UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE FÍSICA DE SÃO CARLOS

DIOGO MACIEL DUARTE DA MOTA

Análise do componente genético do Diabetes Mellitus tipo 2

São Carlos/SP
2024

DIOGO MACIEL DUARTE DA MOTA

Análise do componente genético do Diabetes Mellitus tipo 2

Tese apresentada ao Programa de Pós-Graduação em Física do Instituto de Física de São Carlos da Universidade de São Paulo, para a obtenção do título de Doutor em Ciências.

Área de concentração: Física Aplicada
Opção: Física Biomolecular
Orientador: Prof. Dr. João Carlos Setubal
Coorientador: Prof. Dr. Alexander Augusto de Lima Jorge

Versão Corrigida

(Versão original disponível na Unidade que aloja o Programa)

São Carlos/SP

2024

Mota, Diogo Maciel Duarte da
Análise do componente genético do Diabetes Mellitus tipo 2 / Diogo Maciel Duarte da Mota; orientador João Carlos Setubal; co-orientador Alexander Augusto de Lima Jorge - versão corrigida -- São Carlos, 2024.
114 p.

Tese (Doutorado - Programa de Pós-Graduação em Física Aplicada Biomolecular) -- Instituto de Física de São Carlos, Universidade de São Paulo, 2024.

1. Diabetes Mellitus Tipo 2. 2. Genome-Wide association studies. 3. Single nucleotide polymorphism. 4. Escores de risco poligênico. 5. População brasileira. I. Setubal, João Carlos, orient. II. Jorge, Alexander Augusto de Lima, co-orient. III. Título.

AGRADECIMENTOS PESSOAIS

Primeiramente, gostaria de agradecer aos meus orientadores e coorientadores, o Prof. Dr. Ricardo De Marco (*in memoriam*), o Prof. Dr. João Carlos Setubal e o Prof. Dr. Alexander Augusto de Lima Jorge, cuja orientação, apoio constante e incansável paixão pela pesquisa foram fundamentais para este trabalho. Suas orientações não apenas me ajudaram a desenvolver minhas habilidades acadêmicas, mas também me inspiraram a sempre buscar o melhor.

Ao meu pai Gilberto (*in memoriam*), à minha mãe Margarete, e ao meu sobrinho, Lorenzo, meu sincero agradecimento pelo apoio e encorajamento ao longo de todos esses anos. Suas palavras de incentivo e amor me fortaleceram nos momentos mais difíceis.

À minha esposa, Tatiane, meu maior agradecimento por sua paciência, compreensão e apoio. Seu amor e companheirismo foram meu alicerce durante esta jornada.

Ao meu amigo, Raphael Montanari, pelo constante apoio nas análises computacionais realizadas ao longo desses anos.

Aos meus amigos, Diego e Stefany, e os seus filhos Emma e Isaac, por todo o suporte emocional durante minha jornada em São Carlos. E a todos os amigos e colegas que compartilharam ideias, debates e momentos ao longo desta trajetória.

Aos meus demais professores, por todo o compartilhamento de conhecimento e experiências que foram fundamentais para o meu desenvolvimento como pesquisador.

Aos técnicos administrativos da pós-graduação do Instituto de Física de São Carlos (IFSC), Sílvio e Ricardo, por todo o suporte e esclarecimentos necessários ao longo desse processo.

Muito obrigado a todos.

AGRADECIMENTOS INSTITUCIONAIS

Gostaria de agradecer à Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) pela bolsa de doutorado direto concedida, sob o processo 2018/11907-0, que tornou possível a realização desta pesquisa.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo período de bolsa de mestrado, sob o processo 88887.313571/2019-00.

Ao Instituto de Física de São Carlos (IFSC) por fornecer um ambiente acadêmico estimulante e os recursos essenciais para que a pesquisa fosse realizada.

Muito obrigado.

"Seja curioso. Leia muito. Experimente coisas novas. Acredito que muito do que as pessoas chamam de inteligência apenas se resume a curiosidade."

Aaron Swartz

RESUMO

MOTA, D. M. D. **Análise do componente genético do Diabetes Mellitus tipo 2.** 2023. 114 p. Tese (Doutorado em Ciências) – Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos, 2024.

O Diabetes Mellitus Tipo 2 DM2 representa um conjunto de complexas condições metabólicas caracterizadas pela hiperglicemia devido a defeitos na secreção ou na ação da insulina. A hiperglicemia crônica relacionada ao DM2 tem efeitos prejudiciais a longo prazo em diversos órgãos e é uma das principais causas de mortalidade em adultos. Em seu aspecto multifatorial, a complexidade genética do DM2, com múltiplos *loci* gênicos e sua interação com o ambiente desempenham um papel relevante na manifestação da doença. Através de *Genome-Wide Association Studies* (GWAS), já foram descritos mais de 800 *single nucleotide polymorphisms* (SNPs), distribuídos em mais de 500 *loci* de suscetibilidade ao DM2. O presente estudo tem como objetivos investigar a alta prevalência do DM2 a partir de características estruturais e evolutivas das regiões que apresentam SNPs associados ao fenótipo, além de realizar um estudo destes SNPs em uma população brasileira (SABE), do Arquivo Brasileiro Online de Mutações – ABraOM, para identificar variantes com características específicas. Os resultados revelam que os genes com SNPs associados ao DM2 estão concentrados em regiões de menor variação da expressão e menor especificidade de tecido, sugerindo perfis de expressão mais estáveis, característicos de genes de *housekeeping*. Além disso, foi identificado o enriquecimento de variantes raras em 41 genes, 24 genes com variantes sinônimas e 17 com variantes não-sinônimas, relacionadas à perda de função e com potencial deletério. Foram validados, também, 63 SNPs com *odds ratio* (OR) estatisticamente significativo, em que 11 deles tiveram um efeito protetivo na população brasileira, em oposição aos resultados dos GWAS originais. A heterogeneidade nas variantes sugere a existência de efeitos genéticos específicos para a população do SABE. Porém, é destacada a necessidade de amostras de dados maiores para obter resultados mais precisos, com tamanho ideal na ordem de centenas de milhares a um milhão de indivíduos. Foi avaliado o desempenho de modelos de Escore de Risco Poligênico (PRS) em relação ao DM2 na população brasileira, os modelos apresentaram desempenho variado, com os resultados representando um importante para avanço para estudos de análise de risco do DM2 na população brasileira. Por fim, este estudo contribui para a compreensão da complexidade genética do DM2 e a importância de estudar populações específicas, destacando a necessidade de amostras maiores para obter resultados mais robustos e o potencial dos

modelos de PRS na identificação de riscos genéticos associados ao DM2 na população brasileira.

Palavras-chave: Diabetes Mellitus Tipo 2. *Genome-Wide association studies*. *Single nucleotide polymorphism*. Escores de risco poligênico. População brasileira.

ABSTRACT

MOTA, D. M. D. **Analysis of genetic component of type 2 Diabetes Mellitus.** 2023. 114 p. Thesis (Doctor in Science) – Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos, 2024.

Type 2 Diabetes Mellitus (T2D) represents a set of complex metabolic conditions characterized by hyperglycemia due to defects in insulin secretion or its action. The chronic hyperglycemia associated with T2D has long-term detrimental effects on various organs and is among the top causes of mortality in adults. In its multifactorial nature, the genetic complexity of T2D, involving multiple gene loci and their interaction with the environment, plays a relevant role in disease manifestation. Through Genome-Wide Association Studies (GWAS), more than 800 single nucleotide polymorphisms (SNPs) distributed across over 500 susceptibility loci for T2D have been described. This study aims to investigate the high prevalence of T2D based on structural and evolutionary characteristics of the genomic regions that present SNPs associated with the phenotype, in addition to carrying out a study of these SNPs in a brazilian population (SABE), from the Brazilian Online Mutation Archive – ABraOM, to identify variants with specific characteristics. The results reveal that genes with SNPs associated with T2D are concentrated in regions with lower tissue specificity, lower expression variation, suggesting more stable expression profiles, characteristic of housekeeping genes. Furthermore, the study identified an enrichment of rare variants in 41 genes, with 24 genes having synonymous variants and 17 with non-synonymous variants related to loss of function and potential deleterious effects. 63 SNPs with statistically significant odds ratios (OR) were also validated, in which 11 of them had a protective effect in the brazilian population, in contrast to the results of the original GWAS. The heterogeneity in variants suggests the existence of population specific genetic effects in the SABE population. However, the need for larger data samples to obtain more accurate results is emphasized, with an ideal sample size in the order of hundreds of thousands to one million individuals. The performance of Polygenic Risk Score (PGS) models for T2D in the brazilian population was evaluated, with models showing varied performance. These results represent a significant step forward for T2D risk analysis studies in the brazilian population. In conclusion, this study contributes to understanding the genetic complexity of T2D and the importance of studying specific populations, emphasizing the need for larger sample sizes to obtain more robust results, and highlighting the potential of PGS models in identifying genetic risks associated with T2D in the brazilian population.

Keywords: Type 2 diabetes mellitus. Genome-wide association studies. Single nucleotide polymorphism. Polygenic risk score. Brazilian population.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE GRÁFICOS

SUMÁRIO

1 INTRODUÇÃO GERAL

Com o intuito de evidenciar os princípios que direcionam esta pesquisa, nesta seção inicial serão apresentadas a contextualização do tema, a definição dos objetivos, bem como a descrição da estrutura que norteará o estudo.

1.1 Diabetes Mellitus

1.1.1 Conceito e Características

Diabetes Mellitus (DM) representa um conjunto de condições metabólicas complexas caracterizadas por hiperglicemia causada por defeitos na secreção de insulina, na sua ação, ou ambos os casos. A hiperglicemia crônica associada ao DM está relacionada a danos de longo prazo, como disfunção e falência de diversos órgãos (1-2). O DM possui um grande impacto na vida e no bem-estar dos indivíduos, sendo classificada entre as 10 (dez) principais causas de morte em adultos, com uma estimativa de 6,7 milhões de mortes em 2021, em todo o mundo (3).

O desenvolvimento do DM envolve vários processos patogênicos que variam desde a destruição autoimune das células β (beta) do pâncreas, no Diabetes Mellitus Tipo 1 (DM1), resultando em deficiência de insulina grave, até anormalidades que levam à resistência à ação da insulina com deficiência relativa (2). No Diabetes Mellitus Tipo 2 (DM2) a diminuição na secreção da insulina e os defeitos na sua ação frequentemente coexistem no mesmo indivíduo (1-2) e estão associados, muitas vezes, a uma resistência pré-existente à insulina, principalmente, nos tecidos adiposos, musculares esqueléticos e do fígado (4).

Os fatores de risco para o DM2 incluem um conjunto complexo de fatores genéticos, metabólicos e ambientais que contribuem para sua prevalência (5). Dentre esses fatores estão histórico familiar ou predisposição genética, ancestralidade e idade, classificados como não modificáveis; e obesidade, sedentarismo e dieta pouco saudável, como modificáveis (1,5).

Os sintomas característicos de uma hiperglicemia acentuada, incluem poliúria (micção frequente), polidipsia (sede excessiva), perda de peso, às vezes acompanhada de polifagia (aumento do apetite) e visão turva. Além disso, a hiperglicemia crônica pode aumentar a suscetibilidade a certas infecções. As complicações de longo prazo, são classificadas em dois grupos: as microvasculares, que incluem a retinopatia, que pode causar perda da visão, a

nefropatia, que pode levar à insuficiência renal, a neuropatia periférica, que pode resultar em úlceras nos pés e amputações, e a neuropatia autonômica, que pode causar sintomas gastrointestinais, geniturinários, cardiovasculares e disfunções sexuais; e as macrovasculares, provocando um risco aumentado para doença cardiovascular aterosclerótica, doença arterial periférica e doenças cerebrovasculares (6-7), além de hipertensão e anormalidades no metabolismo das lipoproteínas (2).

1.1.2 Classificação e Diagnóstico

O DM pode ser categorizado, principalmente, nas seguintes classificações gerais: DM1, que corresponde a 5-10% dos casos, incluindo o diabetes autoimune latente do adulto (*Latent Autoimune Diabetes in Adults* – LADA), e DM2, compreendendo 90-95% dos casos (2). Existem também tipos específicos e mais raros de DM decorrentes de outras causas, como diabetes neonatal (DMN), diabetes tipo MODY (*Maturity Onset Diabetes of Young*), uma forma monogênica que apresenta uma herança autossômica dominante e alta penetrância, mas não apresenta caráter autoimune, além de outras síndromes com padrão monogênico, doenças do pâncreas exócrino (como fibrose cística e pancreatite) e DM induzido por medicamentos ou produtos químicos (como o uso de glicocorticoides no tratamento de HIV/AIDS ou após transplante de órgãos) (1). Por fim, há também o diabetes mellitus gestacional (DMG), que é diagnosticado no segundo ou terceiro trimestre da gravidez, em mulheres que não apresentavam evidências claras de DM antes da gestação (2,8).

Com relação aos níveis de glicemia, eles podem alterar com o tempo, dependendo da extensão do processo, que pode causar glicemia de jejum alterada ou tolerância diminuída à glicose, sem preencher os critérios para o diagnóstico do DM, representando um processo progressivo, onde alterações pontuais na glicemia, seja em jejum ou pós-prandial, precedem uma hiperglicemia sustentada. Em alguns indivíduos, o controle glicêmico adequado pode ser alcançado com redução de peso, prática de exercícios e uso de medicamentos orais. Outros indivíduos, que apresentam baixa secreção residual de insulina, necessitam do uso de insulina exógena, associado com outras formas de tratamento, para obter um controle glicêmico adequado. Já os indivíduos com uma destruição acentuada das células β, sem secreção residual, necessitam do uso contínuo de insulina exógena para sobreviver (2).

Classificar indivíduos quanto ao tipo de DM depende das circunstâncias presentes no momento do diagnóstico, muitos indivíduos não enquadram facilmente em uma classe específica, tanto o DM1 quanto o DM2 são condições complexas, nas quais a apresentação

clínica e a progressão podem variar significativamente, uma classificação adequada é essencial para determinar o plano terapêutico. No entanto, em alguns casos, é difícil fazer uma distinção clara no momento do diagnóstico. É importante saber que a classificação correta do tipo de DM nem sempre é direta no momento da apresentação e erros de diagnósticos são comuns (2). Em todos os tipos de DM, diversos fatores genéticos e ambientais podem levar à perda progressiva de massa ou função das células β, resultando em hiperglicemia clinicamente manifesta. Uma vez que a hiperglicemia se estabelece, todos os indivíduos estão em risco de desenvolver as mesmas complicações crônicas, embora as taxas de progressão possam variar (9).

Os testes utilizados para o rastreamento e diagnóstico do DM podem ser feitos com base em critérios de glicose plasmática, através do valor da glicemia de jejum e da glicemia de 2 horas (2hPG) durante um teste oral de tolerância à glicose com 75 g de glicose anidra (2), ou por critérios de hemoglobina glicada (HbA1c), que apresenta um índice integrado de glicemia durante a vida útil de 120 dias dos glóbulos vermelhos, indicando o percentual dessas células que possuem hemoglobina aderida pelo excesso de glicemia no sangue (10-11) (Tabela 1.1). No caso de um indivíduo assintomático, são utilizados os critérios de uma glicemia de jejum igual ou superior a 126 mg/dL, uma glicemia de 2 horas após uma sobrecarga de 75 g de glicose anidra igual ou superior a 200 mg/dL ou um percentual de HbA1c igual ou superior a 6,5%. É importante que, pelo menos, dois dos exames clínicos estejam alterados para confirmar o diagnóstico. Se apenas um exame apresentar alteração, é necessário repeti-lo para confirmação (2).

Tabela 1.1 - Critérios para o diagnóstico de DM.

Critérios para o Diagnóstico de DM
Glicemia de Jejum* ≥ 126 mg/dL.
Glicemia de 2 horas em um Teste Oral de Tolerância à Glicose ≥ 200 mg/dL.
Hemoglobina Glicada (HbA1c) ≥ 6,5%.
Glicemia Eventual** ≥ 200 mg/dL.

*O jejum é definido como um período mínimo de 8 horas sem a ingestão de calorias.
**Valores observados em indivíduos que apresentam sintomas clássicos de hiperglicemia.
Fonte: Adaptada de ELSAYED *et al* (2).

1.1.3 Epidemiologia do Diabetes Mellitus Tipo 2

1.1.3.1 Diabetes Mellitus Tipo 2 no Mundo

A manifestação do DM2 é complexa e influenciada pela união de diversos fatores de risco. Para tentar reduzir, de maneira significativa, a enorme morbidade e mortalidade

prematura que o DM2 causa, é necessário que a prevenção e o controle de suas complicações possuam uma abordagem ampla, integrada e global (3,12-13).

A incidência mundial de DM2 tem apresentado um crescimento contínuo ao longo de mais de quatro décadas, atingindo proporções pandêmicas. Em 2021, estimou-se que a prevalência de DM2 diagnosticado e não diagnosticado fosse de 10,5%, o que representa cerca de 537 milhões de pessoas. Projeções indicam que esse número deverá aumentar para 11,3% (643 milhões) até 2030 e 12,2% (783 milhões) até 2045 (Figura 1.1) (3). O aumento dessa incidência de DM2 está diretamente ligado ao envelhecimento da população e à adesão mais abrangente a hábitos de vida pouco saudáveis, o que resulta em uma maior proporção de pessoas com obesidade (1,12).

Para atingir objetivos de prevenção, cuidado e tratamento do DM2, assim como reduzir sua crescente ameaça, é fundamental ter a capacidade de mensurar sua prevalência e incidência, bem como seus determinantes, como fatores de risco. Além disso, é importante compreender suas consequências, que incluem complicações, mortalidade prematura, redução da qualidade de vida e aumento dos custos com saúde. Essa capacidade de medição é um pré-requisito essencial para avaliar de forma abrangente o panorama do DM2 e seus efeitos (14).

Paralelo ao aumento da prevalência, também é observado um notável crescimento nos custos econômicos para os sistemas de saúde atribuíveis ao DM2. Em 2007, estimou-se que os gastos diretos com saúde decorrentes do DM2 foram de US$ 232 bilhões, enquanto em 2021 esse valor subiu para aproximadamente US$ 966 bilhões. Atualmente, cerca de 80% dos casos de DM2 ocorrem em países de baixa e média renda, e as projeções indicam que os maiores aumentos continuarão a ocorrer nessas regiões nas próximas décadas (Figura 1.1) (3).

As consequências do DM2 para a saúde são graves e debilitantes, frequentemente resultando em morte prematura e redução da produtividade no trabalho. Conter o crescimento do DM2 em escala global é de suma importância para mitigar os custos econômicos e melhorar a saúde e o bem-estar de indivíduos e populações (3,5,13,15-16).

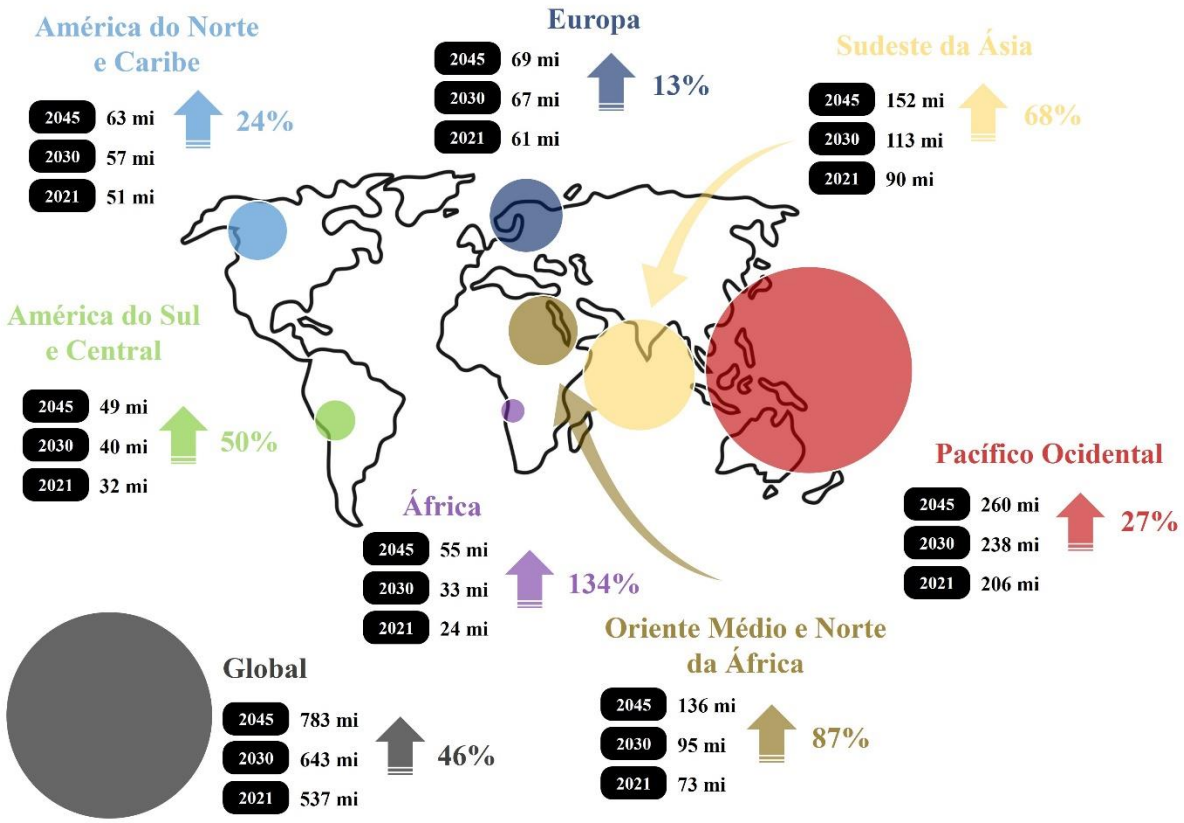

Figura 1.1 – Número de casos de DM2, global e por região, em 2021 e projetados para 2030 e 2045.
Fonte: Adaptada de SUN *et al*. (3).

1.1.3.2 Diabetes Mellitus Tipo 2 no Brasil

De acordo com o 10º *Diabetes Atlas* da *International Diabetes Federation* (IDF), de 2021, estima-se que existam aproximadamente 15,7 milhões de pessoas com DM2 no Brasil (~10,5%), na faixa etária de 20 a 79 anos. Além disso, há aproximadamente 5,0 milhões de casos não diagnosticados, o que representa uma proporção de subdiagnóstico de 31,9% (3), além de vários estudos revelarem informações alarmantes sobre o pré-diabetes no Brasil (17).

A prevalência de DM no Brasil, para todas as faixas etárias, em 2017, foi de 4,4%, sendo 6,2% ao considerar indivíduos com faixa etária entre 20 e 79 anos, com DM2 representando 96% dos casos. Entre os anos de 1990 e 2017, houve um aumento nessa prevalência de 30% em indivíduos do sexo masculino e 26% do sexo feminino, apresentando também uma considerável variabilidade geográfica. Os estados do Nordeste apresentam uma prevalência geralmente maior de DM2 do que outras regiões, além das regiões Norte, Nordeste e Centro-Oeste sofrerem os maiores aumentos percentuais durante esse período. Embora exista também um aumento na incidência do DM2 durante esse tempo, ao padronizar esses números por idade, essa incidência permaneceu estável. Isso mostra que os aumentos da prevalência do DM2, no

Brasil, está relacionado, principalmente, com o envelhecimento populacional e uma maior taxa de sobrevivência dos indivíduos (18).

A Pesquisa Nacional de Saúde (PNS), realizada entre os anos de 2014 e 2015, que utilizou a HbA1c como ferramenta de diagnóstico em uma amostra de mais de 8.500 indivíduos, identificou uma prevalência de pré-diabetes entre 6,8% e 16,9%, dependendo dos critérios analisados (19). O Estudo Longitudinal de Saúde do Adulto (ELSA) (http://elsabrasil.org/), que envolveu adultos brasileiros, revelou uma prevalência de pré-diabetes variando de 20% a 59% na amostra recrutada, composta por profissionais universitários (20).

Além dos desafios relacionados ao diagnóstico, as taxas de controle do DM2 no Brasil continuam insatisfatórias, um estudo abrangente realizado em âmbito nacional para avaliar o controle da glicemia realizado em 2006, constatou que 75% dos indivíduos com DM2, atendidos em serviços públicos ou privados, por especialistas ou não especialistas, apresentavam níveis de HbA1c acima de 7% (21).

1.1.4 Bases Genéticas do Diabetes Mellitus Tipo 2

A manifestação do DM2 possui caráter multifatorial e é dependente de polimorfismos em múltiplos *loci* gênicos, em que a interação entre eles e o ambiente serão determinantes para o surgimento e o desenvolvimento da síndrome (4-5,22). É um fenótipo poligênico, em que as contribuições individuais de cada variante genética associada, presentes em diversos genes diferentes, tendem a ser moderadas, sugerindo que a maioria delas não gerará um cenário catastrófico, como a perda de atividade de uma proteína ou a completa desestruturação de um gene. Devido à alta prevalência dos alelos associados ao DM2, pressupõe-se que, durante a evolução da espécie humana, esses apresentaram caráter próximo de neutro ou positivo (23-25).

