = c("#a1e9f0", "#d9b1f0"),bty="n")
text(x=c(bp), y=SNPs,labels=paste0(round(prop.table(SNPs, margin = 2) * 100,2),"%")
,pos=3)
```

```
### PARTE 2
MAF_12<- cbind(maf_12, rep(12,length(maf_12)))
MAF_13<- cbind(maf_13, rep(13,length(maf_13)))
MAF_14<- cbind(maf_14, rep(14,length(maf_14)))
MAF_15<- cbind(maf_15, rep(15,length(maf_15)))
MAF_16<- cbind(maf_16, rep(16,length(maf_16)))
MAF_17<- cbind(maf_17, rep(17,length(maf_17)))
MAF_18<- cbind(maf_18, rep(18,length(maf_18)))
MAF_19<- cbind(maf_19, rep(19,length(maf_19)))
MAF_20<- cbind(maf_20, rep(20,length(maf_20)))
MAF_21<- cbind(maf_21, rep(21,length(maf_21)))
MAF_22<- cbind(maf_22, rep(22,length(maf_22)))
MAF_tds_2<- rbind(MAF_12,MAF_13,MAF_14,MAF_15,MAF_16,
MAF_17,MAF_18,MAF_19,MAF_20,MAF_21,MAF_22)
MAF_tds_2<- as.data.frame(MAF_tds_2)
colnames(MAF_tds_2)<- c('MAF','Cr')


### CONSTRUINDO BOXPLOT PARA O MAF PARA OS CROMOSSOMOS DE 12 A 22
red_to_lilac_palette <- colorRampPalette(c("paleturquoise4","palegreen" ))(12)
boxplot(MAF_tds_2$MAF~MAF_tds_2$Cr, xlab = "Cromossomo", ylab= "MAF",
col=red_to_lilac_palette)


### CONSTRUINDO GRÁFICO DE BARRAS PARA OS CROMOSSOMOS DE 12 A 22
SNPs_12<- rbind(ncol(SNPs_comuns_cr12),ncol(SNPs_raros_cr12))
SNPs_13<- rbind(ncol(SNPs_comuns_cr13),ncol(SNPs_raros_cr13))
SNPs_14<- rbind(ncol(SNPs_comuns_cr14),ncol(SNPs_raros_cr14))
SNPs_15<- rbind(ncol(SNPs_comuns_cr15),ncol(SNPs_raros_cr15))
SNPs_16<- rbind(ncol(SNPs_comuns_cr16),ncol(SNPs_raros_cr16))
SNPs_17<- rbind(ncol(SNPs_comuns_cr17),ncol(SNPs_raros_cr17))
SNPs_18<- rbind(ncol(SNPs_comuns_cr18),ncol(SNPs_raros_cr18))
SNPs_19<- rbind(ncol(SNPs_comuns_cr19),ncol(SNPs_raros_cr19))
```

```
SNPs_20<- rbind(ncol(SNPs_comuns_cr20),ncol(SNPs_raros_cr20))
SNPs_21<- rbind(ncol(SNPs_comuns_cr21),ncol(SNPs_raros_cr21))
SNPs_22<- rbind(ncol(SNPs_comuns_cr22),ncol(SNPs_raros_cr22))
SNPs_v2<- cbind(SNPs_12,SNPs_13,SNPs_14,SNPs_15,SNPs_16,
SNPs_17,SNPs_18,SNPs_19,SNPs_20,SNPs_21,SNPs_22)
rownames(SNPs_v2)<- c("SNPs comuns", "SNPs raros")
colnames(SNPs_v2)<- c("Cr 12","Cr 13", "Cr 14", "Cr 15","Cr 16",
"Cr 17","Cr 18","Cr 19","Cr 20","Cr 21","Cr 22")
bp_2<- barplot(SNPs_v2,horiz=F,beside=T, col=c("#a1e9f0", "#d9b1f0"),ylim=c(0,2500),
xlab="Cromossomo",ylab="Frequência absoluta")
legend("topright",inset = c(-0.13, -0.115),"#a1e9f0", "#d9b1f0",
legend=c("SNPs comuns","SNPs raros"), fill = c("#a1e9f0", "#d9b1f0"),bty="n")
text(x=c(bp_2), y=SNPs_v2,labels=paste0(round(prop.table(SNPs_v2, margin = 2) *
100,2),"%"),pos=3)