Embora a base genética desempenhe um papel fundamental no risco de desenvolver o DM2 (5), diversos outros fatores devem ser considerados, ela é suficiente apenas nas formas mendelianas, como os casos monogênicos (26), essa característica é dada devido a herdabilidade, uma medida quantitativa que expressa o quanto da variabilidade populacional presente em uma característica específica se deve a variações genéticas (27). Estudos de gêmeos e familiares há muito tempo sugerem um componente genético para a suscetibilidade ao DM2, demonstrando, para gêmeos monozigóticos, uma concordância alta de 70% a 90% (26). Porém, conforme apoiado pelos resultados recentes dos estudos de associação ampla do genoma (*Genome-Wide Association Studies* – GWAS), a heterogeneidade genética é uma explicação

mais provável, em que o estilo de vida e o ambiente também são críticos para o desenvolvimento do DM2, resultando em uma herdabilidade estimada entre 26% e 69% (26,28-29).

Uma abordagem para categorizar os dados genéticos ao risco de desenvolvimento do DM2 é o escore de risco poligênico (*polygenic risk score* – PRS), que possui seu resultado dependente da herdabilidade estimada por variantes genéticas. Isso significa que, por melhor que seja o PRS aplicado, a herdabilidade irá limitar o percentual total explicado do fenótipo. Apesar dos estudos avaliarem a importância da hereditariedade na etiologia do DM2, a identificação de variantes genéticas utilizando GWAS explica apenas cerca de 10% da sua herdabilidade, indicando que o restante deste percentual pode ser atribuído aos fatores de interação dos genes com o ambiente (26).

A alta prevalência do DM2 representa um enigma evolutivo, visto que seria esperado que variantes com caráter deletério e forte influência genética fossem eliminadas pela seleção natural. Como resposta a essa questão, foram levantadas diversas hipóteses que sugerem que características associadas ao DM2 poderiam ser evolutivamente vantajosas em certos contextos da evolução humana (24,30-31). Um exemplo é a hipótese do *Thrifty Genotype* que sugere que populações que apresentavam alto risco de fome eram favorecidas por um fenótipo de resistência à insulina e tendência à obesidade (Figura 1.2) (32-36). Essas hipóteses são contestadas por outros cientistas que sugerem um cenário evolutivo onde tais mutações adquiririam um caráter próximo à neutralidade e se espalharam na população humana através de fenômenos de deriva genética (22-23,37).

O DM2 aumentou rapidamente em prevalência nos últimos anos e representa um componente importante na carga global de doenças (38). As mudanças no estilo de vida e no comportamento humano, que ocorreram no último século, provocaram um aumento significativo na incidência de DM2 ao redor do mundo. A epidemia é comumente constatada, surgindo algumas condições associadas como “diabesidade” e síndrome metabólica (1,12,39).

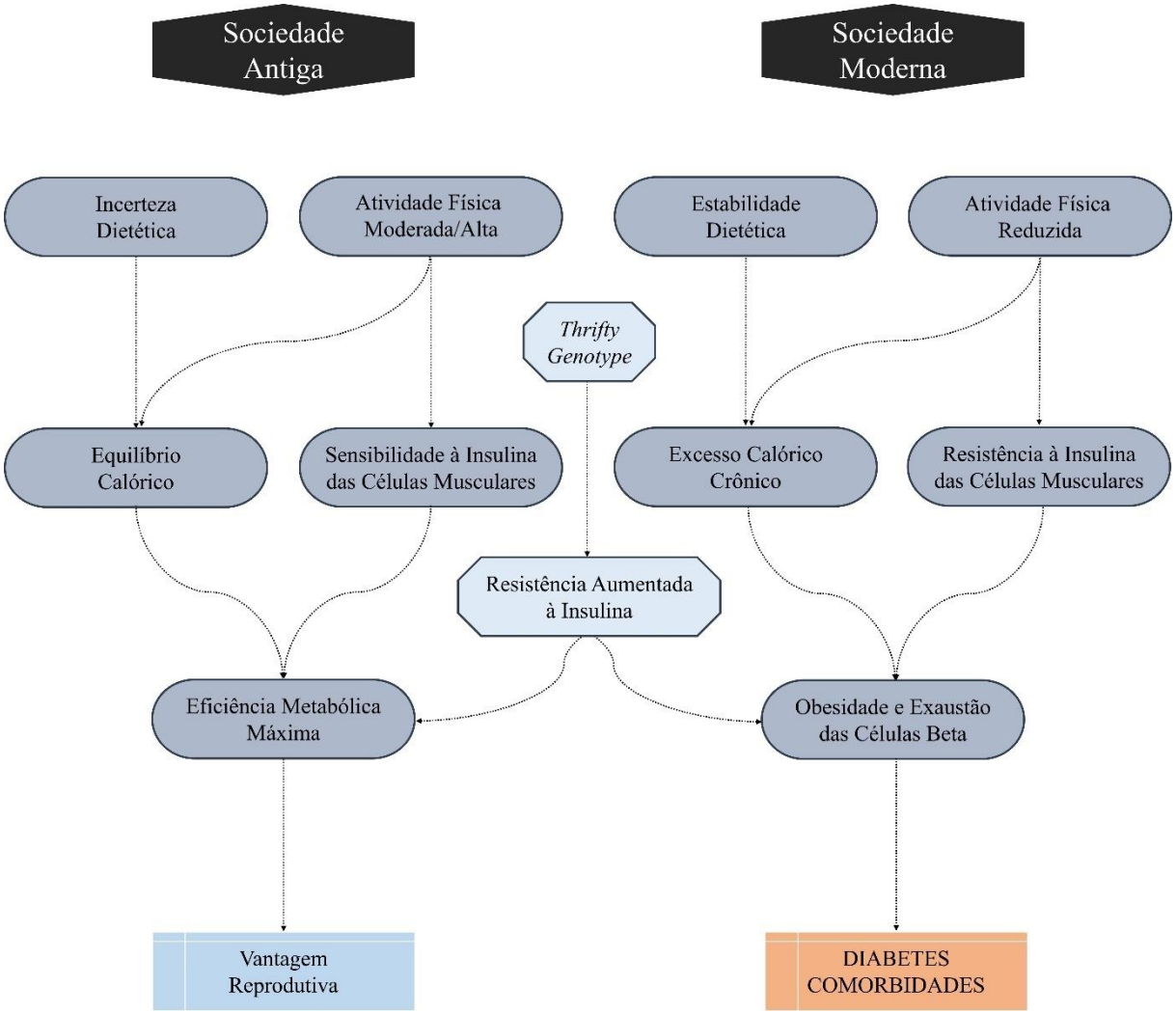

Figura 1.2 – Hipótese do *Thrifty Genotype* e sua relação ao ambiente e o estilo de vida em dois momentos distintos da sociedade. A sociedade antiga representa momentos de escassez de alimento e a sociedade moderna representa uma época em que os alimentos passaram a ser mais abundantes. Além da diferença entre a frequência de atividade física entre esses dois momentos.
Fonte: Adaptada de BINDON; BAKER (33).

1.2 Doenças Complexas e Estudos de Associação Ampla do Genoma

1.2.1 Características Genéticas

Uma característica genética pode ser considerada importante devido à sua desejabilidade entre os indivíduos ou sua vantagem concedida, enquanto existem características desejáveis, como inteligência, aptidão física, entre outras, existem também características completamente indesejáveis, como câncer e outras doenças graves. O papel do componente genético na origem ou associação dessas características pode ser investigado de várias

maneiras, incluindo estudos com gêmeos, agregação familiar (40-41) ou estudos de associação de variantes genéticas (42-44).

Uma vez confirmada a existência de um componente genético para a manifestação de um determinado fenótipo, o próximo passo é identificar os *loci* genéticos e as variantes envolvidas (42,44-46). A determinação da relação entre genótipo e fenótipo pode ser complexa devido à participação e efeitos de múltiplos genes, variantes e fatores ambientais. Além disso, a própria mensuração do fenótipo pode ser desafiadora com a presença de fenocópias, que são características que surgem quando uma resposta ao ambiente gera um fenótipo com efeitos semelhantes aos de um determinado grupo de genes, como ocorre no DM2, que pode ser controlado por dieta, estilo de vida e o uso medicamentos. Outro fator desafiador é a variabilidade da expressividade do fenótipo, que indica o quanto uma característica é manifestada entre os indivíduos, podendo ser influenciada pelo ambiente e pela interação de vários genes e é mais comum em casos poligênicos ou oligogênicos (41).

Os fatores etiológicos compreendem mecanismos genéticos, epigenéticos, eventos relacionados ao comportamento e exposição a substâncias químicas, além de fatores como idade. Apesar da existência desses diversos fatores, eles ainda não conseguem explicar a etiologia da doença em sua totalidade, havendo também alguns fatores desconhecidos (Figura 1.3) (47).

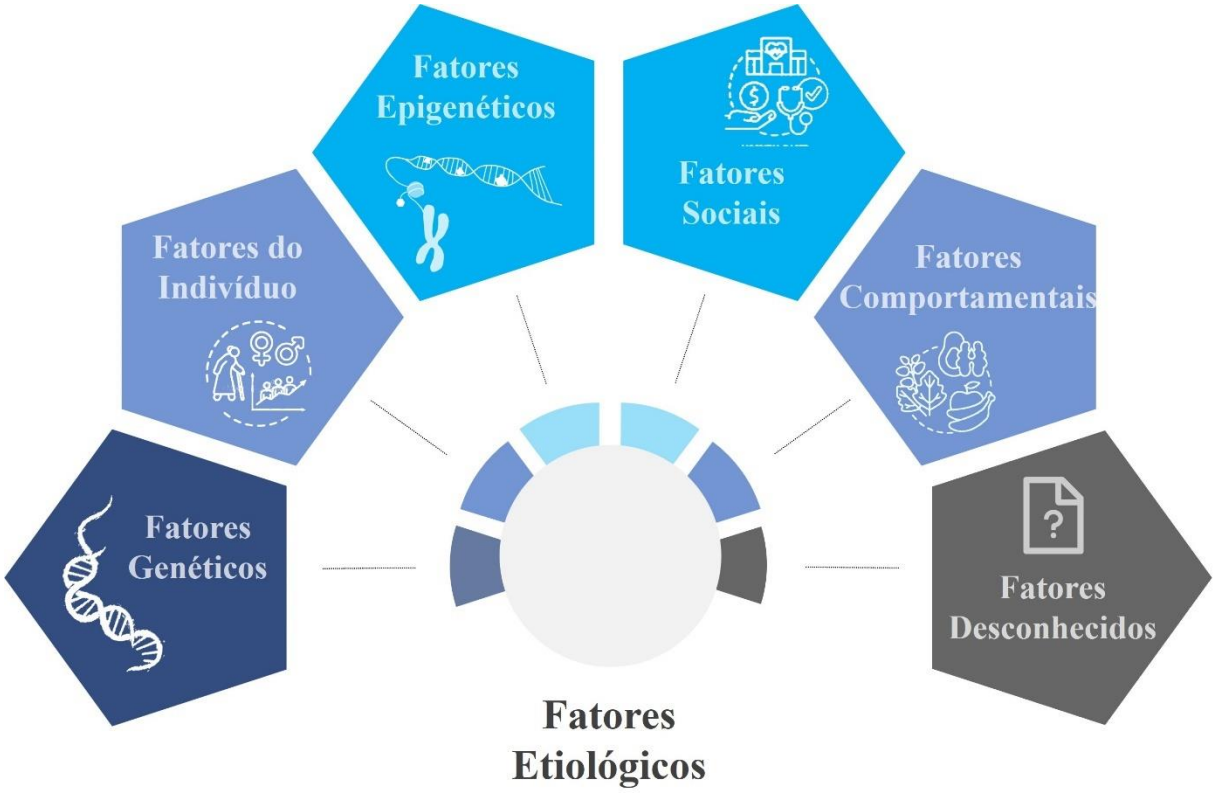

Figura 1.3 – Fatores etiológicos das doenças complexas.
Fonte: Adaptada de SCHRIML *et al*. (47).

Do ponto de vista genético, essas características podem ser divididas em três categorias: monogênicas, oligogênicas e poligênicas. No caso das características monogênicas, a expressão do fenótipo é determinada por um único gene. Nas características oligogênicas, existem alguns genes envolvidos, geralmente um gene principal acompanhado por um ou mais genes modificadores. Já nas características poligênicas, há a presença de diversas variantes em múltiplos genes que exercem, individualmente, um pequeno efeito, com interações complexas (41). Na categoria poligênica encontra-se o DM2 (2).

As suscetibilidades às doenças e respostas ao tratamento são influenciadas pela combinação única de variantes genéticas presentes em cada indivíduo. As informações obtidas por meio da análise dessas variações no genoma são utilizadas para identificar e gerenciar riscos de saúde antes do surgimento de sintomas, auxiliar no diagnóstico de distúrbios existentes, aprimorar as previsões prognósticas e orientar o tratamento (40).

A obtenção dessas informações foi possibilitada através dos GWAS, que revolucionaram as pesquisas sobre doenças complexas nos últimos 20 anos (48-49). Estes estudos permitem a identificação de variações genéticas comuns que influenciam características, doenças e

respostas clínicas a medicamentos (41-42), fornecendo uma nova visão sobre o embasamento biológico subjacente ao diagnóstico.

Embora a patogenicidade das doenças complexas possua um caráter multifatorial, os fundamentos genéticos permitem avaliar, de maneira independente, a validade das características e classificações de doenças já descritas (50). Os GWAS, a partir de análises envolvendo milhares de variantes, fornecem informações sobre a arquitetura genética de doenças complexas através da descoberta de novas associações, da identificação de *loci* de suscetibilidade e de vias biológicas que permitem a descoberta de biomarcadores (49).

1.2.2 Polimorfismos de Nucleotídeo Único

Os GWAS analisam informações ligando variantes comuns, os polimorfismos de nucleotídeo único (*Single Nucleotide Polimorphism* – SNP) ao risco de doença (Figura 1.4) (45; 46). Os SNPs são modificações de um único par de bases numa sequência genômica que ocorrem com uma alta frequência no genoma humano (42,45-46) (Figura 1.4). São frequentemente usados como marcadores de uma região do genoma, possuindo, em sua grande maioria, um impacto mínimo nos sistemas biológicos, representando a forma mais frequente de variantes (51). Um SNP possui, geralmente, dois alelos, significando que para uma determinada população, existe duas possibilidades de pares de bases que ocorrem comumente, tendo sua frequência baseada na frequência do alelo menor (*Minor Allele Frequency* – MAF) (45).

Através de um conjunto de SNPs especialmente selecionados para identificar todas as variantes genéticas comuns conhecidas no genoma, os GWAS revelam conexões entre *loci* genômicos específicos e características genéticas (51).

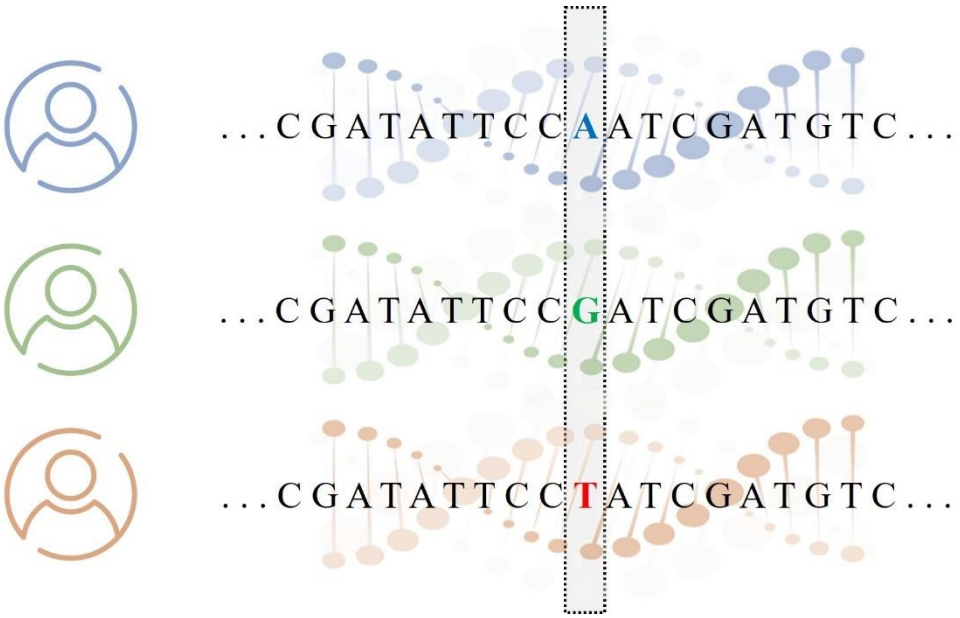

Figura 1.4 – Representação de um *Single Nucleotide Polimorphism* – SNP.
Fonte: Elaborada pelo autor.

1.2.3 Estudos de Associação e Diabetes Mellitus Tipo 2

Com os GWAS tornou-se possível identificar as variantes comuns que aumentam o risco de DM2 (52). Os primeiros estudos publicados, em 2007, identificaram 10 *loci* explorando um conjunto de SNPs candidatos (53-55).

Foi a partir da colaboração de diversos consórcios de pesquisa, que resultou em um grande banco de dados de genotipagem, que os GWAS atingiram maiores patamares de tamanho de amostra e de resultados (52). Em 2012, um estudo com uma amostra de 150.000 indivíduos aumentou o número de *loci* para 56 (56).

À medida que os GWAS foram ficando cada vez mais robustos, ampliando o número de indivíduos e variantes analisadas, eles passaram a fornecer detalhes importantes sobre a arquitetura genética do DM2 (52). Relatando que variantes associadas ao DM2 estão mais relacionadas com à diminuição na secreção da insulina do que com à sua sensibilidade (57-58), e classificando três grupos patológicos para o DM2, um relacionado com a secreção e o processamento de insulina prejudicados, outro com a resistência à insulina e um terceiro grupo relacionado com dislipidemia (59), além de gerar dados importantes para estudos futuros.

Os GWAS relataram, até 2023, mais de 700 *loci* de risco ao DM2. Esses estudos demonstram que o aumento do tamanho da amostra e a inclusão de participantes de diversas ancestralidades aumentam substancialmente o poder estatístico para identificar novos sinais.

Consequentemente, variantes de menor efeito ao risco de DM2 passaram a ser detectadas por GWAS, indicando que essas variantes podem ser estatisticamente significativas, mas sua contribuição para a compreensão da fisiopatologia do DM2 é branda (52,60). O principal banco de dados que incorpora esses estudos é o GWAS *Catalog*, que engloba mais de 45.000 estudos publicados (49).

1.3 Objetivos do Estudo

Nesta tese tivemos dois objetivos principais:

1) Investigar a relação entre o DM2 e as características das regiões genômicas onde estão localizados os SNPs associados ao fenótipo;

2) Analisar os SNPs associados ao DM2 em um conjunto de dados de uma população brasileira, buscando encontrar características específicas das variantes nesta população em relação àquelas já conhecidas para outras populações com ancestralidades diferentes.

1.3.1 Detalhamento do Primeiro Objetivo

Dentro deste objetivo tivemos vários subobjetivos, conforme descrito a seguir.

Verificar se os SNPs associados ao DM2 tendem a se localizar em regiões mais permissivas do genoma, o que poderia indicar efeitos em funções associadas a pressões evolutivas mais brandas.

Verificar como esses SNPs podem influenciar os padrões de *splicing* e a regulação da expressão dos genes envolvidos.

Analisar os SNPs, do ponto de vista, evolutivo e estrutural, com o intuito de obter parâmetros que determinem a essencialidade ou dispensabilidade dos éxons localizados em suas proximidades. Essa análise verifica se os éxons que possuem SNPs associados ao DM2, ou próximos a íntrons que compartilham essa característica, demonstram uma maior propensão a estarem ausentes ou mutados em diferentes variantes resultantes de *splicing* alternativo e em genes ortólogos.

Investigar se esses mesmos éxons apresentam maior tendência a codificar regiões desestruturadas das proteínas correspondentes.

Investigar se os genes associados ao DM2 apresentam uma menor variação em termos de valores absolutos, ranqueamento e especificidade de expressão, se comparados aos outros genes. Desta forma, buscou-se compreender se esses SNPs estão relacionados a padrões de expressão mais estáveis e específicos.

Os resultados das investigações acima descritas são apresentados no Capítulo 2.

1.3.2 Detalhamento do Segundo Objetivo

Dentro deste objetivo tivemos vários subobjetivos, conforme descrito a seguir.

Analisar os fatores genéticos relacionados ao risco de desenvolvimento do DM2 em uma população miscigenada brasileira, considerando a variação da prevalência entre grupos ancestrais, e comparando os dados dos SNPs associados ao DM2 com os extraídos em amostras populacionais com diferentes ancestralidades e etnias.

Identificar quais SNPs são mais prevalentes na população brasileira, e quais apresentam frequências padrões ou constantes em todos os grupos étnicos considerados.

Identificar, em uma coorte brasileira, SNPs potencialmente deletérios para função de proteína em genes que apresentam sinal em GWAS, genes associados ao MODY e outras formas de insulinopenia e genes de lipodistrofia e resistência à insulina.

Verificar o risco de DM2 em relação à presença de variante raras nesses mesmos grupos de genes, buscando compreender os mecanismos genéticos subjacentes ao desenvolvimento do DM2 nesse contexto.

Avaliar o risco de DM2, na mesma coorte, com o objetivo de entender como variantes genéticas comuns amplamente distribuídas nas populações estão relacionadas com a manifestação e progressão desse fenótipo.

Avaliar a capacidade discriminatória, na população brasileira, dos escores de risco poligênicos para DM2 existentes, com o objetivo de determinar o quão eficazes esses escores são na identificação de indivíduos em risco de desenvolver DM2 dentro do contexto específico da amostra.

Os resultados das investigações acima descritas são apresentados no Capítulo 3.

2 ANÁLISES EVOLUTIVAS E ESTRUTURAIS DOS GENES ASSOCIADOS AO DIABETES MELLITUS TIPO 2

2.1 Introdução

Com os avanços em genômica, após o término do Projeto Genoma Humano, no início dos anos 2000, teve-se uma evolução crescente no conhecimento de como os SNPs afetam a saúde humana, além da redução significativa nos custos para sequenciamentos de genomas inteiros, facilitando o desenvolvimento da medicina genética individualizada. Esse novo modelo clínico preconiza a informação preditiva e o conhecimento de variantes genômicas para compreender, antecipar, diagnosticar e gerenciar o tratamento de uma determinada doença (40, 61).

Os GWAS permitiram a descoberta de variantes genéticas que contribuem para características normais e patológicas, e respostas a medicamentos clínicos, porém reconhecer os alvos precisos dessas associações é um desafio que vem sendo enfrentado (41-42).

O fato de que a maioria dos SNPs detectados por GWAS estarem localizados em regiões intrônicas ou intergênicas (4,43,62), aproximadamente 75,92% para DM2 (49), sugere uma seleção negativa mais amena comparada aos SNPs em éxons. Isso provavelmente se deve a um provável efeito regulatório em relação a expressão e *splicing* dos genes associados a essas regiões, em contraste com o efeito direto na codificação de proteínas derivadas de variantes em éxons (42-43,63). Esse padrão de localização destes SNPs também pode ser explicado pelo fato de se tratar da forma de variante mais frequente no genoma humano, que serão mais associadas entre si e podendo reportar efeitos de regiões vizinhas, indicando um desequilíbrio de ligação, em que os SNPs associados são encontrados mais frequentemente juntos em uma determinada população do que o esperado caso eles segregassem independentemente (62).

Evidências indicam que SNPs em sequências codificantes e não-codificantes podem ter efeitos no processamento do RNA mensageiro, podendo ocasionar doenças ao afetar o *splicing* constitutivo ou alternativo (4,46). Deve-se notar, no entanto, que é extremamente difícil a detecção de elementos regulatórios de *splicing* baseado unicamente na sequência de DNA, de maneira que em muitas ocasiões pode-se subestimar o efeito desses SNPs relacionados ao fenômeno. Muitas das mudanças relevantes no padrão de *splicing* induzidas por SNPs podem ser sutis, como enriquecimento de uma isoforma em condições específicas (4,64).

A manifestação do DM2 está relacionada com a redução da plasticidade fenotípica e a dificuldade de manter a homeostase metabólica quando exposto a diferentes ambientes. A

identificação dos principais fatores que alteram a plasticidade fenotípica pode definir a suscetibilidade individual para o desenvolvimento do DM2, podendo ter um grande potencial preditivo (65). Há evidências de que, sob condições metabólicas adversas como obesidade, resistência à insulina, entre outras, a maquinaria de *splicing* sofre uma desregulação na maioria dos tecidos (66-67), e que essas desregulações estão associadas ao desenvolvimento de doenças (66,68-69).