####### RESULTADOS

##### PRIMEIRA ABORDAGEM: MODELO AJUSTADO SÓ COM O INTERCEPTO

#### RESULTADOS DOS SNPS RAROS SELECIONADOS COMO SIGNIFICATIVOS PARA O BURDEN NO
MODELO AJUSTADO SÓ COM INTERCEPTO

resul_modelo_s_cov_burden<-
rbind(gene_raro_sign_bf_sem_cov_cr1_burden,gene_raro_sign_bf_sem_cov_cr2_burden,
t(gene_raro_sign_bf_sem_cov_cr3_burden),gene_raro_sign_bf_sem_cov_cr4_burden,
t(gene_raro_sign_bf_sem_cov_cr5_burden),t(gene_raro_sign_bf_sem_cov_cr6_burden),
gene_raro_sign_bf_sem_cov_cr7_burden,gene_raro_sign_bf_sem_cov_cr8_burden,
gene_raro_sign_bf_sem_cov_cr9_burden,gene_raro_sign_bf_sem_cov_cr10_burden,
gene_raro_sign_bf_sem_cov_cr11_burden,gene_raro_sign_bf_sem_cov_cr12_burden,
t(gene_raro_sign_bf_sem_cov_cr13_burden),gene_raro_sign_bf_sem_cov_cr14_burden,
```

```
gene_raro_sign_bf_sem_cov_cr15_burden,t(gene_raro_sign_bf_sem_cov_cr16_burden),
t(gene_raro_sign_bf_sem_cov_cr17_burden),t(gene_raro_sign_bf_sem_cov_cr18_burden)
,gene_raro_sign_bf_sem_cov_cr19_burden,gene_raro_sign_bf_sem_cov_cr20_burden,
gene_raro_sign_bf_sem_cov_cr21_burden,t(gene_raro_sign_bf_sem_cov_cr22_burden))

resul_modelo_s_cov_burden_completo<- as.data.frame(resul_modelo_s_cov_burden)
colnames(resul_modelo_s_cov_burden_completo)<- c("gene","Valor-p")
resul_modelo_s_cov_burden_gene <- as.data.frame(resul_modelo_s_cov_burden[,1])
colnames(resul_modelo_s_cov_burden_gene)<- "gene"
resul_modelo_s_cov_burden_gene<- resul_modelo_s_cov_burden_gene %>%
left_join(snp_info,by='gene')
resul_modelo_s_cov_burden_gene<- resul_modelo_s_cov_burden_gene[,c(1,3)]
resul_modelo_s_cov_burden_gene<- distinct(resul_modelo_s_cov_burden_gene)

resul_modelo_s_cov_burden_completo<- resul_modelo_s_cov_burden_completo
%>% left_join(resul_modelo_s_cov_burden_gene,by='gene')


#### RESULTADOS DOS SNPS RAROS SELECIONADOS COMO SIGNIFICATIVOS PARA O
SKAT NO MODELO AJUSTADO SÓ COM INTERCEPTO

resul_modelo_s_cov_skat<-
rbind(gene_raro_sign_bf_sem_cov_cr1_skat,gene_raro_sign_bf_sem_cov_cr2_skat,
gene_raro_sign_bf_sem_cov_cr3_skat,gene_raro_sign_bf_sem_cov_cr4_skat,
t(gene_raro_sign_bf_sem_cov_cr5_skat),t(gene_raro_sign_bf_sem_cov_cr6_skat),
t(gene_raro_sign_bf_sem_cov_cr7_skat),gene_raro_sign_bf_sem_cov_cr8_skat,
gene_raro_sign_bf_sem_cov_cr9_skat,gene_raro_sign_bf_sem_cov_cr10_skat,
t(gene_raro_sign_bf_sem_cov_cr11_skat),gene_raro_sign_bf_sem_cov_cr12_skat,
t(gene_raro_sign_bf_sem_cov_cr13_skat),gene_raro_sign_bf_sem_cov_cr14_skat,
gene_raro_sign_bf_sem_cov_cr15_skat,gene_raro_sign_bf_sem_cov_cr16_skat,
t(gene_raro_sign_bf_sem_cov_cr17_skat),t(gene_raro_sign_bf_sem_cov_cr18_skat),
gene_raro_sign_bf_sem_cov_cr19_skat,gene_raro_sign_bf_sem_cov_cr20_skat,
gene_raro_sign_bf_sem_cov_cr21_skat,t(gene_raro_sign_bf_sem_cov_cr22_skat))
```

```
resul_modelo_s_cov_skat_completo<- as.data.frame(resul_modelo_s_cov_skat)
colnames(resul_modelo_s_cov_skat_completo)<- c("gene","Valor-p")
resul_modelo_s_cov_skat_gene <- as.data.frame(resul_modelo_s_cov_skat[,1])
colnames(resul_modelo_s_cov_skat_gene)<- "gene"
resul_modelo_s_cov_skat_gene<- resul_modelo_s_cov_skat_gene %>%
left_join(snp_info,by='gene')
resul_modelo_s_cov_skat_gene<- resul_modelo_s_cov_skat_gene[,c(1,3)]
resul_modelo_s_cov_skat_gene<- distinct(resul_modelo_s_cov_skat_gene)
resul_modelo_s_cov_skat_completo<- resul_modelo_s_cov_skat_completo %>%
left_join(resul_modelo_s_cov_skat_gene,by='gene')
```

```
#### RESULTADOS DOS SNPS RAROS SELECIONADOS COMO SIGNIFICATIVOS PARA O SKAT-O
NO MODELO AJUSTADO SÓ COM INTERCEPTO

resul_modelo_s_cov_skat_o<-
rbind(gene_raro_sign_bf_sem_cov_cr1_skat_o,gene_raro_sign_bf_sem_cov_cr2_skat_o,
t(gene_raro_sign_bf_sem_cov_cr3_skat_o),gene_raro_sign_bf_sem_cov_cr4_skat_o,
gene_raro_sign_bf_sem_cov_cr5_skat_o,gene_raro_sign_bf_sem_cov_cr6_skat_o,
t(gene_raro_sign_bf_sem_cov_cr7_skat_o),gene_raro_sign_bf_sem_cov_cr8_skat_o,
gene_raro_sign_bf_sem_cov_cr9_skat_o,gene_raro_sign_bf_sem_cov_cr10_skat_o,
t(gene_raro_sign_bf_sem_cov_cr11_skat_o),gene_raro_sign_bf_sem_cov_cr12_skat_o,
t(gene_raro_sign_bf_sem_cov_cr13_skat_o),gene_raro_sign_bf_sem_cov_cr14_skat_o,
gene_raro_sign_bf_sem_cov_cr15_skat_o,t(gene_raro_sign_bf_sem_cov_cr16_skat_o),
t(gene_raro_sign_bf_sem_cov_cr17_skat_o),t(gene_raro_sign_bf_sem_cov_cr18_skat_o),
gene_raro_sign_bf_sem_cov_cr19_skat_o,gene_raro_sign_bf_sem_cov_cr20_skat_o,
gene_raro_sign_bf_sem_cov_cr21_skat_o,t(gene_raro_sign_bf_sem_cov_cr22_skat_o))
```

```
resul_modelo_s_cov_skat_o_completo<- as.data.frame(resul_modelo_s_cov_skat_o)
colnames(resul_modelo_s_cov_skat_o_completo)<- c("gene","Valor-p","Rho")
resul_modelo_s_cov_skat_o_gene <- as.data.frame(resul_modelo_s_cov_skat_o[,1])
colnames(resul_modelo_s_cov_skat_o_gene)<- "gene"
resul_modelo_s_cov_skat_o_gene<- resul_modelo_s_cov_skat_o_gene %>%
left_join(snp_info,by='gene')
resul_modelo_s_cov_skat_o_gene<- resul_modelo_s_cov_skat_o_gene[,c(1,3)]
resul_modelo_s_cov_skat_o_gene<- distinct(resul_modelo_s_cov_skat_o_gene)
resul_modelo_s_cov_skat_o_completo<- resul_modelo_s_cov_skat_o_completo %>%
left_join(resul_modelo_s_cov_skat_o_gene,by='gene')