Como o DM2 está associado com alterações nos padrões de *splicing*, as desregulações nessa maquinaria podem preceder, contribuir e prever o desenvolvimento do fenótipo. Um estudo, CORDIOPREV *study* (70), com 215 pacientes diagnosticados com doença cardiovascular e que não possuíam diagnóstico de DM2 no início do estudo, em que 107 casos desenvolveram DM2 ao longo de 5 anos, revelou a existência de alterações na maquinaria de *splicing* que precedem e predizem o desenvolvimento do DM2 em pacientes com doença cardiovascular, fortalecendo a hipótese de que a expressão alterada de componentes e variantes de fatores de *splicing* pode significar uma relação com a manifestação do fenótipo (71). Além disso, já foi descrito que SNPs também podem afetar os níveis de expressão gênica de um gene próximo, em cis, ou um gene distante, em trans, alterando a sua regulação (4,72).

Aproximadamente, 84,56% dos sinais associados ao DM2, relatados nos GWAS, mapeiam para sequências não codificantes (49), sendo provável que seu efeito seja devido à sua localização dentro ou próximo a regiões reguladoras que modulam a expressão dos genes nesses locais, atuando em conjunto e afetando processos fisiológicos chave, contribuindo assim para as características multifatoriais de doenças complexas com o DM2 (73). Isso dificulta os avanços para relacionar os SNPs com as transcrições e redes em que eles exercem seus efeitos. (74). Uma abordagem para enfrentar esse desafio variante para função é usar o mapeamento de expressão para caracterizar o impacto de SNPs, associados aos fenótipos, na expressão dos genes (75).

Diversos GWAS de DM2 foram realizados e permitiram a identificação de uma grande quantidade de regiões relacionadas com uma maior suscetibilidade ao fenótipo. Atualmente, há, pelo menos, 944 sinais distintos associados ao DM2, relacionados com 503 genes (49). Porém, o progresso no entendimento do mecanismo da patologia, a partir desses dados, tem sido lento (43). Uma análise ampla, analisando variantes de diferentes frequências, sugeriu que a maior contribuição genética deriva de variantes de alta frequência que individualmente possuem impactos modestos no risco ao DM2 (76), essas variantes de baixa penetrância e pequeno tamanho de efeito, reveladas nos GWAS, se combinam para conferir o risco associado (61,63).

2.2 Métodos

2.2.1 Alinhamento das Sequências de Proteínas e Regiões Exônicas

A partir dos dados de pares de genes ortólogos (conforme explicado a seguir), foi realizada a busca dos dados da sequência de proteínas correspondente aos genes, para duas espécies. As sequências foram retiradas do *Universal Protein Resource* – UniProt (https://www.uniprot.org/) (77), e foram selecionados os proteomas correspondentes as espécies *Homo sapiens* (*Entry*: UP000005640) e *Mus musculus* (*Entry*: UP000000589), as duas espécies que possuem o maior número de proteínas cadastradas no banco.

Primeiramente, para cada gene selecionado no genoma humano, foi realizada uma busca por genes homólogos no camundongo, através do InParanoiDB 9 *Protein and Domain Ortholog Groups* (https://inparanoidb.sbc.su.se/), que consiste em um banco de dados composto por grupos de ortólogos e inparálogos para 640 espécies que usa as pontuações de similaridade *pairwise* entre dois proteomas completos para construir grupos de ortologia, calculadas usando o DIAMOND, um alinhador de sequência para buscas de proteínas e DNA traduzido, projetado para análise de alto desempenho com grandes quantidades de sequências (78-79).

Um grupo de ortologia é inicialmente composto por dois ortólogos chamados de sementes, encontrados através dos melhores resultados bidirecionais entre os dois proteomas. Para refinar a análise, mais sequências são adicionadas ao grupo, e se houver sequências nos dois proteomas que estão mais próximas do ortólogo semente, esses membros de um grupo de ortologia são chamados de inparálogos. O InParanoiDB 9 fornece um valor de confiança para cada inparálogo que mostra o quanto ele está intimamente relacionado ao seu ortólogo inicial (80). Foram selecionados os inparálogos que apresentaram um maior valor de confiança e de correspondência para fazer parte de cada par de ortólogos.

As sequências de proteínas das duas espécies foram alinhadas utilizando a ferramenta *Basic Local Alignment Search Tool* – BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi) (81-82). As regiões exônicas, foram delimitadas de acordo com a posição inicial e final de cada éxon. Esses dados foram extraídos do GENCODE (https://www.gencodegenes.org/), que identifica e classifica todas as características genéticas do genoma humano e do camundongo com alta precisão, com base em evidências biológicas, promovendo informações que auxiliam na interpretação do genoma (83). Para a nossa análise, foi utilizada a *Release* 43 (GRCh38.p13) para o genoma humano e a *Release* M32 (GRCm39) para o camundongo.

2.2.2 Coeficiente de Tendência de *Indels*

Ao realizar os alinhamentos de proteínas descritos na seção anterior, foi verificada a existência de *indels* em cada sequência de gene de humanos em relação ao ortólogo em camundongos, indicando fenômenos de inserção ou deleção. No entanto, devido a função e arquitetura específica de cada proteína codificada pelo gene em questão, é esperado que certas proteínas sejam mais tolerantes ao surgimento de *gaps* do que outras. Neste sentido o número de *indels* de um éxon pode, simplesmente, refletir as características gerais do gene e não uma peculiaridade daquele éxon. Outro ponto importante é que éxons maiores possuem uma maior probabilidade de apresentar *indels* ao acaso, se comparados com éxons menores.

Para realizar uma comparação que permita verificar a tendência específica de cada éxon, considerando toda a extensão do gene, foi calculado um coeficiente de tendência de *indels* que normaliza sua quantidade por esses fatores:

$$CTI = f - F * \left(\frac{l}{L}\right)$$

CTI = Coeficiente de Tendência de *Indels*;
f = Frequência de *Indels* no Éxon;
F = Frequência de *Indels* no Gene;
l = Comprimento do Éxon;
L = Comprimento do Gene.

2.2.3 Fatores de Estrutura Secundária das Proteínas

Através da ferramenta IUPred3 (https://iupred3.elte.hu/), as regiões intrinsecamente desordenadas de cada éxon foram identificadas. Essa ferramenta atribui, para cada resíduo de aminoácido, valores entre 0 e 1 para dois tipos de escores diferentes: o IUPRED, que representa a probabilidade do resíduo está localizado em uma região desordenada da proteína, e o ACNHOR, que representa a probabilidade do resíduo está localizado em uma região de ligação desordenada (84-86). As regiões proteicas, correspondentes a cada éxon, foram delimitadas e foi calculada a mediana das probabilidades dos resíduos pertencentes em cada região.

A natureza desordenada de uma região proteica depende de diversos contextos, com algumas regiões alternando entre um estado ordenado e um estado desordenado. O IUPred3

detecta esse tipo de distúrbio dependente de contexto, em casos em que os fatores externos sejam representados pela presença de um parceiro de ligação ordenado (84,87). O método estima a energia total de interação *pairwise*, com base em uma forma quadrática na composição de aminoácidos da proteína (85,88).

2.2.4 Bancos de Dados de Expressão Gênica

Os dados de expressão em *transcripts per million* (TPM), para diversos tecidos, foram obtidos no *Genotype-Tissue Expression* – GTEx v8 (https://gtexportal.org/home/datasets) (89). A caracterização das amostras que constitui o GTEx são demostradas nas tabelas 2.1, 2.2 e 2.3, a seguir.

Tabela 2.1 – Estatísticas das amostras do GTEx v8.

GTEx v8	Tecidos	Indivíduos	Amostras
Genótipos	54	838	15.253
Total	54	948	17.382

Fonte: Elaborada pelo autor.

Tabela 2.2 – Estatísticas das ancestralidades dos indivíduos no GTEx v8.

Ancestralidade	Indivíduos	Percentual
Caucasianos	802	84,6%
Afro-americanos	122	12,9%
Asiáticos	12	1,3%
Ameríndios	2	0,2%
Outros	10	1,1%

Fonte: Elaborada pelo autor.

Tabela 2.3 – Estatísticas das faixas de idade dos indivíduos no GTEx v8.

Faixa de Idade	Indivíduos	Percentual
20-29	81	8,5%
30-39	76	8,0%
40-49	146	15,4 %
50-59	304	32,1%
60-70	341	36,0%

Fonte: Elaborada pelo autor.

2.2.5 Padrões de Tecido Especificidade

Foi calculado o padrão de tecido especificidade da expressão de cada gene, através do parâmetro TAU (90), utilizando o valor da mediana de expressão em cada tecido. Esse parâmetro possui valores que variam de 0 a 1, onde o 0 indica que o gene possui o mesmo nível de expressão em todos os tecidos avaliados, e 1 indica que o gene possui sua expressão apenas em um único tecido. O TAU foi considerado a melhor métrica para medir a tecido especificidade de um gene, onde o parâmetro foi comparado com 8 outros métodos (91).

Os genes são frequentemente caracterizados em duas classificações, uma como genes tecido específicos e outra como genes de *housekeeping*, comumente descritos como essenciais para a existência celular, independentemente de sua função específica no tecido ou organismo, e que sua expressão se mantém estável, independentemente do tipo de tecido, estágio de desenvolvimento, ciclo celular ou estímulos externos (92). Porém, algumas informações funcionais importantes estão presentes em genes com perfis de expressão de médio alcance (90).

Com esses parâmetros, podemos verificar se os genes que possuem SNPs associados ao DM2 possuem uma distribuição da sua expressão mais uniforme em todos os tecidos, ou se a sua expressão está concentrada em alguns tecidos específicos.

2.2.6 Coeficiente de Variação da Expressão

Os padrões de expressão dos genes selecionados foram caracterizados a partir do Coeficiente de Variação da Expressão (CVE) entre os indivíduos. Genes que possuem maiores valores de CVE, indicam uma maior permissividade de fenômenos que provocam instabilidades na sua expressão. Esse tipo de instabilidade é reflexo de uma heterogeneidade entre os indivíduos ou o comportamento de um gene induzível.

Primeiramente, para obter valores que refletem essa característica, foi selecionado, para cada gene, o tecido em que ele é mais expresso (maior valor de TPM). Em seguida, foi calculado o desvio padrão relativo (coeficiente de variação) da distribuição da expressão de cada um dos genes em seus respectivos tecidos.

A escolha para calcular o valor baseado somente no tecido de maior expressão, teve o objetivo de eliminar dados ruidosos ao incluir a baixa expressão dos genes em muitos tecidos.

2.2.7 Especificidade da Expressão nos Indivíduos

Foi calculado um parâmetro para avaliar a Especificidade da Expressão nos Indivíduos (EEI). Genes que apresentam maiores valores de EEI, indicam genes que são expressos em uma pequena quantidade de indivíduos de uma população, considerando o tecido em que ele é mais expresso.

Primeiramente, para obter valores que descrevam a especificidade, para cada gene, foi selecionado o tecido em que ele é mais expresso, e foi realizado o cálculo semelhante ao parâmetro TAU, em que, dessa vez, o 0 indica que o gene possui o mesmo nível de expressão em todas as amostras, em diferentes indivíduos, e 1 indica que o gene possui expressão apenas em uma única amostra, um único indivíduo.

Esse parâmetro indica se, para o tecido em que o gene é mais expresso, a distribuição da sua expressão é mais uniforme em relação aos indivíduos representados na amostra, ou se essa distribuição está concentrada em indivíduos com características específicas. Valores baixos de EEI indicam genes com uma expressão mais homogênea em uma determinada amostra populacional.

2.2.8 Correlação da Expressão das Isoformas

Foi verificada a Correlação da Expressão das Isoformas (CEI) de cada gene selecionado. Os dados de expressão dos transcritos foram retirados no GTEx. Maiores valores de CEI indicam uma maior coordenação da expressão das isoformas entre si, determinando, possivelmente, padrões de *splicing* alternativo dos genes.

Inicialmente, em cada gene, foram selecionadas todas as isoformas que possuem mediana geral de expressão acima de 2 TPM, considerando todos os tecidos analisados. Em seguida, foi identificado o tecido em que o gene é mais expresso e, para cada par de isoformas, com seus dados de expressão nos indivíduos, foi calculado o coeficiente de correlação de Spearman (ρ), uma medida não-paramétrica da dependência do ranqueamento das variáveis analisadas, e que possui considerada robustez em casos com *outliers* (93), uma característica bem comum ao analisar a expressão das isoformas, pois grande parte dos genes possuem uma isoforma de referência com uma expressão mais frequente.

Essa análise permite avaliar se, para os pares de isoformas consideradas, há uma alta coordenação das suas expressões ($\rho > 0{,}70$), coordenação moderada ($0{,}30 \leq \rho \leq 0{,}70$), ou se a expressão das isoformas não possuem um padrão de coordenação ($\rho < 0{,}30$) (93).

Foi calculada a média de todos os pares de isoformas selecionadas, e utilizou-se o módulo desse resultado para avaliar o nível de coordenação da expressão das isoformas. Menores valores de CEI indicam genes que sofrem menos *splicing* alternativo.

2.3 Resultados

2.3.1 Análise Evolutiva e Estrutural

Para cada análise realizada, os éxons dos genes selecionados foram classificados em duas amostras, uma contendo os éxons adjacentes, e a outra contendo os éxons não-adjacentes, chamados de éxons distantes.

Para essa classificação, SNPs intrônicos são flanqueados por dois éxons adjacentes, enquanto SNPs exônicos apresentam o seu próprio éxon como adjacente. Todos os éxons que não possuem SNP em sua extensão, ou não estão flanqueando algum SNP associado ao DM2, são classificados como éxons distantes. Nesta análise, foi verificado se os éxons adjacentes ao SNPs associados ao DM2 possuem uma maior tendência em sofrer fenômenos de inserção ou deleção, além de avaliar o padrão de estruturação dessas regiões, com o intuito de verificar se os éxons adjacentes estão, preferencialmente, localizados em regiões desordenadas das proteínas correspondentes (Figura 2.1).

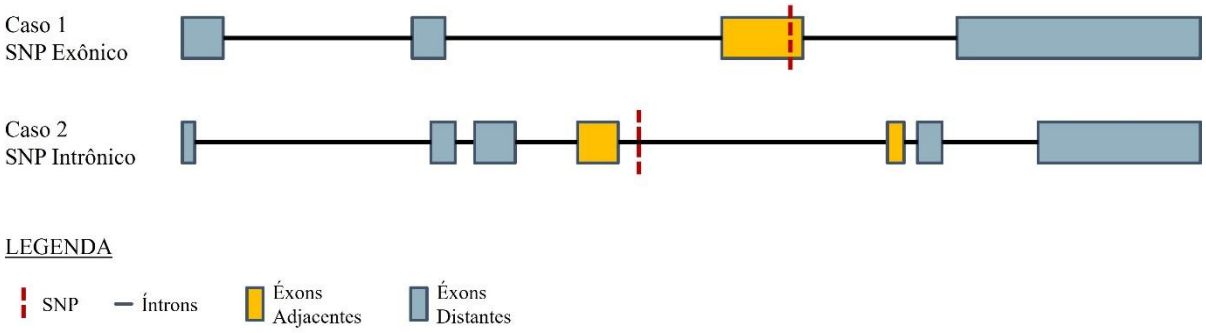

Figura 2.1 – Classificação dos éxons em adjacentes e distantes.
Fonte: Elaborada pelo autor.

2.3.1.1 Seleção dos Genes Associados ao Diabetes Mellitus Tipo 2

Os genes selecionados como associados ao DM2 foram pesquisados no GWAS *Catalog* (https://www.ebi.ac.uk/gwas/), que, em resposta ao rápido aumento de GWAS, fornece um banco de dados consistente e disponível gratuitamente sobre a associação de SNPs e os respectivos fenótipos (49). Até a última atualização nos dados iniciais, dia 31/03/2023, o GWAS *Catalog* contava com 8.724 estudos publicados, totalizando 368.980 sinais reportados.

Inicialmente, foram selecionados 3.772 sinais reportados ao DM2, representando 2.522 variantes únicas. Por se tratar de GWAS compreendidos em diversos estudos diferentes, alguns SNPs são reportados em diferentes estudos catalogados. Desses SNPs, foram selecionados 944 SNPs que possuíam um *odds ratio* (OR) reportado no estudo original, um dado essencial para o desenvolvimento das análises. A partir disso, foram selecionados 503 genes que apresentam SNPs associados ao DM2 em sua extensão, e, posteriormente, apenas os genes que apresentavam SNPs localizados em íntrons e éxons, representando uma amostra inicial de 276 genes (Figura 2.2).

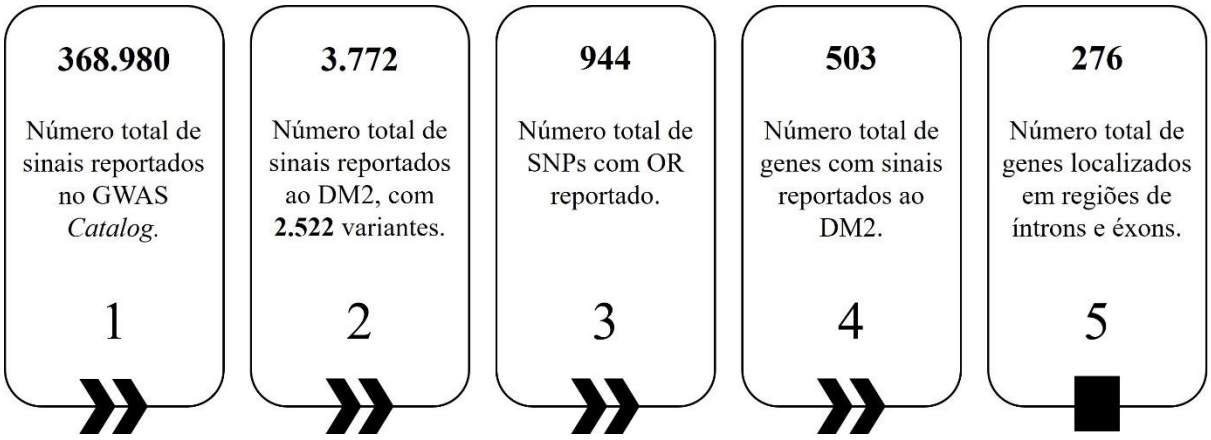

Figura 2.2 – Seleção dos genes associados ao DM2.
Fonte: Elaborada pelo autor.

2.3.1.2 Seleção dos Dados das Proteínas

Com o objetivo de verificar uma possível influência dos SNPs associados ao DM2 com o fenômeno de *splicing* alternativo, foram analisadas as estruturas genômicas localizadas nas regiões onde os SNPs estão presentes. O padrão de essencialidade de cada éxon foi verificado para quatro fatores, um deles relacionado a tendência de surgimento de *gaps* na região, e outros

três relacionados com o padrão de estruturação da proteína codificada pelos genes selecionados na seção 2.3.1.1.

Para observar a tendência de surgimento de *gaps* nas regiões determinadas, as sequências do genoma e das proteínas correspondentes aos genes foram alinhadas. Preliminarmente, foram selecionadas as espécies *Pan troglodytes, Mus musculus, Gallus gallus, Xenopus tropicalis* e *Danio rerio*. Ao definir os critérios utilizados sobre os parâmetros de qualidade dos sequenciamentos dessas espécies e a distância evolutiva de cada uma delas em relação ao *Homo sapiens* (humano), a espécie *Mus musculus* (camundongo) foi escolhida como a mais consistente para as análises.

2.3.1.3 Essencialidade dos Éxons Adjacentes

A amostra partiu de 944 SNPs com OR reportado, distribuídos em 276 genes com SNPs intrônicos e exônicos. Foram analisados um total de 5779 éxons, 888 adjacentes e 4291 distantes. Foram desconsiderados dessa análise 600 éxons classificados como iniciais e finais.

Seguindo a metodologia de cada parâmetro, e devido a característica de cada análise, em alguns casos, os éxons não apresentaram valores do parâmetro correspondente, sendo descartados da amostra final. A composição das amostras de éxons adjacentes e distantes, para cada tipo de análise, pode ser visualizada na Tabela 2.4, a seguir.

Tabela 2.4 – Composição das amostras dos éxons adjacentes e distantes das análises realizadas.

Classificação da Análise	**Adjacentes**	**Distantes**
Coeficiente Tendência de *Indels* (CTI)	888	4291
Escores de Regiões Desordenadas	622	4049

Fonte: Elaborada pelo autor.

Foi calculado o CTI através da diferença entre a quantidade real de *indels* encontrados no alinhamento das sequências correspondentes ao éxon e a quantidade esperada ao acaso, considerando o número total de *indels* no gene, como descrito na seção 2.2.2. Valores positivos desse parâmetro indicam que o éxon apresenta mais *indels* que o esperado ao acaso, enquanto valores negativos indicam que o número de *indels* no éxon é abaixo do esperado.

A distribuição dos valores de CTI para a amostra de éxons adjacentes apresenta mediana de -0,1542, enquanto os éxons distantes, uma mediana de -0,1663 (Tabela 2.5). Esses valores indicam que, tanto para as amostras de éxons adjacentes como as de éxons distantes, a tendência

de ocorrer fenômenos de inserção ou deleção é abaixo do esperado, considerando a metodologia utilizada nessa análise.

Os éxons adjacentes apresentam valores para os coeficientes levemente maiores que os éxons distantes, mas um p-valor de 0,1852 indica que não houve diferença estatisticamente significativa entre as duas amostras.

Foi realizada uma análise de distribuição desses valores utilizando o *Kernel Density Estimator* (KDE), que consiste em um método não paramétrico que utiliza estimativas de funções de probabilidade, alocando os dados de acordo com a densidade em cada local (94) (Gráfico 2.1).

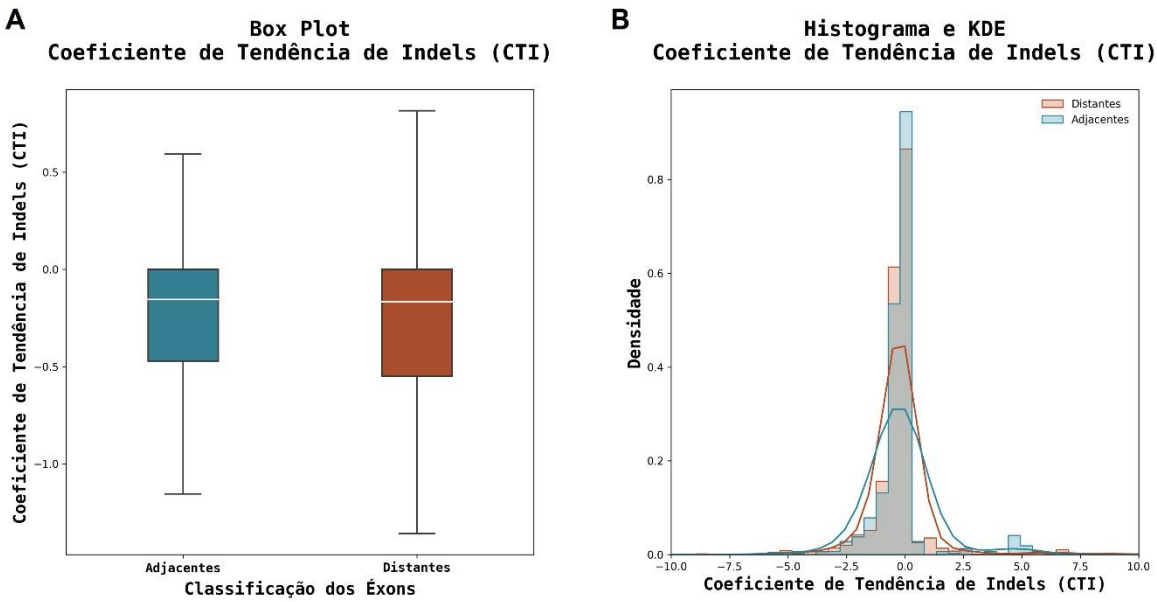

Gráfico 2.1 – Distribuição dos Coeficientes de Tendência de *Indels* dos éxons dos genes que apresentam SNPs associados ao DM2. Os éxons estão divididos em duas amostras: éxons adjacentes, em azul, e éxons distantes, em laranja. **A)** Distribuição dos parâmetros dos coeficientes. Mann-Whitney pValor = 0,1852. **B)** Histograma e KDE com a densidade da distribuição dos dois grupos de éxons.
Fonte: Elaborado pelo autor.

Para a análise de regiões desordenadas das proteínas nos éxons adjacentes e distantes, foram avaliados dois escores, IUPRED e ANCHOR, que foram calculados através do programa IUPred3, como descrito na seção 2.2.3. Valores maiores desses parâmetros indicam uma maior probabilidade do éxon estar localizado em segmentos que codificam regiões desordenadas da proteína, em que os impactos da perda desse éxon, em uma possível forma de *splicing* alternativo, são menores.

O grupo de éxons adjacentes apresenta maiores valores dos escores de regiões desordenadas, se comparado com os éxons distantes. Os escores calculados através do IuPred3

apresentam distribuições com uma diferença estatisticamente significativa, com p-valor de 8,7124 e-10 para o IUPRED, e p-valor de 1,5835 e-10 para o ANCHOR (Gráfico 2.2). A mediana para o IUPRED foi de 0,3267 para os éxons adjacentes e 0,2565 para os éxons distantes, enquanto o ANCHOR apresentou um valor de mediana de 0,4035 para os éxons adjacentes e 0,3603 para os éxons distantes (Tabela 2.5).