##### SEGUNDA ABORDAGEM: MODELO AJUSTADO COM AS COVARIÁVEIS
SELECIONADAS PELO LASSO



#### RESULTADOS DOS SNPS RAROS SELECIONADOS COMO SIGNIFICATIVOS PARA O BURDEN
NO MODELO AJUSTADO COM AS COVARIÁVEIS SELECIONADAS PELO LASSO

resul_modelo_c_cov_burden<-
rbind(gene_raro_sign_bf_com_cov_cr1_burden,gene_raro_sign_bf_com_cov_cr2_burden,
gene_raro_sign_bf_com_cov_cr3_burden,gene_raro_sign_bf_com_cov_cr4_burden,
gene_raro_sign_bf_com_cov_cr5_burden,gene_raro_sign_bf_com_cov_cr6_burden,
gene_raro_sign_bf_com_cov_cr7_burden,gene_raro_sign_bf_com_cov_cr8_burden,
gene_raro_sign_bf_com_cov_cr9_burden,gene_raro_sign_bf_com_cov_cr10_burden,
gene_raro_sign_bf_com_cov_cr11_burden,t(gene_raro_sign_bf_com_cov_cr12_burden),
t(gene_raro_sign_bf_com_cov_cr13_burden),gene_raro_sign_bf_com_cov_cr14_burden,
gene_raro_sign_bf_com_cov_cr15_burden,gene_raro_sign_bf_com_cov_cr16_burden,
```

```
t(gene_raro_sign_bf_com_cov_cr17_burden),t(gene_raro_sign_bf_com_cov_cr18_burden),
gene_raro_sign_bf_com_cov_cr19_burden,t(gene_raro_sign_bf_com_cov_cr20_burden),
gene_raro_sign_bf_com_cov_cr21_burden,gene_raro_sign_bf_com_cov_cr22_burden)

resul_modelo_c_cov_burden_completo<- as.data.frame(resul_modelo_c_cov_burden)
colnames(resul_modelo_c_cov_burden_completo)<- c("gene","Valor-p")
resul_modelo_c_cov_burden_gene <- as.data.frame(resul_modelo_c_cov_burden[,1])
colnames(resul_modelo_c_cov_burden_gene)<- "gene"
resul_modelo_c_cov_burden_gene<- resul_modelo_c_cov_burden_gene %>%
left_join(snp_info,by='gene')
resul_modelo_c_cov_burden_gene<- resul_modelo_c_cov_burden_gene[,c(1,3)]
resul_modelo_c_cov_burden_gene<- distinct(resul_modelo_c_cov_burden_gene)
resul_modelo_c_cov_burden_completo<- resul_modelo_c_cov_burden_completo %>%
left_join(resul_modelo_c_cov_burden_gene,by='gene')


#### RESULTADOS DOS SNPS RAROS SELECIONADOS COMO SIGNIFICATIVOS PARA O SKAT
NO MODELO AJUSTADO COM AS COVARIÁVEIS SELECIONADAS PELO LASSO

resul_modelo_c_cov_skat<-
rbind(gene_raro_sign_bf_com_cov_cr1_skat,t(gene_raro_sign_bf_com_cov_cr2_skat),
gene_raro_sign_bf_com_cov_cr3_skat,gene_raro_sign_bf_com_cov_cr4_skat,
gene_raro_sign_bf_com_cov_cr5_skat,gene_raro_sign_bf_com_cov_cr6_skat,
gene_raro_sign_bf_com_cov_cr7_skat,gene_raro_sign_bf_com_cov_cr8_skat,
t(gene_raro_sign_bf_com_cov_cr9_skat),gene_raro_sign_bf_com_cov_cr10_skat,
gene_raro_sign_bf_com_cov_cr11_skat,t(gene_raro_sign_bf_com_cov_cr12_skat),
t(gene_raro_sign_bf_com_cov_cr13_skat),gene_raro_sign_bf_com_cov_cr14_skat,
gene_raro_sign_bf_com_cov_cr15_skat,gene_raro_sign_bf_com_cov_cr16_skat,
t(gene_raro_sign_bf_com_cov_cr17_skat),t(gene_raro_sign_bf_com_cov_cr18_skat),
gene_raro_sign_bf_com_cov_cr19_skat,gene_raro_sign_bf_com_cov_cr20_skat,
gene_raro_sign_bf_com_cov_cr21_skat,gene_raro_sign_bf_com_cov_cr22_skat)

resul_modelo_c_cov_skat_completo<- as.data.frame(resul_modelo_c_cov_skat)
```

```
colnames(resul_modelo_c_cov_skat_completo)<- c("gene","Valor-p")
resul_modelo_c_cov_skat_gene <- as.data.frame(resul_modelo_c_cov_skat[,1])
colnames(resul_modelo_c_cov_skat_gene)<- "gene"
resul_modelo_c_cov_skat_gene<- resul_modelo_c_cov_skat_gene %>%
left_join(snp_info,by='gene')
resul_modelo_c_cov_skat_gene<- resul_modelo_c_cov_skat_gene[,c(1,3)]
resul_modelo_c_cov_skat_gene<- distinct(resul_modelo_c_cov_skat_gene)
resul_modelo_c_cov_skat_completo<- resul_modelo_c_cov_skat_completo %>%
left_join(resul_modelo_c_cov_skat_gene,by='gene')


#### RESULTADOS DOS SNPS RAROS SELECIONADOS COMO SIGNIFICATIVOS PARA O SKAT-O
NO MODELO AJUSTADO COM AS COVARIÁVEIS SELECIONADAS PELO LASSO

resul_modelo_c_cov_skat_o<-
rbind(gene_raro_sign_bf_com_cov_cr1_skat_o,gene_raro_sign_bf_com_cov_cr2_skat_o,
gene_raro_sign_bf_com_cov_cr3_skat_o,gene_raro_sign_bf_com_cov_cr4_skat_o,
gene_raro_sign_bf_com_cov_cr5_skat_o,gene_raro_sign_bf_com_cov_cr6_skat_o,
gene_raro_sign_bf_com_cov_cr7_skat_o,gene_raro_sign_bf_com_cov_cr8_skat_o,
t(gene_raro_sign_bf_com_cov_cr9_skat_o),gene_raro_sign_bf_com_cov_cr10_skat_o,
gene_raro_sign_bf_com_cov_cr11_skat_o,gene_raro_sign_bf_com_cov_cr12_skat_o,
t(gene_raro_sign_bf_com_cov_cr13_skat_o),gene_raro_sign_bf_com_cov_cr14_skat_o,
gene_raro_sign_bf_com_cov_cr15_skat_o,gene_raro_sign_bf_com_cov_cr16_skat_o,
t(gene_raro_sign_bf_com_cov_cr17_skat_o),t(gene_raro_sign_bf_com_cov_cr18_skat_o),
gene_raro_sign_bf_com_cov_cr19_skat_o,t(gene_raro_sign_bf_com_cov_cr20_skat_o),
gene_raro_sign_bf_com_cov_cr21_skat_o,gene_raro_sign_bf_com_cov_cr22_skat_o)

resul_modelo_c_cov_skat_o_completo<- as.data.frame(resul_modelo_c_cov_skat_o)
colnames(resul_modelo_c_cov_skat_o_completo)<- c("gene","Valor-p","Rho")
resul_modelo_c_cov_skat_o_gene <- as.data.frame(resul_modelo_c_cov_skat_o[,1])
colnames(resul_modelo_c_cov_skat_o_gene)<- "gene"
resul_modelo_c_cov_skat_o_gene<- resul_modelo_c_cov_skat_o_gene %>%
left_join(snp_info,by='gene')
```