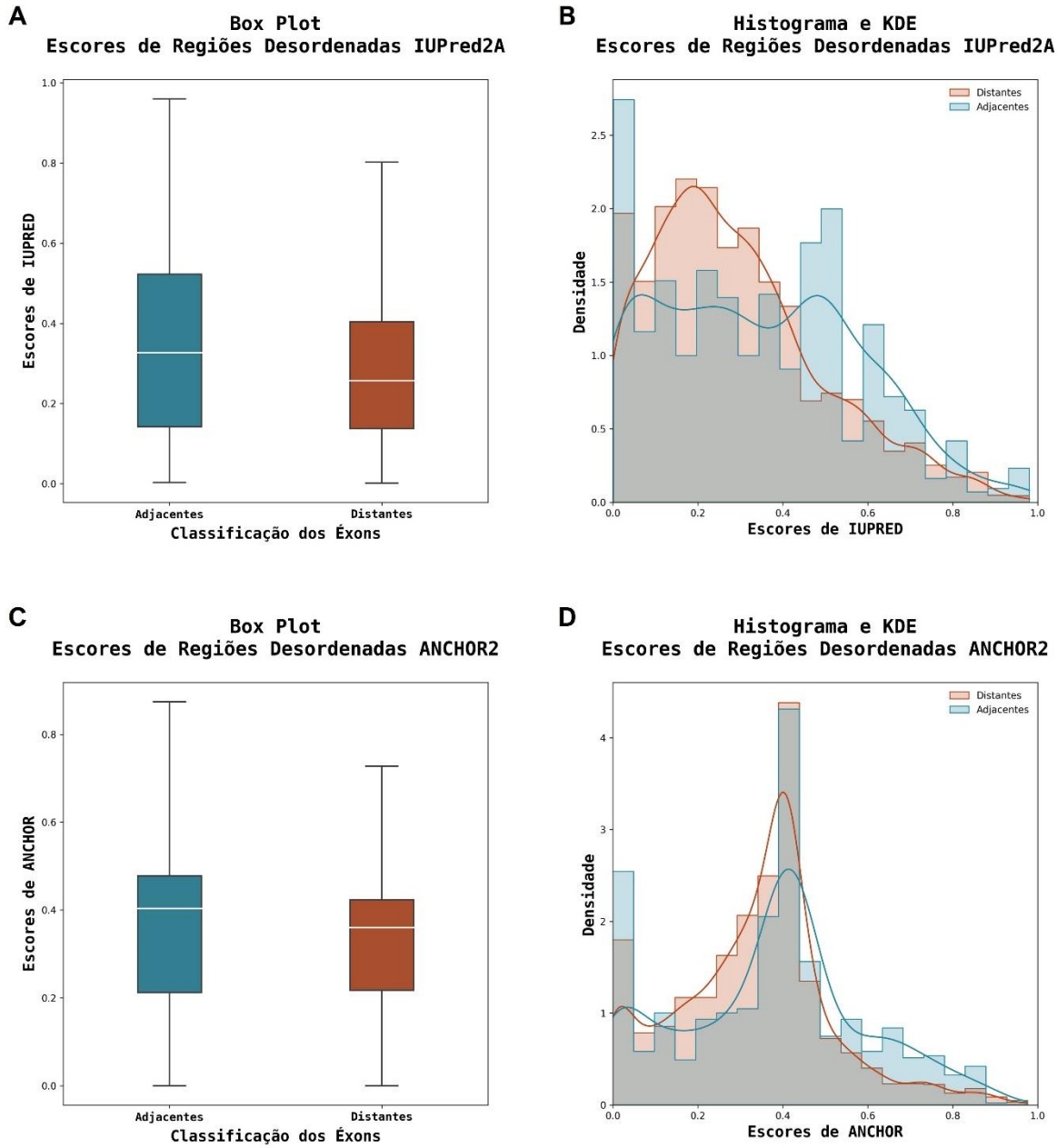

Gráfico 2.2 – Distribuição dos escores de regiões desordenadas das proteínas codificadas pelos éxons dos genes que apresentam SNPs associados ao DM2. Os éxons estão divididos em duas amostras: éxons adjacentes, em azul, e éxons distantes, em laranja. Distribuição dos escores de **A)** IUPRED, Mann-Whitney pValor = 8,7124 e-10; e C) ANCHOR. Mann-Whitney pValor = 1,5835 e-10. Histograma e KDE com a densidade da distribuição dos dois grupos de éxons, para o **C)** IUPRED e **D)** ANCHOR.

Fonte: Elaborado pelo autor.

Tabela 2.5 – Estatística dos parâmetros analisados. Medianas da distribuição dos parâmetros e Mann-Whitney P-valor.

Classificação dos Éxons	Adjacentes	Distantes	p-Valor
Coeficiente de Tendência de *Indels*	-0,1542	-0,1663	0,18
Escore de IUPRED	0,3267	0,2565	8,71 e-10
Escore de ANCHOR	0,4035	0,3603	1,58 e-10

Fonte: Elaborada pelo autor.

Esses resultados, principalmente os escores de regiões desordenadas, mostram que os éxons adjacentes aos SNPs associados ao DM2 possuem uma probabilidade significativamente maior de estarem localizados em regiões mais permissivas desses genes, que codificam regiões intrinsicamente desordenadas das proteínas.

2.3.2 Análise de Expressão Gênica

Foi realizado um estudo sistemático dos SNPs associados ao DM2, comparando com os SNPs associados a outras variações fenotípicas, analisando o padrão de expressão dos genes selecionados.

Primeiramente, os diferentes genes foram caracterizados com os padrões de alterações fenotípicas e a variação dos padrões de expressão entre os indivíduos. Genes que possuem maiores padrões de variação entre os indivíduos, em princípio, podem indicar uma maior tolerância a variações da expressão. Este tipo de variação poderia ser reflexo de uma heterogeneidade entre os indivíduos ou refletir o fato de que se trata de um gene induzível. As correlações entre o genótipo e os níveis de expressão gênica específicos do tecido podem identificar regiões do genoma que tem maior influência no processo de expressão gênica.

Com isso, foi verificado se os genes que possuem SNPs associados ao DM2 possuem uma maior ou menor permissividade a variações nos perfis de expressão, se comparados com genes associados aos outros fenótipos descritos nos GWAS, apresentados na 2.3.2.2, a seguir.

2.3.2.1 Seleção dos Tecidos das Amostras

Dos 54 tecidos encontrados no GTEx, foram excluídos os tecidos que possuem menos de 50 (n < 50) amostras documentadas, os tecidos com os códigos *'Bladder'* [21], *'Cervix – Endocervix'* [10], *'Cervix – Ectocervix'* [9], *'Fallopian Tube'* [8] e *'Kidney – Medulla'* [4].

Outros tecidos foram excluídos dessa análise devido ao comportamento bimodal da expressão desses tecidos nos genes. Esse comportamento foi identificado através do cálculo do coeficiente de correlação de Spearman da expressão de 3 amostras de 1000 genes, selecionados aleatoriamente, para cada par de indivíduos. Em seguida, foi encontrada a mediana correspondente.

Esse comportamento representa tecidos que possuem uma heterogeneidade tecidual, que pode depender de qual região do corpo ou órgão a amostra foi retirada, ou representam tecidos em que a expressão dos seus genes está relacionada com características específicas do indivíduo, como sexo, idade ou causa da morte. Foram excluídos os tecidos com códigos *'Colon – Transverse', 'Small Intestine - Terminal Ileum', 'Stomach', 'Vagina', 'Whole Blood'*.

Utilizando esses critérios, foram selecionados um total de 44 tecidos para essa análise (Figura 2.3) (Tabela 2.6).

Tabela 2.6 – Lista dos tecidos selecionados no GTEx com suas respectivas quantidade de amostras.

Tecido	**Amostras**	**Tecido**	**Amostras**
Muscle – Skeletal	706	*Spleen*	227
Skin - Sun Exposed (Lower leg)	605	*Prostate*	221
Artery – Tibial	584	*Artery - Coronary*	213
Adipose - Subcutaneous	581	*Brain - Cerebellum*	209
Thyroid	574	*Liver*	208
Nerve – Tibial	532	*Brain - Cortex*	205
Skin - Not Sun Exposed (Suprapubic)	517	*Brain - Nucleus accumbens (basal ganglia)*	202
Lung	515	*Brain - Caudate (basal ganglia)*	194
Esophagus - Mucosa	497	*Brain - Cerebellar Hemisphere*	175
Cells - Cultured fibroblasts	483	*Brain - Frontal Cortex (BA9)*	175
Adipose - Visceral (Omentum)	469	*Brain - Hypothalamus*	170
Esophagus - Muscularis	465	*Brain - Putamen (basal ganglia)*	170
Breast - Mammary Tissue	396	*Ovary*	167
Artery – Aorta	387	*Brain - Hippocampus*	165
Heart - Left Ventricle	386	*Brain - Anterior cingulate cortex (BA24)*	147
Heart - Atrial Appendage	372	*Cells - EBV-transformed lymphocytes*	147
Esophagus - Gastroesophageal Junction	330	*Minor Salivary Gland*	144
Testis	322	*Brain - Amygdala*	129
Colon – Sigmoid	318	*Uterus*	129
Pancreas	305	*Brain - Spinal cord (cervical c-1)*	126
Pituitary	237	*Brain - Substantia nigra*	114
Adrenal Gland	233	*Kidney - Cortex*	73

Fonte: Elaborada pelo autor.

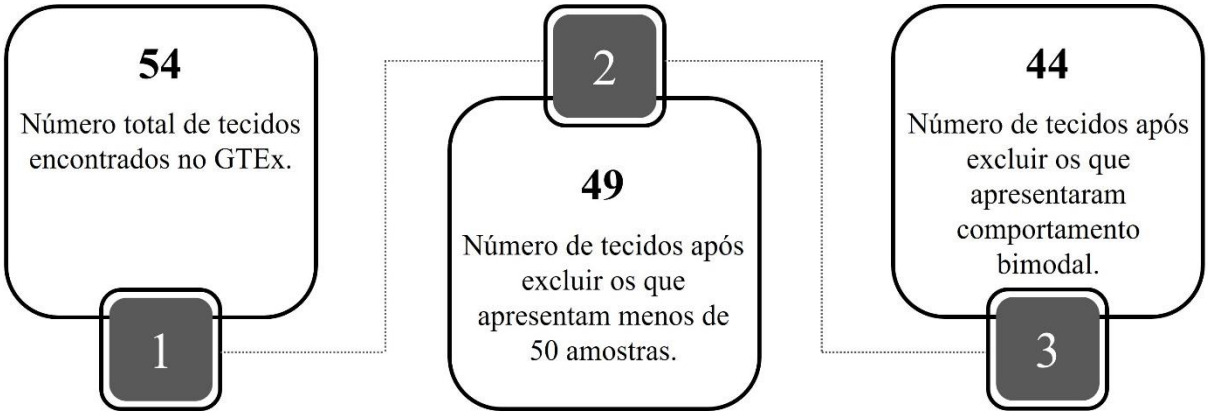

Figura 2.3 – Seleção dos tecidos do GTEx.
Fonte: Elaborada pelo autor.

2.3.2.2 Seleção dos Genes Associados ao Diabetes Mellitus Tipo 2 e Outros Fenótipos

Para realizar as análises descritas nesta seção, foram selecionadas três listas de genes associados. Uma lista corresponde aos genes associados ao DM2 (GWAS_DM2), outra lista compreende os genes associados a outras doenças multifatoriais (GWAS_MULTIFATORIAL), enquanto a última lista contém os genes associados a todos os fenótipos reportados no GWAS *Catalog* (GWAS_TOTAL).

Para a lista GWAS_DM2, foram selecionados 3772 sinais associados ao DM2, distribuídos em 1250 SNPs, contabilizados em 404 genes (Tabela 2.8).

Na segunda lista, GWAS_MULTIFATORIAL, foram identificados 6.198 sinais, e slecionads 1.101 SNPs que possuíam um *odds ratio* (OR) reportado no estudo original. Foram selecionados 365 genes (Tabela 2.8) presentes nos seguintes fenótipos: *'Amyotrophic lateral sclerosis'*, *'Asthma'*, *'Autism'*, *'Autism spectrum disorder'*, *'Bipolar disorder'*, *'Depression'*, *'Hypertension'*, *'Multiple sclerosis'*, *'Obesity'*, *'Schizophrenia'*, *'Epilepsy'*, *'Hypothyroidism'* (Tabela 2.7).

Tabela 2.7 – Quantidade de SNPs com OR reportado para cada fenótipo multifatorial selecionado.

Fenótipo	SNPs
Asthma	419
Bipolar disorder	185
Schizophrenia	144
Depression	142
Multiple sclerosis	53
Hypertension	39
Amyotrophic lateral sclerosis	36
Autism spectrum disorder	36
Hypothyroidism	20
Autism	9
Obesity	9
Epilepsy	9

Fonte: Elaborada pelo autor.

A outra lista descrita, GWAS_TOTAL, compreende todos os genes que possuem SNPs associados a todos os fenótipos reportados. Foram identificadas 368.980 sinais, em que 23.045 SNPs possuíam um *odds ratio* (OR) reportado no estudo original, totalizando 4.180 genes (Tabela 2.8).

Tabela 2.8 – Quantidade de genes selecionados para cada grupo.

Grupo de Genes	Sinais Reportados	OR Reportado	Genes
GWAS_DM2	3772	1250	404
GWAS_MULTIFATORIAL	5198	1101	365
GWAS_TOTAL	368980	23045	4180

Fonte: Elaborada pelo autor.

2.3.2.3 Padrões de Expressão dos Genes

Nesta análise, foram avaliados os parâmetros descritos nas seções 2.2.5, 2.2.6, 2.2.7 e 2.2.8.

Ao verificar a tecido especificidade e classificar os genes, para os três grupos, em alta especificidade (TAU > 0,8) e baixa tecido especificidade (TAU < 0,8), foi observado um enriquecimento dos genes em regiões de alta especificidade, para todos os grupos.

O grupo GWAS_DM2 apresenta 51,24% dos genes em regiões altas de TAU, enquanto GWAS_MULTIFATORIAL e GWAS_TOTAL apresentam uma proporção de 54,79% e 56,53%, respectivamente (Tabela 2.9). Há uma diferença significativa entre as proporções dos

grupos GWAS_DM2 e GWAS_TOTAL (p-valor = 0,0349), indicando que os genes com SNPs associados ao DM2 possuem uma menor proporção de genes em regiões de alta tecido especificidade.

Tabela 2.9 – Quantidade de genes selecionados para cada grupo.

Grupo de Genes	**Alta Especificidade**	**Baixa Especificidade**	**Total de Genes**
GWAS_DM2	207	197	404
GWAS_MULTIFATORIAL	200	165	365
GWAS_TOTAL	2363	1817	4180

Fonte: Elaborada pelo autor.

As distribuições dos parâmetros avaliados para os três grupos de genes indicam que o GWAS_DM2 apresenta valores menores de mediana para todos os parâmetros, exceto para o CEI (Gráfico 2.3) (Tabela 2.10). Através de um teste de Mann-Whitney, foi encontrado diferenças estatisticamente significativa entre a distribuição dos GENES_DM2 e GENES_TOTAL para todos os parâmetros, exceto, novamente, para o CEI (Tabela 2.11).

Esses resultados indicam que os genes que possuem SNPs associados ao DM2 estão concentrados em regiões de menor tecido especificidade, menor variação da expressão e menor especificidade da expressão entre os indivíduos, características indicativas de genes de *housekeeping*.

Tabela 2.10 – Mediana da distribuição de cada parâmetro para os três grupos de genes.

Grupo de Genes	Mediana (TAU)	Mediana (CVE)	Mediana (EEI)	Mediana (CEI)
GWAS_DM2	**0,8095**	**0,3676**	**0,5939**	0,4919
GWAS_MULTIFATORIAL	0,8260	0,3941	0,5995	0,5094
GWAS_TOTAL	0,8318	0,4030	0,6111	**0,4742**

Fonte: Elaborada pelo autor.

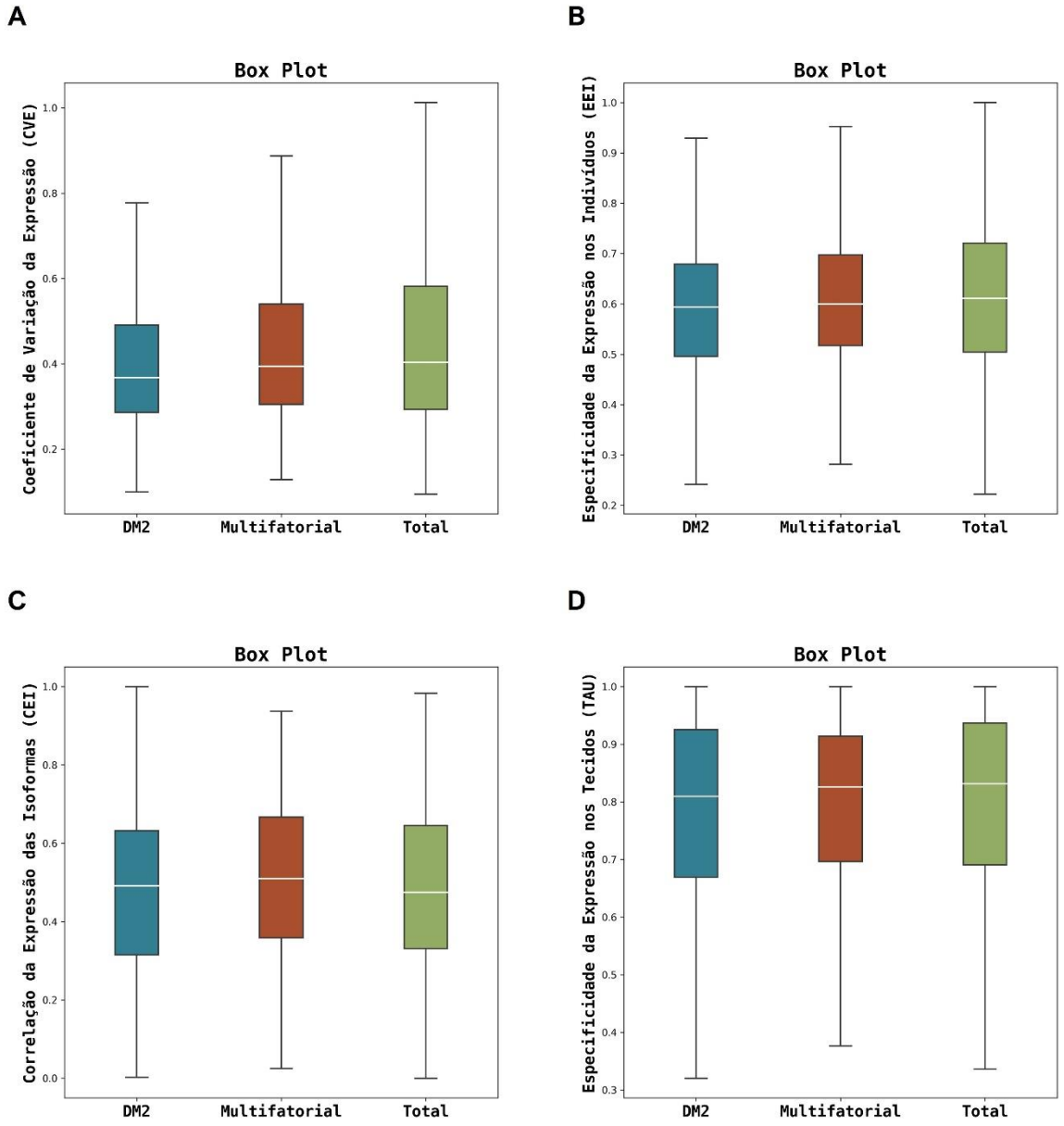

Gráfico 2.3 – Distribuição de cada parâmetro analisado para os três grupos de genes. **A)** Coeficiente de Variação da Expressão (CVE), **B)** Especificidade da Expressão nos Indivíduos (EEI), **C)** Correlação da Expressão das Isoformas (CEI) e **D)** Especificidade da Expressão nos Tecidos (TAU).
Fonte: Elaborado pelo autor.

Tabela 2.11 – Diferença estatística (p-valor) de cada parâmetro para as distribuições dos genes GWAS_DM2 em relação aos outros dois grupos de genes analisados.

	GWAS_DM2			
Grupo de Genes	**P-Valor (TAU)**	**P-Valor (CVE)**	**P-Valor (EEI)**	**P-Valor (CEI)**
GWAS_MULTIFATORIAL	0,4025	0,0281	0,1310	0,2279
GWAS_TOTAL	**0,0199**	**0,0009**	**0,0045**	0,9421

Fonte: Elaborada pelo autor.

2.3.2.4 Coeficiente de Variação da Expressão

Foi verificada uma correlação positiva moderada ($0,30 \leq \rho \leq 0,70$) entre os parâmetros CVE e TAU (Tabela 2.12). O padrão de dispersão e correlação é bem semelhante, principalmente ao analisar os grupos GWAS_DM2 (Gráfico 2.4A) e GWAS_TOTAL (Gráfico 2.4C). Ao avaliar o grupo GWAS_MULTIFATORIAL (Gráfico 2.4B) essa correlação cai um pouco, ainda sendo considerada moderada. Há uma linha de tendência sugestiva de um aumento menos acentuado do CVE, para o grupo GWAS_DM2 (Gráfico 2.4A), em regiões de baixa especificidade, com valores mais baixos de TAU, indicando que genes mais uniformemente expressos (genes de *housekeeping*) sejam menos sensíveis as variações da expressão. Esse comportamento é confirmado ao comparar o valores de ρ para cada classificação (Tabela 2.12).

Tabela 2.12 – Coeficiente de correlação de Spearman (ρ) do CVE e TAU para os três grupos de genes em geral e considerando apenas as regiões de baixa (TAU < 0,8) e alta (TAU > 0,8) tecido especificidade.

	CVE x TAU					
Grupo de Genes	**ρ Geral**	**P-Valor**	**ρ Baixa**	**P-Valor**	**ρ Alta**	**P-Valor**
GWAS_DM2	0,5799	1,12 e-37	0,2961	2,39 e-05	**0,3444**	3,74 e-07
GWAS_MULTIFATORIAL	0,4883	2,84 e-23	**0,4707**	1,77 e-10	0,1665	0,01
GWAS_TOTAL	0,5899	0,000	**0,3897**	5,50 e-67	0,2715	3,39 e-41

Fonte: Elaborada pelo autor.

Foi realizada, também, uma análise de distribuição desses valores utilizando o KDE, em sua versão para duas variáveis. Nos KDEs correspondentes aos grupos GWAS_MULTIFATORIAL (Gráfico 2.5B) e GWAS_TOTAL (Gráfico 2.5C), foi verificado um pico de densidade de genes localizados em região de alta especificidade (TAU > 0,8). No KDE que representa o grupo GWAS_DM2 (Gráfico 2.5A), foram observados dois picos, um mais evidente, na região de alta especificidade, assim como nos outros grupos, e outro mais sutil, na região de baixa especificidade (TAU < 0,8).

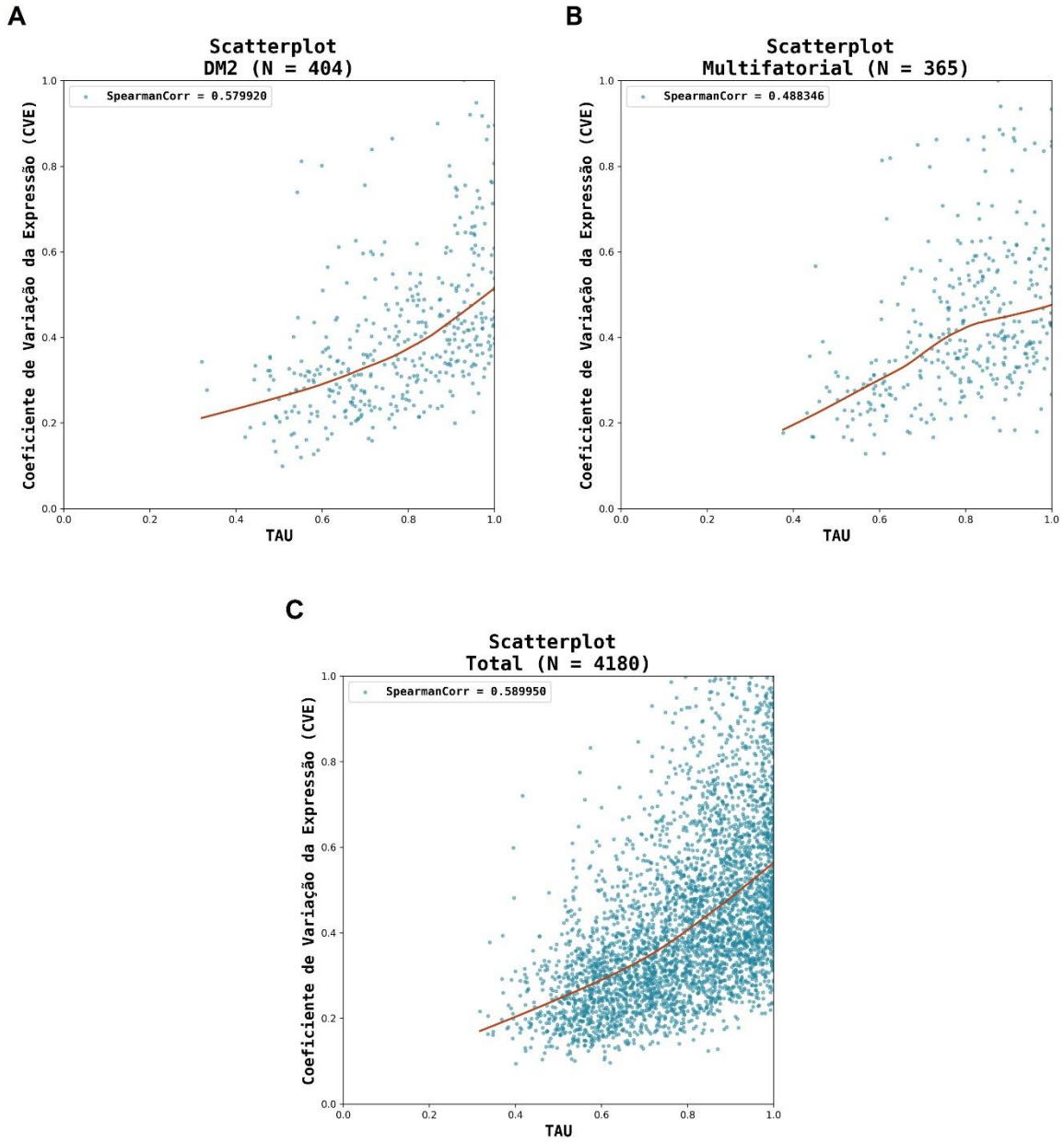

Gráfico 2.4 – Dispersão do Coeficiente de Variação da Expressão (CVE) por Especificidade da Expressão nos Tecidos (TAU). Cada ponto representa um gene selecionado nos grupos específicos. Para todos os grupos de genes é possível identificar uma correlação positiva e moderada entre os dois parâmetros. **A)** GENES_DM2, ρ = 0,5799. **B)** GENES_MULTIFATORIAL, ρ = 0,4883. **C)** GENES_TOTAL, ρ = 0,5899.

Fonte: Elaborado pelo autor.

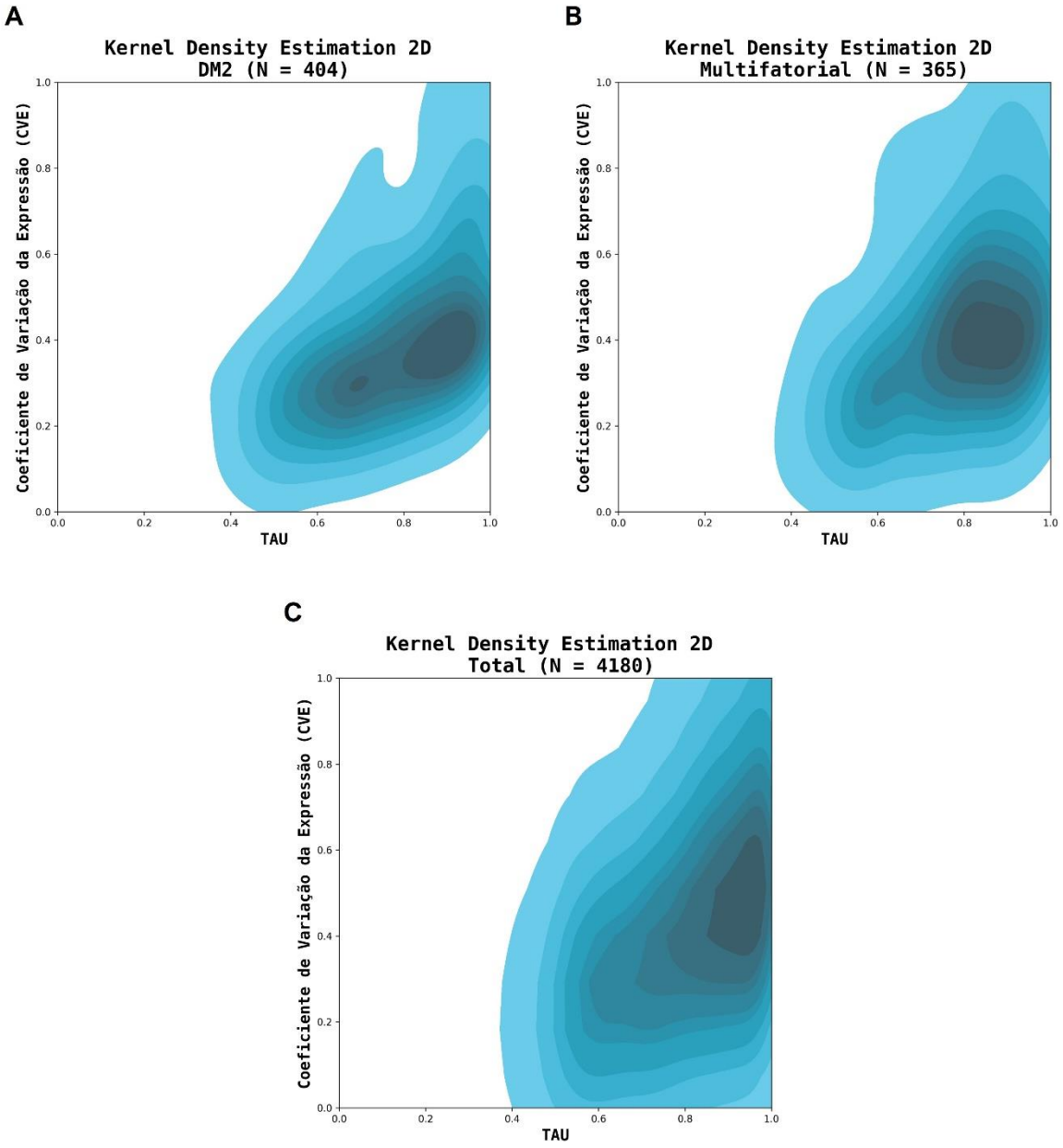

Gráfico 2.5 – KDE do Coeficiente de Variação da Expressão (CVE) por Especificidade da Expressão nos Tecidos (TAU). Para todos os grupos de genes, **A)** GWAS_DM2, **B)** GWAS_MULTIFATORIAL e **C)** GWAS_TOTAL.
Fonte: Elaborado pelo autor.

2.3.2.5 Especificidade da Expressão nos Indivíduos

Foi verificada, também, uma correlação positiva moderada ($0{,}30 \leq \rho \leq 0{,}70$) entre os parâmetros EEI e TAU (Tabela 2.13). Assim como para o CVE, o padrão de dispersão e correlação são mais semelhantes entre os grupos GWAS_DM2 (Gráfico 2.6A) e GWAS_TOTAL (Gráfico 2.6C). Essa semelhança diminui um pouco ao verificar o grupo GWAS_MULTIFATORIAL (Gráfico 2.6B). Para o EEI, a correlação é um pouco mais

acentuada em regiões de baixa especificidade, com valores mais baixos de TAU, para todos os grupos de genes, indicando que genes mais uniformemente expressos (genes de *housekeeping*) entre os tecidos, também apresentam um comportamento de expressão mais uniforme entre os indivíduos.

Nos KDEs correspondentes, foi verificado um pico de densidade de genes localizados em região de alta especificidade (TAU > 0,8) para todos os grupos (Gráfico 2.7).

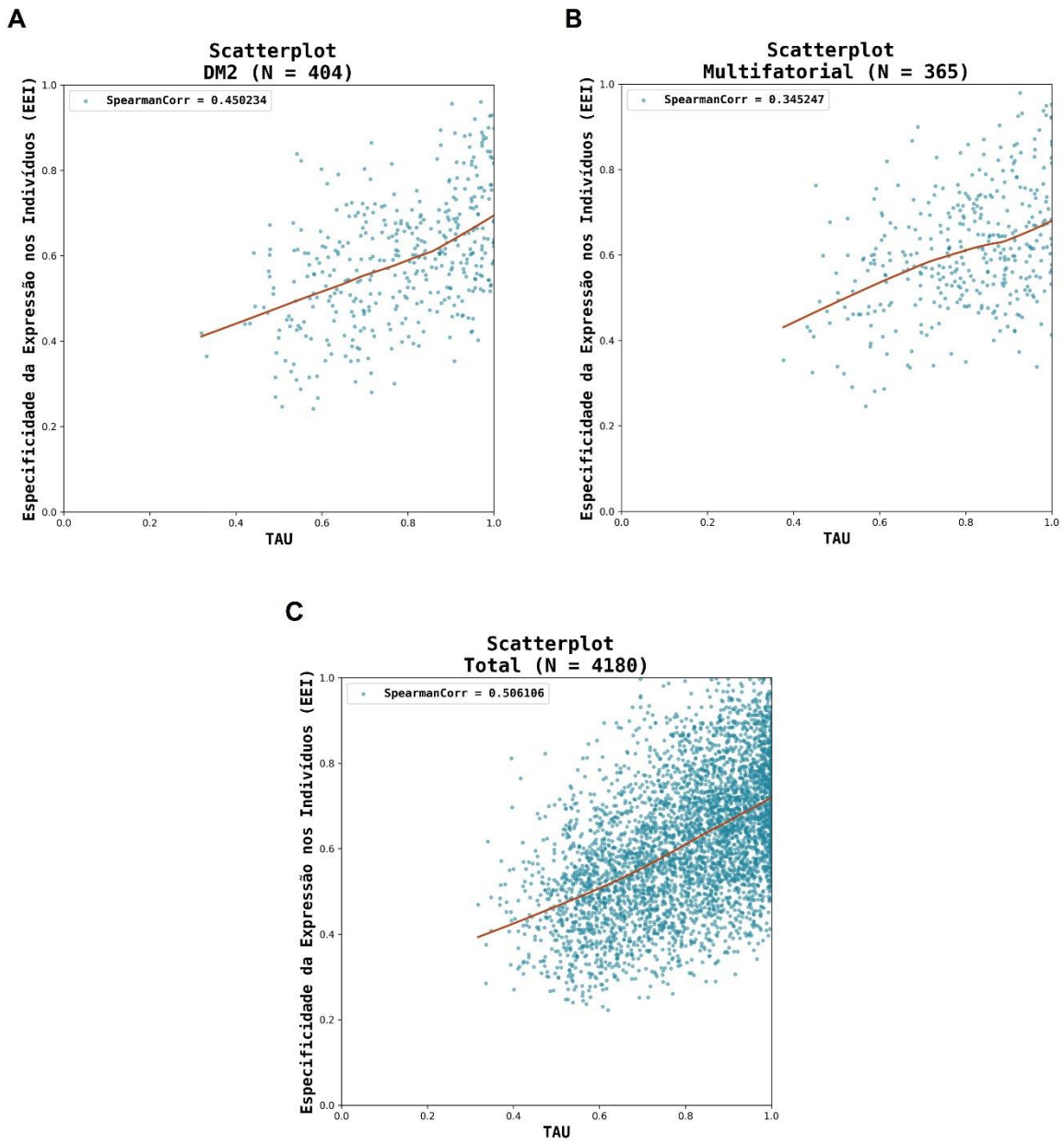

Gráfico 2.6 – Dispersão da Especificidade da Expressão nos Indivíduos (EEI) por Especificidade da Expressão nos Tecidos (TAU). Cada ponto representa um gene selecionado nos grupos específicos. Para todos os grupos de genes é possível identificar uma correlação positiva e moderada entre os dois parâmetros. **A)** Genes DM2, ρ = 0,4502. **B)** Genes Multifatorial, ρ = 0,3454. **C)** Genes Total, ρ = 0,5061.

Fonte: Elaborado pelo autor.

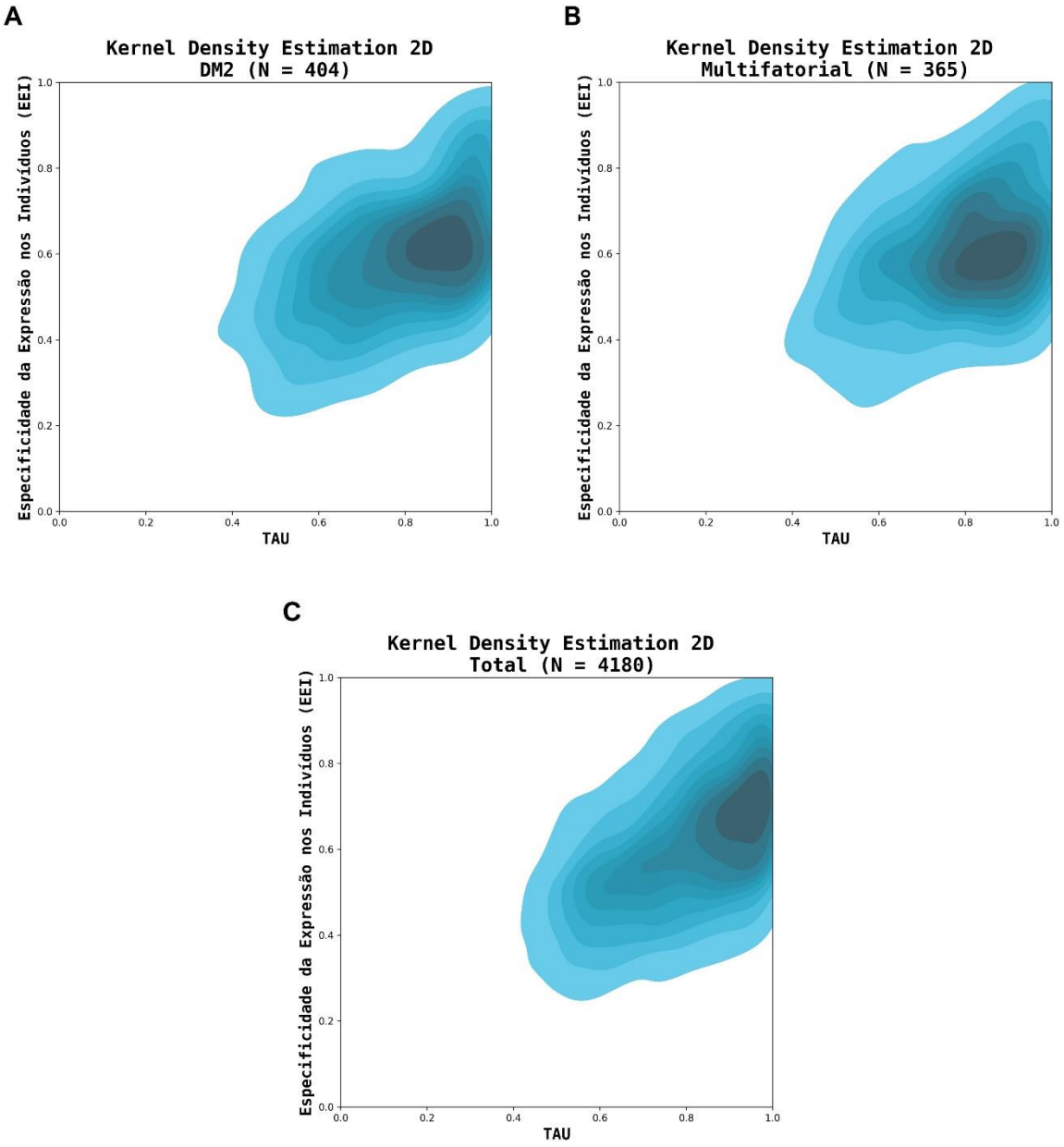

Gráfico 2.7 – KDE da Especificidade da Expressão nos Indivíduos (EEI) por Especificidade da Expressão nos Tecidos (TAU). Para todos os grupos de genes, **A)** GWAS_DM2, **B)** GWAS_MULTIFATORIAL e **C)** GWAS_TOTAL, é possível identificar um enriquecimento de genes na região de alta especificidade.

Fonte: Elaborado pelo autor.

Tabela 2.13 – Coeficiente de correlação de Spearman (ρ) do EEI e TAU para os três grupos de genes em geral e considerando apenas as regiões de baixa (TAU < 0,8) e alta (TAU > 0,8) tecido especificidade.

	EEI x TAU					
Grupo de Genes	**ρ Geral**	**P-Valor**	**ρ Baixa**	**P-Valor**	**ρ Alta**	**P-Valor**
GWAS_DM2	0,4502	1,46 e-21	**0,2809**	6,35 e-05	0,2739	6,48 e-05
GWAS_MULTIFATORIAL	0,3452	1,18 e-11	**0,3370**	9,54 e-06	0,1398	0,04
GWAS_TOTAL	0,5061	8,47 e-98	**0,3133**	1,09 e-42	0,2293	1,41 e-29

Fonte: Elaborada pelo autor.

2.3.2.6 Correlação da Expressão das Isoformas

Ao analisar os parâmetros CEI e TAU é possível observar apenas uma correlação fraca na região de alta especificidade (TAU > 0,8) para o grupo GWAS_MULTIFATORIAL (Tabela 2.14). Esse resultado indica que não há dependência do CEI em relação ao TAU.

Tabela 2.14 – Coeficiente de correlação de Spearman (ρ) do CEI e TAU para os três grupos de genes em geral e considerando apenas as regiões de baixa (TAU < 0,8) e alta (TAU > 0,8) tecido especificidade.

	CEI x TAU					
Grupo de Genes	**ρ Geral**	**P-Valor**	**ρ Baixa**	**P-Valor**	**ρ Alta**	**P-Valor**
GWAS_DM2	0,0974	0,11	-0,0225	0,8	0,0349	0,68
GWAS_MULTIFATORIAL	0,0119	0,85	-0,2498	0,01	0,2059	0,02
GWAS_TOTAL	0,1354	1,99 e-12	0,0864	0,002	0,06	0,02

Fonte: Elaborada pelo autor.

Nos KDEs correspondentes, foi verificado o mesmo comportamento do CVE, em que os grupos GWAS_MULTIFATORIAL (Gráfico 2.9B) e GWAS_TOTAL (Gráfico 2.9C) apresentam um pico de densidade de genes localizados em região de alta especificidade (TAU > 0,8), enquanto o grupo GWAS_DM2 (Gráfico 2.9A) apresenta dois picos, um mais evidente, na região de alta especificidade, assim como nos outros grupos, e outro mais sutil, na região de baixa especificidade (TAU < 0,8).

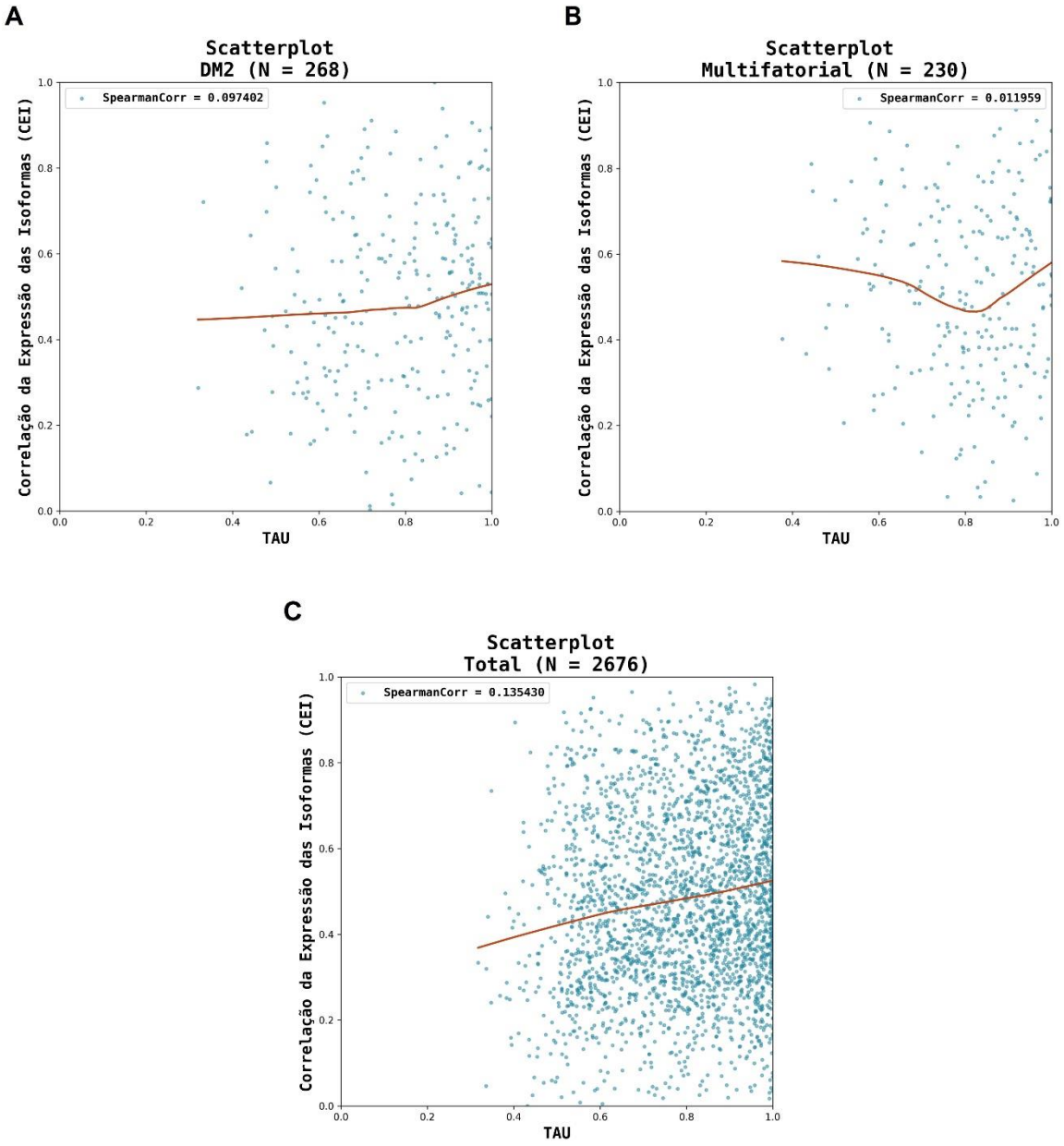

Gráfico 2.8 – Dispersão da Correlação da Expressão nas Isoformas (CEI) por Especificidade da Expressão nos Tecidos (TAU). Cada ponto representa um gene selecionado nos grupos específicos. Não há correlação entre os parâmetros para nenhum dos grupos de genes. **A)** Genes DM2, ρ = 0,0974. **B)** Genes Multifatorial, ρ = 0,0119. **C)** Genes Total, ρ = 0,1354.

Fonte: Elaborado pelo autor.

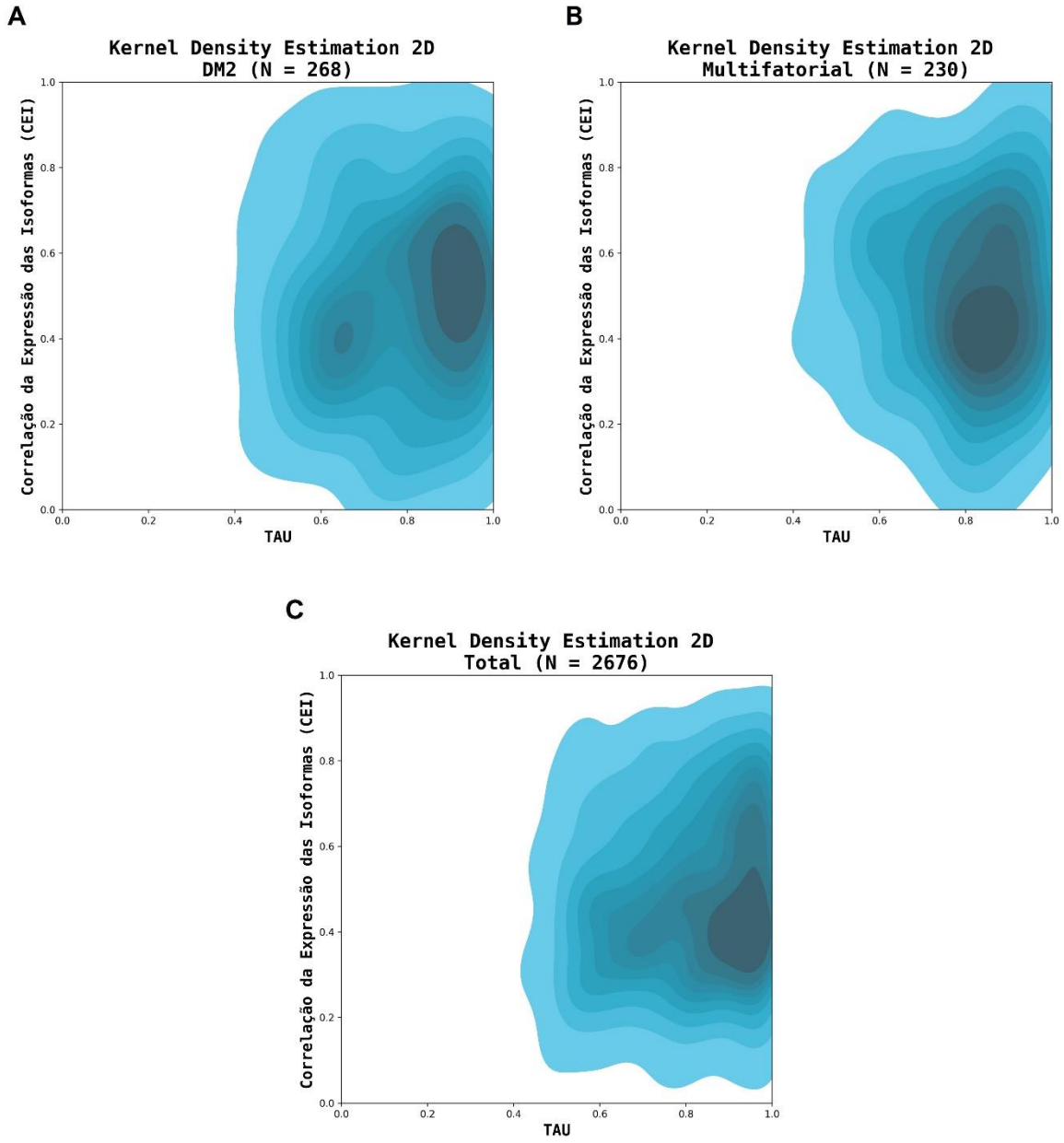

Gráfico 2.9 – KDE do Coeficiente de Variação da Expressão (CVE) por Especificidade da Expressão nos Tecidos (TAU). Para todos os grupos de genes, **A)** GWAS_DM2, **B)** GWAS_MULTIFATORIAL e **C)** GWAS_TOTAL, é possível identificar um enriquecimento de genes na região de alta especificidade. No grupo GWAS_DM2, há também um pequeno enriquecimento na região de baixa especificidade.
Fonte: Elaborado pelo autor.