```
resul_modelo_c_cov_skat_o_gene<- resul_modelo_c_cov_skat_o_gene[,c(1,3)]
resul_modelo_c_cov_skat_o_gene<- distinct(resul_modelo_c_cov_skat_o_gene)
resul_modelo_c_cov_skat_o_completo<- resul_modelo_c_cov_skat_o_completo %>%
left_join(resul_modelo_c_cov_skat_o_gene,by='gene')



#### RESULTADOS DAS VARIÁVEIS E SNPS COMUNS SELECIONADOS COMO
SIGNIFICATIVASPELO LASSO

result_var_comuns_sign<-
rbind(snps_comuns_sign_cr1,snps_comuns_sign_cr2,snps_comuns_sign_cr3,
snps_comuns_sign_cr4,snps_comuns_sign_cr5,snps_comuns_sign_cr6,
snps_comuns_sign_cr7,snps_comuns_sign_cr8,snps_comuns_sign_cr9,
snps_comuns_sign_cr10,snps_comuns_sign_cr11,snps_comuns_sign_cr12,
snps_comuns_sign_cr13,snps_comuns_sign_cr14,snps_comuns_sign_cr15,
snps_comuns_sign_cr16,snps_comuns_sign_cr17,snps_comuns_sign_cr18,
snps_comuns_sign_cr19,snps_comuns_sign_cr20,snps_comuns_sign_cr21,
snps_comuns_sign_cr22)
nomes_var_comuns_sign<- as.data.frame(result_var_comuns_sign[,1])
colnames(nomes_var_comuns_sign)<- "Variáveis comuns significativas"
nomes_var_comuns_sign<- distinct(nomes_var_comuns_sign)



##### ANALISANDO QUANTIDADE DE GENES E SNPS COMUNS
PARA CALCULAR MEDIDAS DE DESEMPENHO



#### QUANTIDADE DE GENES DO GAW 17
gene<- as.data.frame(snp_info$gene)
View(gene)
gene<- distinct(gene)
```

```
#### QUANTIDADE DE SNPS COMUNS E RAROS
qtd_snps_comuns <-
ncol(SNPs_comuns_cr1)+ncol(SNPs_comuns_cr2)+ncol(SNPs_comuns_cr3)+
ncol(SNPs_comuns_cr4)+ncol(SNPs_comuns_cr5)+ncol(SNPs_comuns_cr6)+
ncol(SNPs_comuns_cr7)+ncol(SNPs_comuns_cr8)+ncol(SNPs_comuns_cr9)+
ncol(SNPs_comuns_cr10)+ncol(SNPs_comuns_cr11)+ncol(SNPs_comuns_cr12)+
ncol(SNPs_comuns_cr13)+ncol(SNPs_comuns_cr14)+ncol(SNPs_comuns_cr15)+
ncol(SNPs_comuns_cr16)+ncol(SNPs_comuns_cr17)+ncol(SNPs_comuns_cr18)+
ncol(SNPs_comuns_cr19)+ncol(SNPs_comuns_cr20)+ncol(SNPs_comuns_cr21)+
ncol(SNPs_comuns_cr22)
qtd_snps_comuns

qtd_snps_raros <-
ncol(SNPs_raros_cr1)+ncol(SNPs_raros_cr2)+ncol(SNPs_raros_cr3)+
ncol(SNPs_raros_cr4)+ncol(SNPs_raros_cr5)+ncol(SNPs_raros_cr6)+
ncol(SNPs_raros_cr7)+ncol(SNPs_raros_cr8)+ncol(SNPs_raros_cr9)+
ncol(SNPs_raros_cr10)+ncol(SNPs_raros_cr11)+ncol(SNPs_raros_cr12)+
ncol(SNPs_raros_cr13)+ncol(SNPs_raros_cr14)+ncol(SNPs_raros_cr15)+
ncol(SNPs_raros_cr16)+ncol(SNPs_raros_cr17)+ncol(SNPs_raros_cr18)+
ncol(SNPs_raros_cr19)+ncol(SNPs_raros_cr20)+ncol(SNPs_raros_cr21)+
ncol(SNPs_raros_cr22)
qtd_snps_raros
```