2.4 Discussão e Perspectivas

Os resultados provenientes da análise evolutiva e estrutural, apresentados nas seções 2.3.1, apontam para uma tendência relacionada à essencialidade dos éxons localizados nas proximidades dos SNPs associados ao DM2. Essa tendência indica que os éxons adjacentes a

esses SNPs apresentam uma probabilidade significativamente maior de estarem localizados em regiões genômicas mais permissivas.

Embora pequenas *indels*, com tamanhos de até 50 pares de bases, representem um dos tipos mais comuns de variações genéticas subsequente aos SNPs (95), ao avaliar as duas categorias de éxons, adjacentes e distantes, foi observada uma tendência, em relação à ocorrência de *indels*, menor do que a esperada ao acaso. As medianas das distribuições dos CTIs para esses dois grupos foram de -0,1542 e -0,1663, respectivamente. Ressaltando que não foi identificada diferença estatisticamente significativa entre as tendências observadas nos dois grupos, sugerindo que ambos compartilham padrões semelhantes. Os resultados podem ter sido afetados pela quantidade de espécies analisadas e a distância evolutiva entre elas. Além da dificuldade na identificação dos estados ancestrais em locais onde essas variações ocorrem (95).

Apesar de, inicialmente, serem selecionadas 5 espécies, os dados foram comparados apenas com a espécie do camundongo (*Mus musculus*), devido a qualidade dos genomas depositados e a distância evolutiva entre as espécies. Espécies com distância evolutiva pequena, apresentavam bastante similaridade entre os genomas, indicando que a maioria dos fenômenos de *indels* ocorreram em distâncias evolutivas maiores.

Outro ponto determinante é o fato do camundongo representar uma espécie bastante estudada, com uma quantidade bem maior de proteínas descritas, possibilitando um maior número de alinhamentos para análise de comparação. Dos dados depositados no UniProt, 20.426 sequências correspondem a proteínas humanas e 17.178 correspondem a sequências do camundongo. Enquanto outras espécies razoavelmente selecionáveis, apresentam uma média de menos de 2000 proteínas (77).

O modelo também considerou uma classificação baseada apenas na categorização dos éxons, em adjacentes ou distantes, sem considerar a real distância das extremidades desses éxons em relação ao SNPs, ou a real posição deles ao longo dos éxons. Foram descritos diversos SNPs que atuam em diferentes posições e tipos de íntrons relacionados com a manifestação de doenças (96), através de diferentes mecanismos (97). Aprimorar a abordagem para um modelo contínuo e quantitativo para a categorização dos éxons em relação a localização dos SNPs, além de um modelo sistemático para a distância evolutiva, pode trazer resultados mais substanciais sobre esse comportamento.

Valores mais elevados dos parâmetros de regiões intrinsecamente desordenadas, com mediana de 0,3267 para os éxons adjacentes, e 0,2565 para os éxons distantes, para o IUPRED; e 0,4035 para os éxons adjacentes e 0,3603 para os éxons distantes, para o ACNHOR, indicam uma probabilidade significativamente maior, p-valor de 8,7124 e-10 e 1,5835 e-10,

respectivamente, de que os éxons estejam situados em regiões da sequência genômica que têm um impacto reduzido na funcionalidade da proteína quando o éxon correspondente é perdido, possivelmente devido a uma forma alternativa de *splicing* que pode compensar essa perda.

Esses resultados sugerem que os éxons adjacentes estão mais propensos a estarem envolvidos em *splicing* alternativo, visto que há uma maior ocorrência desse fenômeno em proteínas desestruturadas (98). Uma análise abrangente de dados de RNAseq detectou um enriquecimento de *splicing* alternativo, em particular de exclusão de éxons e inclusão prejudicada de microéxons em ilhotas de camundongos suscetíveis ao DM, inclusive em genes bem conhecidos, como ABCC8 e TCF7L2, sugerindo que um padrão de *splicing* alterado pode contribuir para um risco elevado de DM2 (99). Sabe-se que o *splicing* desregulado provoca a manifestação de doenças (100).

Em relação aos SNPs em regiões intergênicas, eles tenderiam, principalmente, a atuar influenciando níveis de transcrição ao invés de *splicing*, considerando uma atuação em *cis*. Porém, já foi descrito que diversos SNPs obtidos via GWAS estão aparentemente associados a transcritos não codificantes que se encontram em íntrons e em regiões intergênicas (62).

As análises dos parâmetros de expressão gênica, descritas nas seções 2.3.2 indicam que os genes com SNPs associados ao DM2 possuem uma menor tolerância as variações de sua expressão. Ao comparar esses genes com os genes com SNPs associados aos outros fenótipos descritos no GWAS *Catalog*, foi observado que eles apresentam uma maior proporção de genes localizados em regiões de baixa tecido especificidade (TAU < 0,8), 48,76% para os genes GWAS_DM2, contra 43,47% para os genes GWAS_TOTAL, um p-valor de 0,0349 indica uma diferença estatisticamente significativa entre essas duas proporções.

As distribuições dos parâmetros avaliados para os grupos de genes indicam que o GWAS_DM2 apresenta valores menores de mediana para todos os parâmetros, exceto para o CEI, além de mostrar uma correlação positiva entre TAU, CVE e EEI. Esses resultados indicam que os genes com SNPs associados ao DM2 estão concentrados em regiões de menor tecido especificidade, menor variação da expressão e menor especificidade da expressão entre os indivíduos, indicando perfis de expressão mais estáveis, características presentes em genes de *housekeeping*.

Genes de *housekeeping* são definidos como expressos de forma estável, independentemente do tipo de tecido, estágio de desenvolvimento, entre outros aspectos, (101-102), e que os genes expressos dessa maneira não são necessariamente essenciais (92). Embora uma maior concentração dos SNPs associados ao DM2 em genes com essa classificação possa

sugerir alguns indícios relacionados a prevalência, são necessárias análises que considerem os outros critérios de classificação para esses genes, além da estabilidade na expressão entre amostras. Alguns desses critérios são: essencialidade, participação na manutenção celular e conservação evolutiva (92).

3 ANÁLISE GENÉTICA DO DIABETES MELLITUS TIPO 2 EM UMA POPULAÇÃO BRASILEIRA

3.1 Introdução

3.1.1 Diabetes Mellitus Tipo 2 e Ancestralidade

O DM2 possui um caráter fortemente hereditário, e sua prevalência varia entre os grupos de ancestrais humanos continentais (28,103). Hipóteses apoiam-se em diferenças na seleção natural entre os diversos grupos de ascendência para tentar explicar as variações na prevalência de diversos fenótipos (34,104-105). Estudos avaliaram esses *loci* de suscetibilidade estabelecidos para evidências de seleção natural que geralmente envolvem a avaliação de assinaturas genéticas de seleção recente ou comparação de frequências alélicas entre grupos de ascendência (24,103,106).

Um estudo realizado com 734 indivíduos de uma amostra incluindo afro-americanos, índios americanos e europeus americanos, sugeriu que, embora o DM2 e características relacionadas possuam diferenças significativas entre os grupos, essas diferenças são consistentes com expectativas neutras baseadas na herdabilidade e nas distâncias genéticas. Essas análises não excluem o papel sutil da seleção natural diferencial, mas não fortalecem a teoria de que ela é necessária para explicar as diferenças fenotípicas entre esses grupos de ascendência (103).

Enquanto alguns fatores relacionados ao de estilo de vida impulsionam o aumento da prevalência de DM2, outras evidências entre as populações mostram que a variação genética, alavancada por critérios evolutivos, como a seleção natural, que traz alterações no genoma humano, também desempenha um papel importante. Foi relatada a evidência do efeito da seleção em genomas africanos nos mecanismos relacionados ao DM2. Os efeitos de seleção encontrados podem ter permitido que os africanos respondessem aos desafios nutricionais, alterando o metabolismo energético, a biologia do tecido adiposo e a ação da insulina, sendo efeitos que podem ter sido historicamente adaptativos em condições críticas, como fome e inflamação. Algumas outras análises poderiam fortalecer essa hipótese, como uma avaliação funcional adicional de *loci* de risco estabelecidos (31).

3.1.2 Variantes Raras

Variantes raras são aquela que possuem uma frequência alélica abaixo de 1% (MAF < 0,01) e podem desempenhar um papel importante na etiologia de características complexas e explicar a parcela da herdabilidade que não são explicadas por variantes comuns (107-108). Foi constatado que várias características complexas estão associadas a essas variantes menos frequentes (108-110), inclusive o DM2 (111-112)

Para analisar variantes raras, o desenvolvimento estatístico aborda teste de efeitos cumulativos de variantes raras em regiões genéticas, genes ou conjuntos de variantes, os *burden tests* (107,110). Um dos modelos mais populares é o Teste de Soma Alélica de Coorte (*Cohort Allelic Sum Test* – CAST), que utiliza uma pontuação para identificar e contabilizar a quantidade de indivíduos apresentam, pelo menos, uma variante rara em um gene (113).

3.1.3 Variantes Comuns

A frequência alélica das variantes comuns pode variar de 5% a 50% (0,05 < MAF < 0,5) e contribuem coletivamente como um componente importante da herdabilidade de características complexas (114). São frequentemente não codificantes, têm efeitos fracos e podem ser correlacionadas com muitas outras variantes próximas, potencialmente envolvendo vários genes (115).

Enquanto as variantes genéticas individuais ou raras conferem um risco substancial a um indivíduo, as variantes comuns não o fazem. Isto levou a considerar agregar múltiplos efeitos genéticos como em estudo de pontuações de risco, normalmente ponderadas pelo tamanho do efeito alélico específico da variante (115). Uma forma comum de variantes no genoma humano é o SNP. Esses polimorfismos servem como sinais ou marcadores, na busca por variantes comuns que influenciam a suscetibilidade de doenças também comuns.

3.1.4 Escore de Risco Poligênico

Um escore de risco poligênico (*polygenic risk score* – PRS) agrega os efeitos de muitas variantes genéticas em um único número que prediz o risco de manifestação de um determinado fenótipo. PRSs são tipicamente compostos de centenas a milhões de variantes genéticas,

geralmente SNPs, que são combinadas usando uma soma ponderada de alelos multiplicadas por seus correspondentes tamanhos de efeito, estimados a partir de GWAS (116).

Graças a GWAS diversos, a literatura mostra um grande número de SNPs associados a um amplo conjunto de características complexas, como obesidade (117), Alzheimer (118), DM2 (59,119-122), entre outros. Apesar disso, esses SNPs, normalmente, possuem um efeito pequeno e correspondem a uma pequena fração dos SNPs verdadeiramente associados ou causais, limitando o poder preditivo dessas análises (123-126).

Em um estudo de análises de características relacionadas com a herdabilidade da altura (127) a utilização de modelos lineares, além de técnicas estatísticas como o PRS, permitiram estimar a herdabilidade, inferir sobreposição genética entre características e prever fenótipos com base no perfil genético, sobretudo com base em SNPs.

3.1.4.1 Escore de Risco Poligênico e Diabetes Mellitus Tipo 2

Os PRS agregam o risco genético de alelos individuais em todo o genoma e representam um método promissor para prever a ocorrência futura de DM2, melhorando o diagnóstico precoce, a intervenção e a prevenção da doença (128-130). Esses estudos estão fornecendo previsões com maior confiança para populações cada vez mais diversas (126,131). Para o DM2 é possível estratificar os subtipos da doença, entre outras características específicas, como os alvos de tratamento, a compreensão mais detalhada das vias biológicas envolvidas e os perfis de manifestação (61).

Embora a síndrome metabólica e a obesidade sejam os preditores mais fortes para o DM2 (5,132-133), a herdabilidade é a principal motivadora para utilizar uma abordagem de PRS para converter dados genéticos em uma medida preditiva de risco ou suscetibilidade ao DM2, com o objetivo de promover uma maior capacidade preventiva, reduzindo substancialmente a progressão para o DM2 com o uso de medicamentos e melhorias no estilo de vida (134).

3.2 Métodos

3.2.1 Descrição do Coorte SABE

Os dados dos SNPs na população brasileira foram retirados do Arquivo Brasileiro Online de Mutações (ABraOM – https://abraom.ib.usp.br/index.php), da coorte SABE-1171-WGS (SABE), o repositório contém SNPs obtidos com o sequenciamento completo do genoma, de uma amostra de 1171 indivíduos de São Paulo/SP, com idade média de 71,86 anos e uma razão de indivíduos do sexo feminino para o sexo masculino de 1,74. O padrão de ancestralidade autodeclarada dos indivíduos encontrados no SABE pode ser visualizado na tabela 3.1 (135).

Do total de indivíduos, 278 se autodeclararam com DM2 e 803 sem DM2. Porém, ao avaliar os dados clínicos de cada indivíduo, foi realizada uma reclassificação utilizando os critérios de diagnóstico apresentados na Tabela 1.1, da seção 1.1.2. Com isso, a amostra ficou dividida em 303 indivíduos com DM2 e 828 indivíduos sem DM2.

Tabela 3.1 – Estatísticas das ancestralidades dos indivíduos no SABE.

Ancestralidade	Percentual
Europeia	72,6%
Africana	17,8%
Nativa Americana	6,70%
Leste Asiática	2,80%

Fonte: Elaborada pelo autor.

3.2.2 Seleção de Variantes

3.2.2.1 Chamada de Variantes Raras

Para identificar variantes potencialmente deletérios para função de proteína, dividimos os genes em três grupos. Genes que apresentam sinal em GWAS (GENES_GWAS), genes associados ao MODY e outras formas de insulinopenia (GENES_MODY) e genes de lipodistrofia e resistência à insulina (GENES_IR).

Para as variantes raras, foram selecionados as que possuem MAF < 0,01. Após isso, as variantes foram classificados de acordo com a sua função a partir dos dados de "*PredictedFunc.refGene*" descritos no SABE. Variantes sinônimas, utilizadas como controle

negativo, foram selecionadas com a classificação "*synonymous* SNV", e variantes não-sinônimas foram selecionadas com a classificação "*non synonymous* SNV".

Variantes relacionadas com perda de função da proteína (*Loss of Function* – LoF) foram selecionadas com as classificações de "*exonic;splicing*", "*frameshift deletion*", "*frameshift insertion*", "*splicing*", "*splicing;splicing*", "*stopgain*", "*stoploss*". Para as variantes associadas com *splicing*, foram consideradas apenas as formas +1, +2, -1 e -2.

Foram desconsideradas para essas análises todas as outras variantes, classificadas como: "UTR3", "UTR5", "UTR5;UTR3", "*downstream*", "*intergenic*", "*intergenic;intergenic*", "*intronic*", "*intronic;intronic*", "ncRNA_UTR5", "ncRNA_*exonic*", "ncRNA_*exonic;splicing*", "ncRNA_*intronic*", "ncRNA_*intronic*;ncRNA_*intronic*", "ncRNA_*splicing*", "*nonframeshift deletion*", "*nonframeshift insertion*", "*unknown*", "*upstream*" e "*upstream;downstream*".

Também foram identificadas variantes categorizadas em modelos *in sílico*. Utilizou-se o *Combined Annotation Dependent Depletion* (CADD), uma ferramenta que pontua o nível deletério de uma variante no genoma humano. Essa pontuação está fortemente correlacionada com a diversidade alélica, patogenicidade de variantes codificantes e não codificantes e efeitos regulatórios medidos experimentalmente, além de variantes causais dentro de sequências genômicas individuais, priorizando quantitativamente variantes causais funcionais, deletérias e de doenças em uma ampla gama de categorias funcionais, tamanhos de efeito e arquiteturas genéticas. Este método pode ser usado para priorizar a variação causal tanto em pesquisas, quanto em ambientes clínicos (136).

3.2.2.2 Chamada de Variantes Comuns

Sugerimos que os SNPs associados ao DM2 apresentam um aspecto evolutivo próximo do neutro, havendo poucos SNPs predominantemente negativos. Com isso, grande parte desses SNPs estão localizados em regiões mais permissivas dos genes, possuindo um efeito relativamente sutil na estrutura de transcritos e proteínas, pois mudanças mais substanciais tendem a ser mais deletérias. Devido a essas características, esses SNPs tendem a ser mais comuns nas populações em geral.

Com isso, foi realizada uma análise com as variantes comuns, com o objetivo de verificar quais variantes possuem uma maior associação com a manifestação do DM2 na população do SABE. Para isso, primeiramente, os indivíduos foram divididos em caso (indivíduos com DM2) e controle (indivíduos sem DM2), além de classificados em dois grupos,

correspondendo à presença ou não do respectivo SNP. Para essa análise, especificamente, foi utilizado o modelo dominante, onde a presença do SNP em qualquer um dos alelos indica que aquele indivíduo é classificado como caso. Foi criada uma tabela de contingência (Tabela 3.2) com as quatro classificações possíveis e calculada a razão de probabilidade (*odds ratio* – OR) para cada SNP associado ao DM2.

Tabela 3.2 – Tabela de contingência para o cálculo do OR.

	Caso	**Controle**
Exposto aos SNPs	A	B
Não Exposto aos SNPs	C	D

Fonte: Elaborada pelo autor.

$$Razão\ de\ Casos\ Expostos = \frac{Número\ de\ Casos\ Expostos\ (A)}{Número\ de\ Casos\ Não\ Expostos\ (C)}$$

$$Razão\ de\ Controles\ Expostos = \frac{Número\ de\ Controles\ Expostos\ (B)}{Número\ de\ Controles\ Não\ Expostos\ (D)}$$

$$Odds\ Ratio\ (OR) = \frac{Razão\ de\ Casos\ Expostos\ \left(\frac{A}{C}\right)}{Razão\ de\ Controles\ Expostos\ \left(\frac{B}{D}\right)} = \frac{A * D}{B * C}$$

O OR mede quão fortemente um evento está associado à exposição, indicando a proporção de dois grupos de chances: as chances de o evento ocorrer em um grupo exposto e as chances de o evento ocorrer em um grupo não exposto. É o índice de tamanho de efeito mais utilizado para demonstrar aumento ou diminuição na chance de uma doença em estudos epidemiológicos, por exemplo (137).

O tamanho do efeito pode ser classificado, de acordo com o OR em três categorias: efeito baixo; efeito moderada ou efeito forte. Os limites de OR (inverso), para cada categoria são de 1,46 (0,68), 2,50 (0,40) e 4,14 (0,24), respectivamente, considerando uma taxa de 10% de prevalência em pessoas não expostas (138).

3.2.3 Análises de Ancestralidade

Considerando a variação da prevalência do DM2 entre os grupos ancestrais e a influência do contexto evolutivo, o principal objetivo dessa análise é comparar os dados de frequência alélica dos SNPs selecionados.

Os dados de frequência alélica foram retirados do banco de dados do *Genome Aggregation Database Browser* (GNOMAD – https://gnomad.broadinstitute.org/), versão v2.1.1 (GRCh37/hg19), que abrange 125.748 sequências de exoma e 15.708 sequências de genoma completo de indivíduos não relacionados sequenciados como parte de vários estudos genéticos de populações e doenças específicas (139). Os dados de frequência dos SNPs na população brasileira foram retirados do SABE (135).

Os valores de frequência alélica foram selecionados de acordo com a ancestralidade da amostra: *African/African American* (GNOMAD_AFR_FREQ), *Latino/Admixed American* (GNOMAD_AMR_FREQ), *East Asian* (GNOMAD_EAS_FREQ) e *Non-Finnish European* (GNOMAD_NFE_FREQ), além do valor de frequência geral (GNOMAD_FREQ). Os dados de frequência da população brasileira foram classificados como SABE_FREQ. O padrão de ancestralidade das frequências analisadas e encontrados no GNOMAD pode ser visualizado na tabela 3.3, a seguir

Tabela 3.3 – Estatísticas das ancestralidades das amostras do GNOMAD.

Ancestralidade	**Percentual**
Non-Finnish European	44,73%
African/African American	27,23%
Latino/Admixed American	10,04%
East Asian	3,42%
Outras	14,58%

Fonte: Elaborada pelo autor.

Para identificar quais SNPs apresentavam uma maior ou menor frequência na população brasileira, ao ser comparado com as populações do GNOMAD, foi calculado um parâmetro de frequência relativa através da razão entre a maior e menor frequência encontradas nos dois bancos de dados. Primeiramente os SNPs foram divididos em dois grupos: um com os SNPs mais prevalentes no SABE e outro com os mais prevalentes no GNOMAD. Para o primeiro grupo, a frequência relativa foi calculada como SABE_FREQ/GNOMAD_FREQ, enquanto no outro grupo ela era representada pelo inverso desse valor.

3.2.4 Seleção dos Escores de Risco Poligênico

Atualmente, constam no PGS *Catalog* 102 estudos de PRS relacionados ao DM2 (140). No entanto, a maioria dos escores existentes de DM2 foram desenvolvidos e validados em

indivíduos de ascendência europeia (129,141). Considerando que o poder preditivo do PRS é, muitas vezes, atenuado em populações não europeias, e as crescentes taxas de prevalência do DM2 que outros grupos ancestrais estão sofrendo, é extremamente importante avaliar e otimizar a capacidade de transferência de um PRS em diversas populações, inclusive a brasileira.

O principal critério para selecionar os estudos para esta análise foi a característica multiancestral da população, com amostras com indivíduos de diversas ancestralidades, entre europeia, africana, hispânica e latino-americana, asiática, do oriente médio, entre outras (Figura 07). Com isso, selecionamos 4 estudos de PRS: o PGS000804 (PRS0804) (142), o PGS002308 (PRS2308) (129), o PGS002026 (PRS2026) (143) e o PGS003443 (PRS3443) (112).

O PRS0804 e o PRS2308 partiu de dados de GWAS de populações diferentes, em sua grande maioria europeia, 79,2% e 81,7% respectivamente, e seus PRS foram avaliados em populações também diversas. O PRS2026 teve o seu PRS desenvolvido em uma população totalmente europeia, mas foi avaliado em populações com ancestralidades diferentes. Já o PRS3443, apesar de partir de dados de GWAS de populações exclusivamente europeia, o seu PRS foi desenvolvido em populações com ascendências hispânico e latino-americanas.

Todos esses estudos foram realizados ou validados em amostras com milhares ou milhões de indivíduos. Alguns deles avaliam milhões de SNPs, enquanto outros analisam um conjunto mais específico de variantes (Tabela 3.4).

Tabela 3.4 – Quantidade de variantes avaliadas e o tamanho total das amostras dos estudos de PRS.

PRS	**# Variantes Avaliadas**	**Desenvolvimento do PRS (# Indivíduos)**	**Validação do PRS (# Indivíduos)**
PRS0804	582	2.814.564	467.951
PRS2308	1.259.754	1.099.372	173.999
PRS2026	830.783	391.124	46.387
PRS3443	1.092.496	898.130	1.484

Fonte: Elaborada pelo autor.

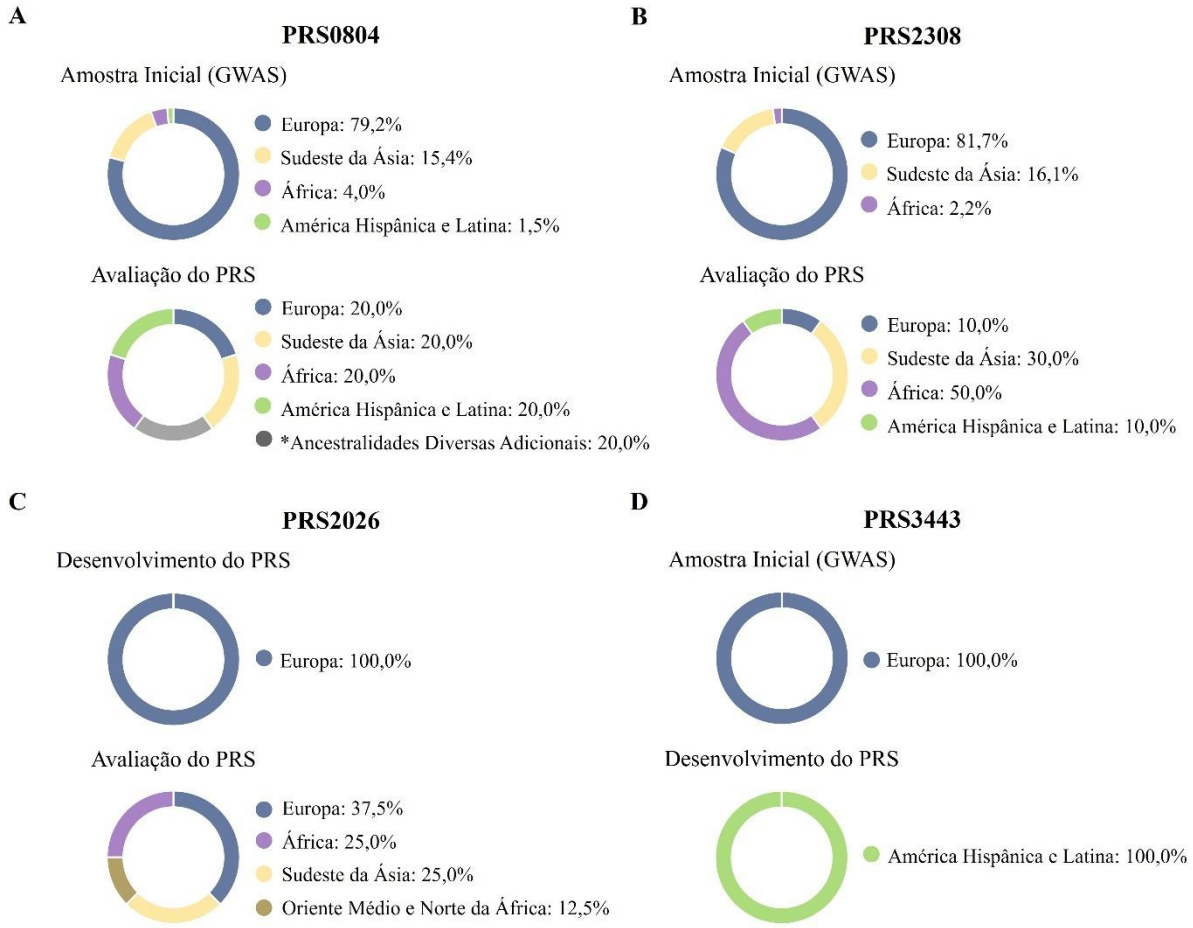

Figura 3.1 – Distribuição das ancestralidades das populações estudadas nos PRS selecionados. **A)** PRS0804, **B)** PRS2308, **C)** PRS2026 e **D)** PRS3443.
Fonte: Elaborada pelo autor.

3.3 Resultados

3.3.1 Coorte do SABE

Para as análises do DM2 no SABE, foram utilizados os dados de um total de 1131 indivíduos, 725 do sexo feminino e 404 do sexo masculino, com idade média de 72,89 anos. A partir dos 944 SNPs associados ao DM2 selecionados na seção 2.3.1.1, foram identificados 915 SNPs no SABE. Dentre eles, 865 SNPs classificados como comuns (MAF > 0,01) e, dentro desse grupo, 455 apresentam o alelo de risco como o mais frequente na população (RAF > 0,5). O total de SNPs raros foi de 50.

3.3.2 Chamada de Variantes Raras

Os resultados apresentados nesta seção correspondem ao métodos especificado na Seção 3.2.2.1.

A lista de GENES_GWAS foi selecionada como mostrado na seção 2.3.1.1, com um total de 503 genes. Os GENES_MODY compreendem um total de 48 genes já descritos como relacionados ao fenótipo, enquanto os GENES_IR correspondem a um total de 11 genes associados (Tabela 3.5).

Tabela 3.5 – Quantidade de genes selecionados para cada grupo.

Grupo de Genes	Genes
GENES_GWAS	503
GENES_MODY	48
GENES_IR	11

Fonte: Elaborada pelo autor.

Variantes sinônimas, utilizadas como controle negativo, foram selecionadas com a classificação de “synonymous SNV”, com um total de 6153 variantes para GWAS, 490 para MODY e 134 para IR (Tabela 3.6).

Variantes não-sinônimas foram selecionadas com a classificação de “non synonymous SNV”, com um total de 8580 variantes para GWAS, 653 para MODY e 184 para IR.

Variantes relacionadas com perda de função da proteína (*Loss of Function* – LOF), foram selecionadas com as classificações descritas na seção 3.2.2.1, para um total de 1212 variantes para GWAS, 77 para MODY e 34 para IR (Tabela 3.6).

A partir das variantes não sinônimas e as relacionadas com perda de função, foram selecionadas apenas as variantes categorizadas em modelos *in sílico*, que apresentam um CADD > 20 (NS_LOF_CADD20). Com isso, as variantes selecionadas para essa análise formam um total de 5481 para GWAS, 435 para MODY e 128 para IR (Tabela 3.6).

Tabela 3.6 – Quantidade de variantes selecionadas para cada grupo de genes.

Grupo de Genes	Variantes Sinônimas	Variantes Não-Sinônimas	Variantes *Loss of Function*	Variantes NS_LOF_CADD20
GENES_GWAS	6153	8580	1212	5481
GENES_MODY	490	653	77	435
GENES_IR	134	184	34	128

Fonte: Elaborada pelo autor.

Após selecionar e classificar todos as variantes, foi contabilizado, nos genótipos do SABE, para cada gene, o número de vezes em que elas estavam presentes nos indivíduos, classificados em caso (indivíduos com DM2) e controle (indivíduos sem DM2). Para essa contagem, foi calculado o OR para cada um dos genes, buscando identificar um maior enriquecimento de variantes raras com potencial deletério em regiões específicas.

Para as variantes sinônimas, foram identificados 22 genes para o grupo GENES_GWAS e 2 para GENES_MODY que apresentam um p-valor significativo nominal (< 0,05), sendo necessário realizar correções para múltiplos testes para verificar um real padrão para corresponder a genes que possuem um enriquecimento de variantes raras em sua região (Tabela 3.7 e 3.8).

Foram identificados 14 genes para o grupo GENES_GWAS e 3 genes do grupo GENES_MODY que apresentam um p-valor significativo nominal (< 0,05), sendo necessário realizar correções para múltiplos testes para verificar um real padrão de corresponder a genes que possuem um enriquecimento de variantes raras em sua região. Não foi identificado nenhum gene do grupo GENES_IR que apresentou significância estatística nessa análise (Tabela 3.9 e 3.10).

Tabela 3.7 – Lista de GENES_MODY que apresentam um enriquecimento de variantes sinônimas raras na população do SABE e que obtiveram significância estatística nominal (p-valor < 0,05).

VARIANTES SINÔNIMAS				
GENE_MODY	**OR**	**P-VALOR**	**DNS**	**DS**
HNF1B	5,4804	0,0162	3	6
COQ2	2,3575	0,0078	21	18

NDS indica o número de alelos em indivíduos que não possuem DM2.
DS indica o número de alelos em indivíduos que não possuem DM2.
Fonte: Elaborada pelo autor.

Tabela 3.8 – Lista de GENE_GWAS que apresentam um enriquecimento de variantes sinônimas raras na população do SABE e que obtiveram significância estatística nominal (p-valor < 0,05).

VARIANTES SINÔNIMAS				
GENE_GWAS	**OR**	**P-VALOR**	**NDS**	**DS**
ZFHX3*	0,4911	0,0001	189	34
LEP	3,0944	0,0010	16	18
KCNU1	2,4497	0,0012	28	25
SPIN2A	7,0528	0,0041	3	8
PCSK1	4,9249	0,0043	5	9
FAM13A*	0,2534	0,0087	43	4
TGFBR3	2,4359	0,0097	18	16
TFRC	2,0568	0,0127	28	21
ZBTB46	2,1462	0,0153	23	18
EPC2	2,5489	0,0154	14	13
APOE	5,4948	0,0160	3	6
HNF1B	5,4804	0,0162	3	6
PDHX	2,4040	0,0169	16	14
CLEC14A	2,1191	0,0203	22	17
ALDH2	3,8316	0,0219	5	7
ZMIZ1*	0,5568	0,0238	88	18
ANK1	1,4461	0,0259	106	56
KCNB2*	0,3205	0,0315	34	4
HLA-B	1,9023	0,0354	27	18
ELFN1	1,4998	0,0397	73	40
PEPD	2,3139	0,0408	13	11
KCNK16	3,2843	0,0497	5	6

NDS indica o número de alelos em indivíduos que não possuem DM2.
DS indica o número de alelos em indivíduos que não possuem DM2.
*Gene que apresenta OR < 1,00, indicando uma característica protetiva ao DM2.
Fonte: Elaborada pelo autor.

Tabela 3.9 – Lista de GENES_MODY que apresentam um enriquecimento de variantes raras não-sinônimas, perda de função e deletérias (CADD > 20) na população do SABE e que obtiveram significância estatística (p-valor < 0,05).

NS_LOF_CADD20*				
GENE_MODY	**OR**	**P-VALOR**	**NDS**	**DS**
EIF2AK3	2,3173	0,0404	12	11
MAFA	2,5139	0,0354	11	10
SLC2A2	3,8341	0,0218	4	7

NDS indica o número de alelos em indivíduos que não possuem DM2.
DS indica o número de alelos em indivíduos que não possuem DM2.
*Variantes não-sinônimas (NS), relacionadas com perda de função (LOF) e que apresentam CADD > 20.
Fonte: Elaborada pelo autor

Tabela 3.10 – Lista de GENE_GWAS que apresentam um enriquecimento de variantes raras não-sinônimas, perda de função e deletérias (CADD > 20) na população do SABE e que obtiveram significância estatística (p-valor < 0,05).

NS_LOF_CADD20*				
GENE_GWAS	**OR**	**P-VALOR**	**NDS**	**DS**
ANO4	6,4037	0,0072	3	7
SACS	1,5093	0,0083	116	64
PTCH1	2,2376	0,0113	22	18
SLC2A2	3,8341	0,0218	5	7
CPQ	2,3744	0,0227	15	13
ZNF703	2,5056	0,0279	12	11
SLX4	1,7712	0,0294	37	24
CRYBA2	4,5682	0,0377	3	5
NUS1	22,1193	0,0390	0	4
LRFN2	2,7389	0,0441	8	8
CACNA2D3	2,0925	0,0453	17	13
USP44**	0,1300	0,0463	21	1
EMB	3,2872	0,0497	5	6
ZBTB46	5,4769	0,0497	2	4

NDS indica o número de alelos em indivíduos que não possuem DM2.
DS indica o número de alelos em indivíduos que não possuem DM2.
*Variantes não-sinônimas (NS), relacionadas com perda de função (LOF) e que apresentam CADD > 20.
**Gene apresenta OR < 1,00, indicando uma característica protetiva ao DM2.
Fonte: Elaborada pelo autor.

3.3.3 Chamada de Variantes Comuns

Os resultados apresentados nesta seção correspondem ao métodos especificados na Seção 3.2.2.2.

Com os resultados desta análise, foram identificados e validados 63 SNPs (7,96%) com um OR estatisticamente significativo para os indivíduos do SABE, indicando uma real associação desses SNPs quando analisados na população brasileira.

Os SNPs foram classificados quanto ao seu direcionamento. OR > 1 indica um caráter de risco, enquanto um OR < 1 indica um caráter protetivo. Com isso, foi verificado quais SNPs possuem direcionamentos de associação iguais, ou opostos, tanto no SABE, quanto no estudo original reportado no GWAS, através do parâmetro SENSE, em que SENSE = 1 indica o mesmo direcionamento de associação e -1 indica direcionamento oposto.

Esta identificação sugere que alguns SNPs podem ter uma característica de risco em determinado banco de dados, e uma característica protetiva em outro, podendo indicar comportamentos diferentes de acordo com as individualidades de cada população, como a ancestralidade.

Do total de SNPs testados, 485 (61,32%) apresentam um direcionamento igual ao relatado no GWAS original. Dos 63 SNPs validados, 52 deles (82,54%) apresentam esse mesmo comportamento. Enquanto 11 SNPs apresentam um direcionamento oposto ao relatado, possuindo caráter protetivo na população brasileira, de acordo com os dados encontrados no SABE (Tabela 3.11).

Tabela 3.11 – Lista de SNPs que apresentam OR com significância estatística (p-valor < 0,05) na população do SABE. OR representa o *odds ratio* da análise dos dados do SABE, OR_GWAS representa o *odds ratio* resgatado dos estudos originais do GWAS *Catalog*, DS é o número de indivíduos com DM2 que apresentam os SNPs, e SENSE indica o direcionamento do SNP.

SNP	DS	OR_GWAS	OR	P-VALOR	SENSE
rs4922793	227	1,04	**1,72**	0,0005	1,00
rs329122	229	1,04	**1,63**	0,0020	1,00
rs2421897	236	1,08	**0,59**	0,0030	**-1,00**
rs201375651	107	1,04	**0,62**	0,0034	**-1,00**
rs7645517	88	1,08	**1,56**	0,0035	1,00
rs1493694	121	1,09	**1,49**	0,0045	1,00
rs11820019	276	1,16	**2,10**	0,0054	1,00
rs2583934	72	1,06	**1,60**	0,0066	1,00
rs2833610	46	1,17	**0,61**	0,0066	**-1,00**
rs62007683	177	1,04	**1,45**	0,0078	1,00
rs96844	63	1,04	**1,58**	0,0083	1,00
rs9892728	101	1,05	**1,47**	0,0096	1,00
rs539515	121	1,05	**1,43**	0,0106	1,00
rs6416749	48	1,05	**1,65**	0,0107	1,00
rs10923931	101	1,13	**1,45**	0,0112	1,00
rs111246699	172	1,06	**1,41**	0,0121	1,00
rs7674212	61	1,07	**0,67**	0,0128	**-1,00**
rs476828	135	1,09	**1,41**	0,0129	1,00
rs523288	128	1,05	**1,41**	0,0130	1,00
rs35906730	100	1,15	**1,54**	0,0136	1,00
rs17265513	83	1,05	**1,46**	0,0141	1,00
rs1260326	258	1,05	**1,63**	0,0144	1,00
rs78020297	35	1,09	**1,74**	0,0145	1,00
rs78840640	23	1,11	**1,98**	0,0150	1,00
rs11793035	83	1,04	**1,46**	0,0153	1,00
rs2283228	243	1,20	**1,52**	0,0158	1,00
rs7403531	30	1,10	**1,78**	0,0177	1,00
rs9465871	156	1,18	**1,38**	0,0191	1,00
rs34773007	21	1,28	**1,98**	0,0203	1,00
rs6723108	290	1,27	**2,78**	0,0203	1,00
rs11073333	31	1,06	**1,73**	0,0211	1,00
rs3217860	169	1,05	**1,37**	0,0223	1,00
rs10938398	204	1,05	**1,38**	0,0271	1,00
rs11496066	193	1,08	**0,73**	0,0292	**-1,00**

(continua)

(continuação)

Tabela 3.11 – Lista de SNPs que apresentam OR com significância estatística (p-valor < 0,05) na população do SABE. OR representa o *odds ratio* da análise dos dados do SABE, OR_GWAS representa o *odds ratio* resgatado dos estudos originais do GWAS *Catalog*, DS é o número de indivíduos com DM2 que apresentam os SNPs, e SENSE indica o direcionamento do SNP.

SNP	DS	OR_GWAS	OR	P-VALOR	SENSE
rs75253922	110	1,05	**1,36**	0,0293	1,00
rs8108269	180	1,06	**1,36**	0,0295	1,00
rs1894299	170	1,13	**1,43**	0,0295	1,00
rs6021276	24	1,06	**0,60**	0,0300	**-1,00**
rs6026382	228	1,70	**1,41**	0,0304	1,00
rs7756992	175	1,15	**1,35**	0,0309	1,00
rs6438234	178	1,05	**1,35**	0,0314	1,00
rs343092	46	1,16	**1,52**	0,0318	1,00
rs184509201	290	1,21	**2,79**	0,0333	1,00
rs76263492	34	1,09	**1,62**	0,0336	1,00
rs34872471	178	1,31	**1,34**	0,0339	1,00
rs34965774	112	1,06	**1,35**	0,0341	1,00
rs4930091	225	1,04	**0,71**	0,0347	**-1,00**
rs60089934	51	1,04	**1,48**	0,0358	1,00
rs10498828	215	1,05	**1,37**	0,0361	1,00
rs780094	258	1,06	**1,51**	0,0364	1,00
rs7572857	228	1,05	**1,39**	0,0403	1,00
rs12027542	234	1,41	**0,70**	0,0407	**-1,00**
rs11680058	251	1,06	**0,67**	0,0410	**-1,00**
rs3772071	264	1,05	**0,63**	0,0432	**-1,00**
rs3115960	68	1,03	**1,40**	0,0440	1,00
rs73347525	187	1,06	**0,75**	0,0461	**-1,00**
rs1016565	216	1,04	**1,35**	0,0464	1,00
rs12277475	21	1,28	**1,77**	0,0466	1,00
rs2237897	246	1,30	**1,44**	0,0468	1,00
rs13266634	189	1,11	**1,32**	0,0479	1,00
rs56200889	189	1,05	**1,32**	0,0482	1,00
rs5945326	156	1,14	**1,35**	0,0493	1,00
rs79687284	23	1,16	**1,71**	0,0499	1,00

Fonte: Elaborada pelo autor.

Os resultados dos ORs calculados para os SNPs na população do SABE foram comparados com os ORs registrado nos estudos originais encontrados no GWAS *Catalog*. Foi identificada, através de um teste de Mann-Whitney, uma diferença estatisticamente significativa entre as duas amostras (p-valor = 4,30 e-06). Enquanto a distribuição dos SNPs reportados no GWAS *Catalog* possuem uma maior concentração de valores de OR próximos de 1, os OR calculados com os dados do SABE possuem uma distribuição menos concentrada (Gráfico 3.1).

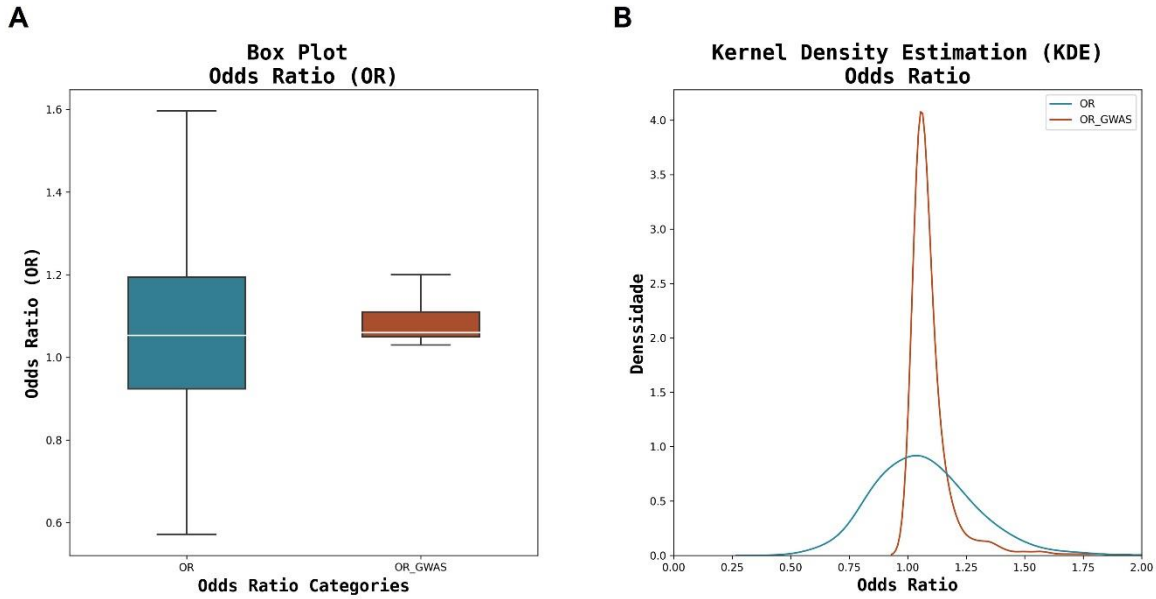

Gráfico 3.1 – Representação em *box plot* e KDE das distribuições de OR das duas amostras, SABE e GWAS.
Fonte: Elaborado pelo autor.

Ao selecionar apenas os SNPs que possuem uma significância estatística para verificar a dispersão de uma distribuição em comparação com a outra, foi possível identificar os SNPs que possuem comportamentos opostos nas duas amostras e os SNPs que possuem um maior efeito em uma população específica. Os SNPs rs184509201, no gene TCF7L2 e rs6723108, no gene TMEM163, foram os que apresentaram o maior efeito, com OR de 2,79 e 2,78 respectivamente. É possível observar também que esses dois SNPs possuem uma alta frequência na população (Gráfico 3.2).

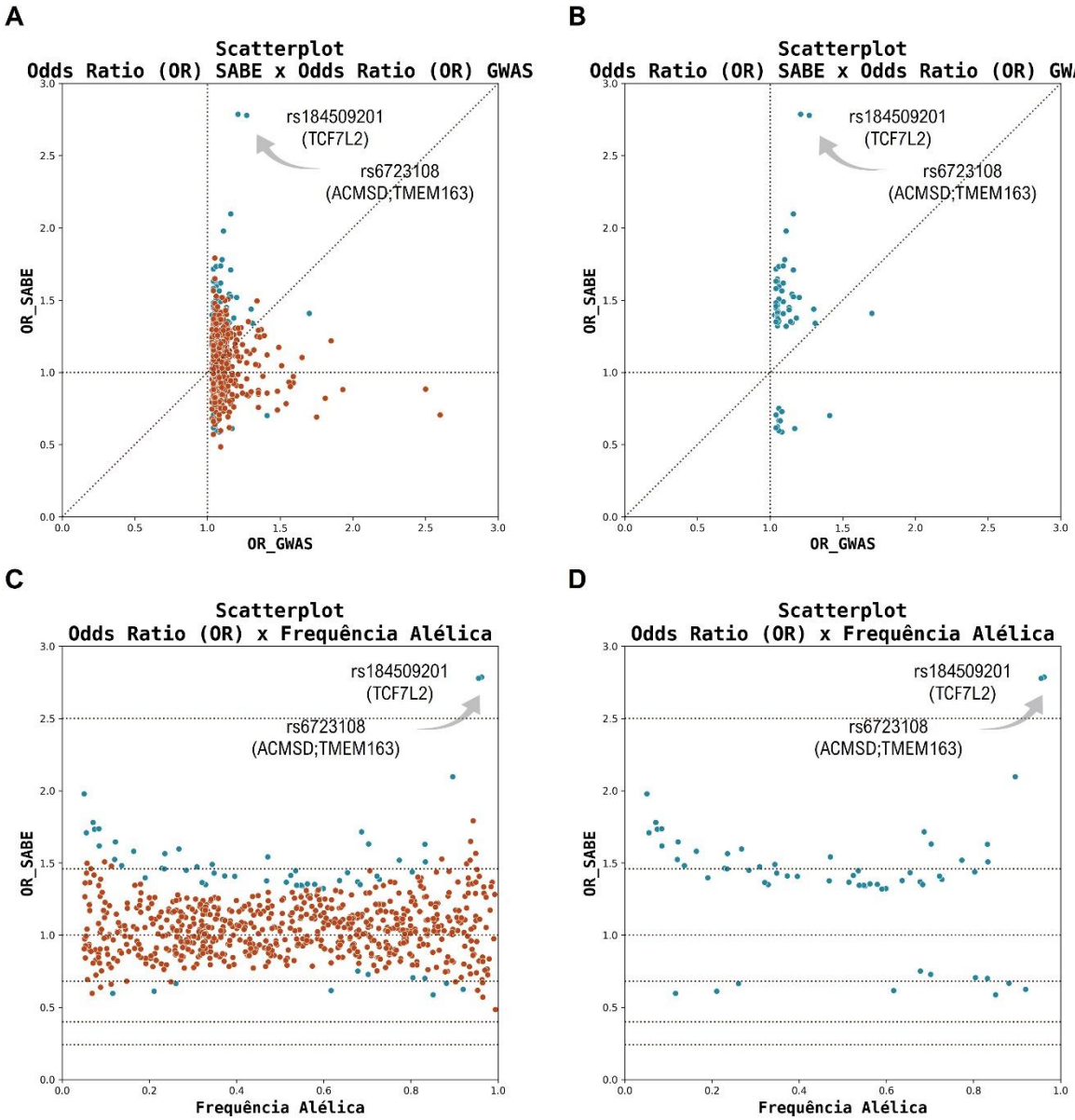

Gráfico 3.2 – Dispersão do OR das duas amostras diferentes, SABE e GWAS. **A)** Total de SNPs analisados e **B)** SNPs que apresentam significância estatística. Dispersão do OR pela frequência alélica de cada SNP. **C)** Total de SNPs analisados e **D)** SNPs que apresentam significância estatística.
Fonte: Elaborado pelo autor.

3.3.3.1 Poder Estatístico da Amostra

Com o objetivo de verificar a robustez estatística da amostra na análise de variantes comuns e projetar um tamanho de amostra considerado ideal para estudos futuros, foi realizada uma análise de poder estatístico.

O poder estatístico é a probabilidade de um teste de significância detectar um efeito quando ele realmente existe. Um alto poder estatístico indica uma grande chance de um teste

detectar um efeito verdadeiro, enquanto um baixo poder significa que seu teste tem uma pequena chance de detectar um efeito verdadeiro ou que os resultados podem ser distorcidos por erros aleatórios e sistemáticos. Ele é influenciado principalmente pelo tamanho da amostra, tamanho do efeito e nível de significância. Uma análise de poder estatístico pode ser usada para determinar o tamanho da amostra necessário para um estudo, para isso é definido um poder estatístico de 80% como parâmetro para a determinação do tamanho ideal da amostra (144).

O poder estatístico e o tamanho da amostra foi calculado através da função epi.sscc (https://rdrr.io/cran/epiR/man/epi.sscc.html), do pacote epiR: *Tools for the Analysis of Epidemiological Data* (https://cran.r-project.org/web/packages/epiR/index.html).

Os resultados de poder estatístico mostram valores baixos para as análises, sugerindo que a quantidade de indivíduos no bando de dados do SABE é abaixo do tamanho ideal de amostra (Gráfico 3.3A). Os resultados do tamanho da amostra mostram uma concentração maior de dados nas regiões de valores de log10 próximo de 2, indicando que o tamanho ideal da amostra deve ser, pelo menos, 100 vezes o tamanho atual encontrado no SABE (Gráfico 3.3B).

Ao verificar a dispersão do poder estatístico dos SNPs analisados, é possível observar que apenas um SNP com significância estatística para o OR apresenta um poder estatístico acima de 0,8 (80%), o rs6026382, no gene APCDD1L (Gráfico 3.3A e B).

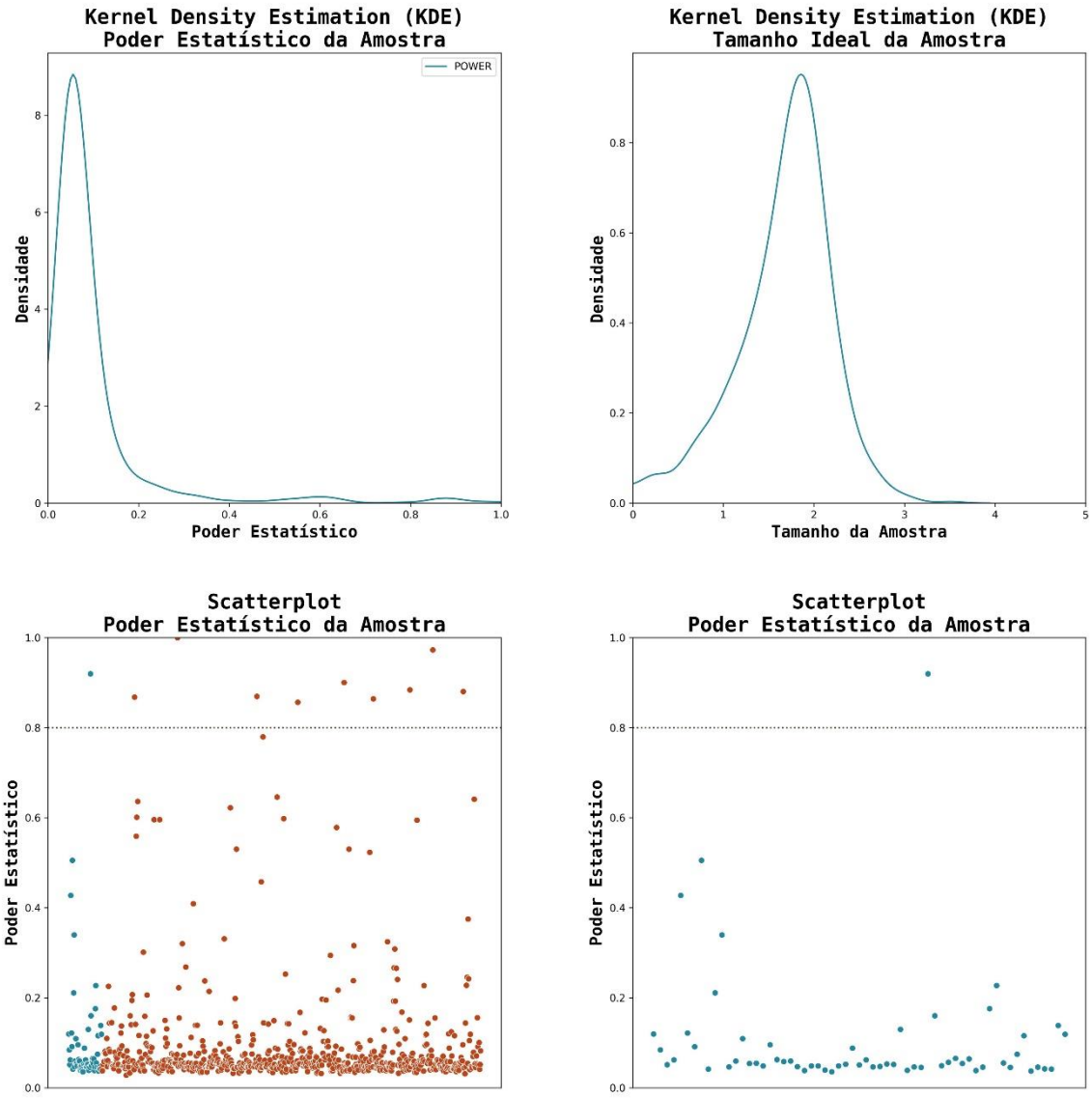

Gráfico 3.3 - **A)** *Kernel Density Estimation* dos poderes estatísticos de cada SNP para a amostra de indivíduos do SABE. **B)** *Kernel Density Estimation* do tamanho ideal da amostra para cada SNP, em escala de log10. **C)** Dispersão dos poderes estatísticos de cada SNP selecionada na análise de variantes comuns. **D)** Dispersão dos SNPs que apresentaram significância estatística.
Fonte: Elaborado pelo autor.

3.3.4 Análises de Ancestralidade

Os resultados apresentados nesta seção correspondem ao objetivo especificado na seção 3.2.3.

Nas análises, foram comparadas as frequências alélicas entre a população do SABE e os grupos de ascendências distintas do GNOMAD, identificando os SNPs que possuem uma frequência que mais difere entre os demais grupos, e os SNPs que possuem frequências mais

próximas, e que representam um padrão de manifestação do fenótipo em todos os grupos de ascendência.

Ao verificar as distribuições das frequências, através de KDE, é possível observar que a distribuição correspondente ao SABE_FREQ apresenta dois picos de densidade próximo as frequências do alelo de risco de 0,3 e 0,7, comportamento não observado na distribuição do GNOMAD_FREQ (Gráfico 3.4A). Apesar disso, não houve diferença significativa entre as duas distribuições (Tabela 3.12).

Ao expandir o KDE para as outras ancestralidades do GNOMAD (Gráfico 3.4B), é possível identificar que as distribuições correspondentes apresentam um padrão bem semelhante, exceto a GNOMAD_FREQ_AFR, que possui uma grande queda de densidade em frequências próximas a 0,4, comportamento inverso ao SABE_FREQ. Porém, como já observado, não há diferença significativa entre as distribuições (Tabela 3.12).

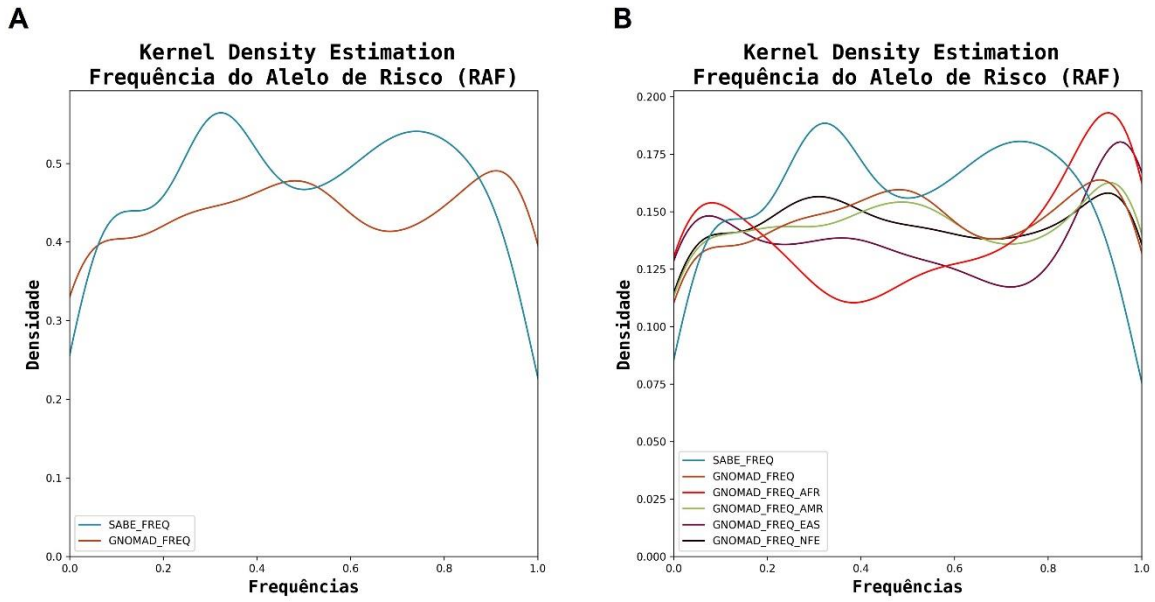

Gráfico 3.4 – KDE das frequências dos SNPs do **A)** SABE (SABE_FREQ) e GNOMAD (GNOMAD_FREQ) e de **B)** diferentes ancestralidades do GNOMAD, *African/African American* (GNOMAD_AFR), *Latino/Admixed American* (GNOMAD_AMR), *East Asian* (GNOMAD_EAS) e *Non-Finnish European* (GNOMAD_NFE).
Fonte: Elaborado pelo autor.

Todas as distribuições de ancestralidade encontradas no GNOMAD apresentaram uma correlação bastante forte entre suas frequências e as encontradas no SABE. Indicando que a grande maioria dos SNPs não divergem muito em termos de frequência alélica (Tabela 3.12).

Tabela 3.12 – Valores de P e ρ para os testes estatísticos de Mann-Whitney e o coeficiente de correlação de Spearman para os pares de amostras populacionais.

		Mann-Whitney P-Valor	**Correlação de Spearman (ρ)**	**P-Valor (ρ)**
SABE_FREQ	GNOMAD_FREQ	0,2165	0,8906	9,70 e-281
SABE_FREQ	GNOMAD_AFR	0,0507	0,8193	8,19 e-199
SABE_FREQ	GNOMAD_AMR	0,2324	0,8684	3,02 e-250
SABE_FREQ	GNOMAD_EAS	0,1459	0,7882	8,30 e-174
SABE_FREQ	GNOMAD_NFE	0,3591	0,8834	3,50 e-270

Fonte: Elaborada pelo autor

Nos gráficos de dispersão, é possível observar esse padrão bem definido de correlação das frequências do SABE_FREQ com as frequências dos outros grupos analisados (Gráfico 3.5), porém há uma correlação moderada entres as frequências desses grupos (Tabela 3.12).

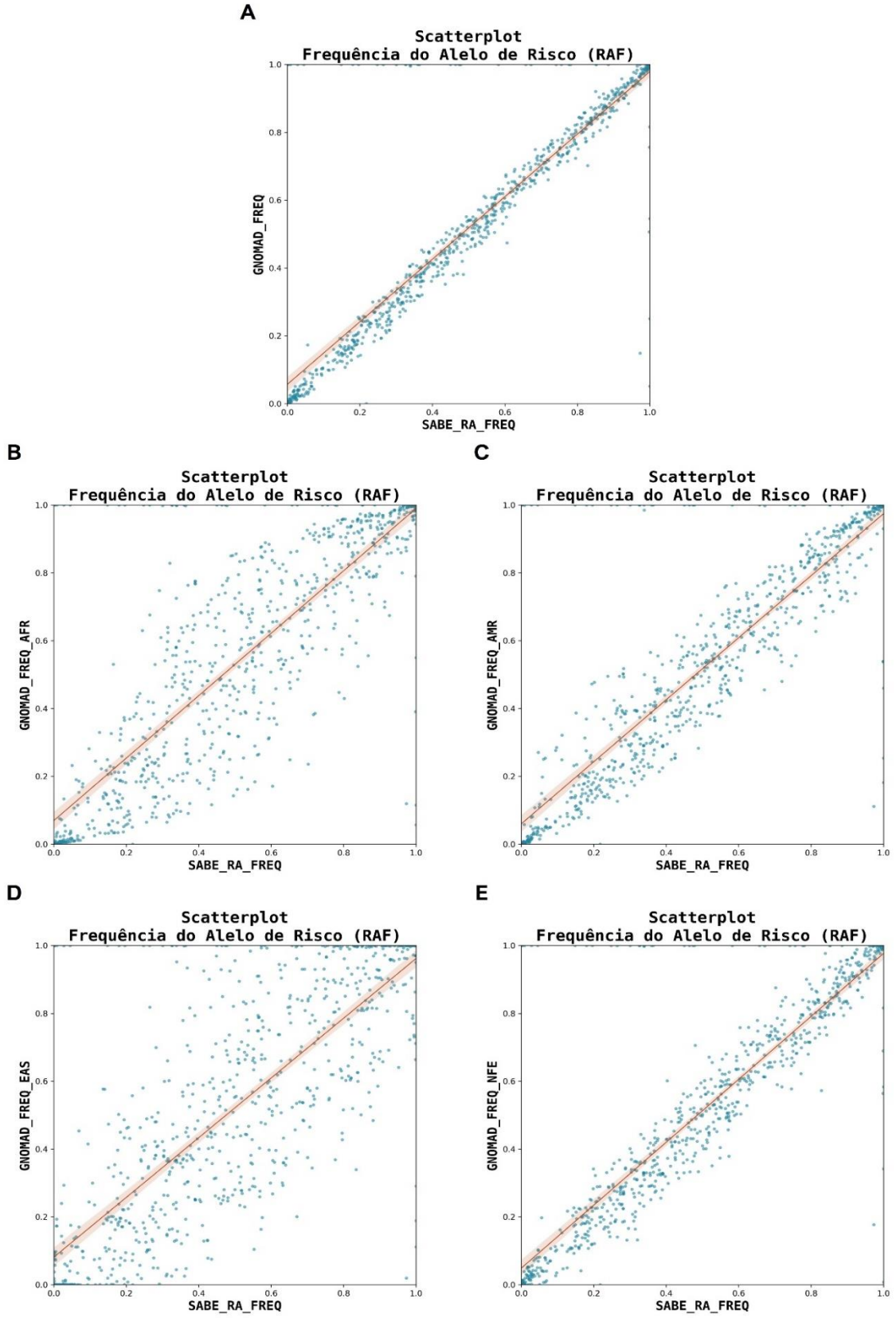

Gráfico 3.5 – Dispersão das frequências dos SNPs do **A)** SABE (SABE_FREQ) e GNOMAD (GNOMAD_FREQ) e das diferentes ancestralidades do GNOMAD, **B)** *African/African American* (GNOMAD_AFR), **C)** *Latino/Admixed American* (GNOMAD_AMR), **D)** *East Asian* (GNOMAD_EAS) e **E)** *Non-Finnish European* (GNOMAD_NFE).

Fonte: Elaborado pelo autor.

Apesar da correlação alta entre as frequências SABE_FREQ e GNOMAD_FREQ, é possível observar, através da frequência relativa, calculada como descrita na seção 3.2.3, que alguns SNPs específicos possuem uma frequência bastante diferente entre essas duas amostras (Gráfico 3.6). Foram identificados 394 SNPs que possuem maior frequência na população do SABE, e 422 SNPs que possuem uma frequência menor, embora 89,58% dos SNPs possuem uma frequência relativa abaixo de 1,50, indicando frequências parecidas para as duas amostras.

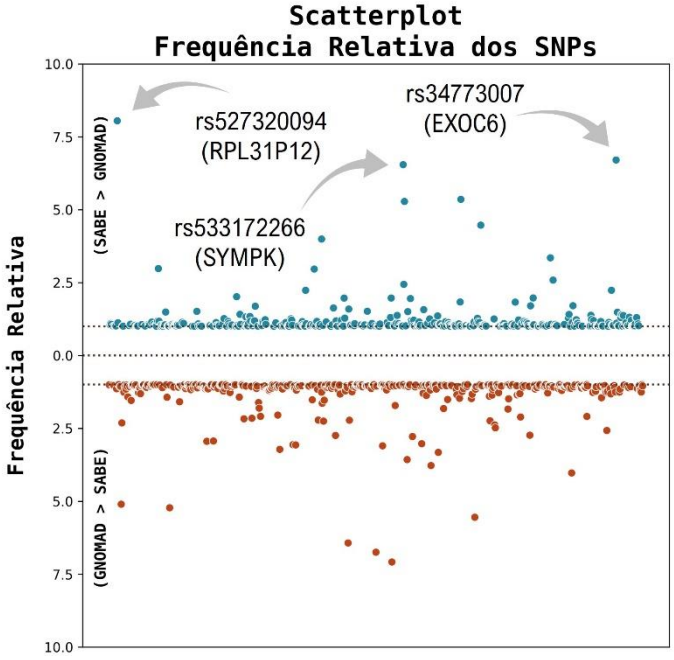

Gráfico 3.6 – Frequência relativa dos SNPs ao comparar SABE_FREQ e GNOMAD_FREQ. A parte superior do gráfico, e em azul, mostra os SNPs que possuem SABE_FREQ maior. A parte inferior, e em laranja, mostra os SNPs que possuem GNOMAD_FREQ maior.
Fonte: Elaborado pelo autor.

É possível destacar 15 SNPs que possuem frequência no SABE pelo menos, duas vezes maior que no GNOMAD (Tabela 3.13).

Tabela 3.13 – SNPs que apresentam os maiores valores de frequência relativa entre SABE_FREQ e GNOMAD_FREQ.

SNP	SABE_FREQ	GNOMAD_FREQ	Frequência Relativa (SABE_FREQ/ GNOMAD_FREQ)
rs527320094	0,002562	0,000318	8,047519
rs533172266	0,000854	0,000127	6,704349
rs34773007	0,973036	0,148711	6,543134
rs576083050	0,000854	0,000160	5,353225
rs565236700	0,000854	0,000162	5,281124
rs557027608	0,000427	0,000096	4,467845
rs6976111	1,000000	0,250511	3,991841
rs551640889	0,000427	0,000128	3,348967
rs184847416	0,000854	0,000287	2,979129
rs551513405	0,001708	0,000576	2,963379
rs199795270	0,008540	0,003304	2,584386
rs78627331	0,001708	0,000701	2,438091
rs745903616	0,000854	0,000382	2,235345
rs78840640	0,024765	0,011088	2,233395
rs17250977	0,045260	0,022455	2,015578

Fonte: Elaborada pelo autor.

3.3.5 Escore de Risco Poligênico

Foi realizada uma análise de PRS nos dados do SABE, para identificar quais SNPs associados ao DM2 possuem uma maior contribuição no risco de manifestação da doença na população brasileira, além de possibilitar a elaboração de um padrão de risco poligênico e outras métricas associadas a herdabilidade e demais fatores de risco.

Os dados de PRS foram retirados do PGS *Catalog*, buscando encontrar como principal característica estudos multiancestrais. Como citado na seção 3.2.4, foram selecionados os PRS0804, PRS2308, PRS2026 e o PRS3443. As quantidades de variantes utilizados para a aplicação de cada PRS no SABE estão descritas na tabela 3.14.

Tabela 3.14 – Quantidade de variantes avaliadas no SABE em comparação a quantidade de variantes registradas nos estudos originais.

PRS	Variantes Estudo Original	Variantes Avaliadas SABE	Percentual SABE/Original
PRS0804	582	576	98,97%
PRS2308	1.259.754	1.258.907	99,93%
PRS2026	830.783	830.507	99,97%
PRS3443	1.092.496	1.085.339	99,34%

Fonte: Elaborada pelo autor.

3.3.5.1 Comparativo dos Estudos

Após selecionar os estudos de PRS, eles foram aplicados na população do SABE, onde foi somado o escore de risco para cada indivíduo. A amostra total, de cada PRS aplicado, foi dividida em percentis e foi verificada a distribuição dos escores em cada um deles (Gráfico 3.7).

Também é possível observar a diferença entre as distribuições dos escores entre o extremo, correspondente ao percentil 90 (P90) e os escores do centro, correspondentes ao valores compreendidos entre os percentis 40 e 60 (P40-P60), além do comparativo do extremo com o restante da amostra (P00-P90), para cada PRS analisado (Gráficos 3.8, 3.9, 3.10 e 3.11).

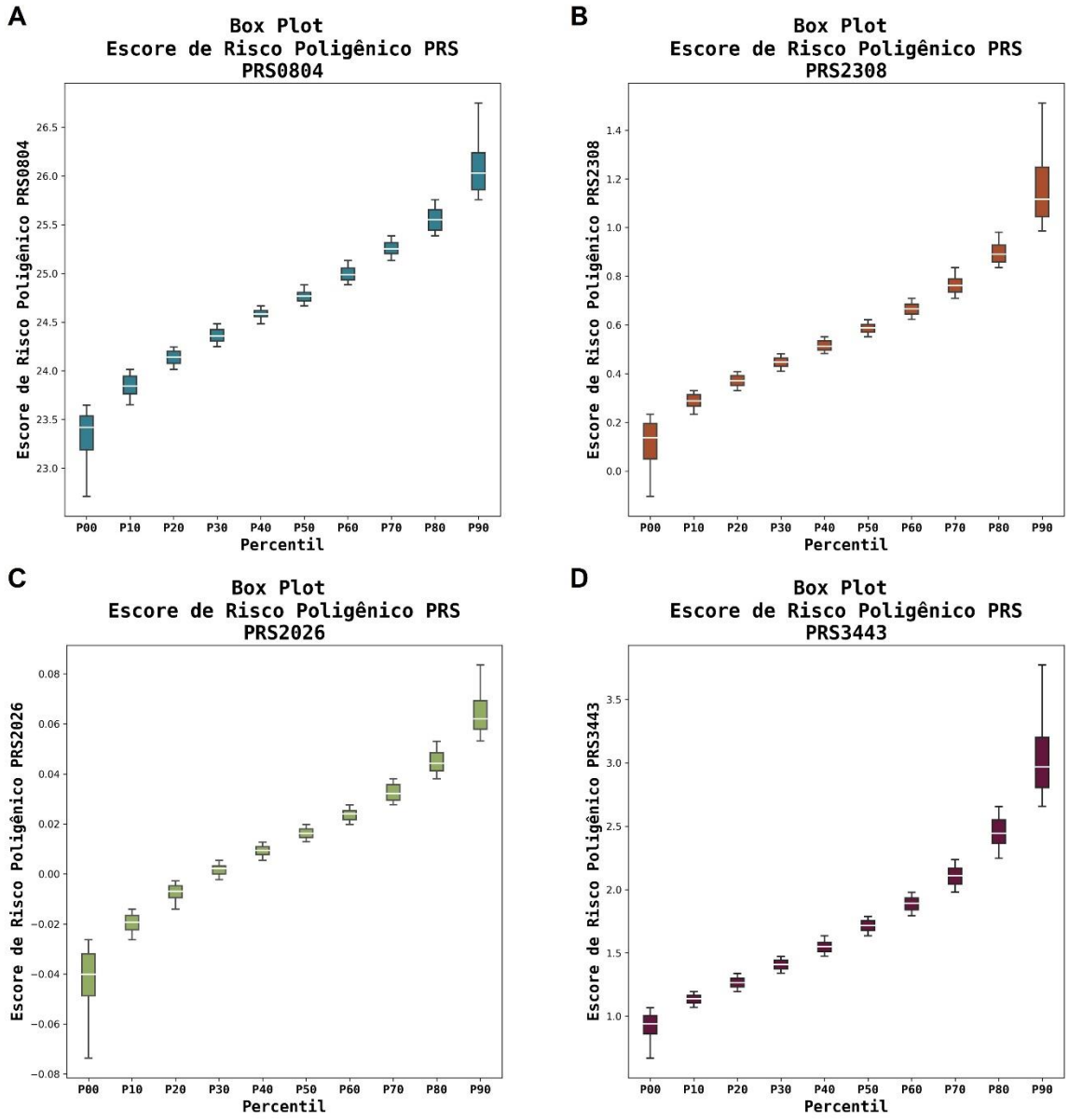

Gráfico 3.7 – Distribuição dos PRS por percentil de PRS. **A)** PRS0804, **B)** PRS2308, **C)** PRS2026 e **D)** PRS3443.

Fonte: Elaborado pelo autor.

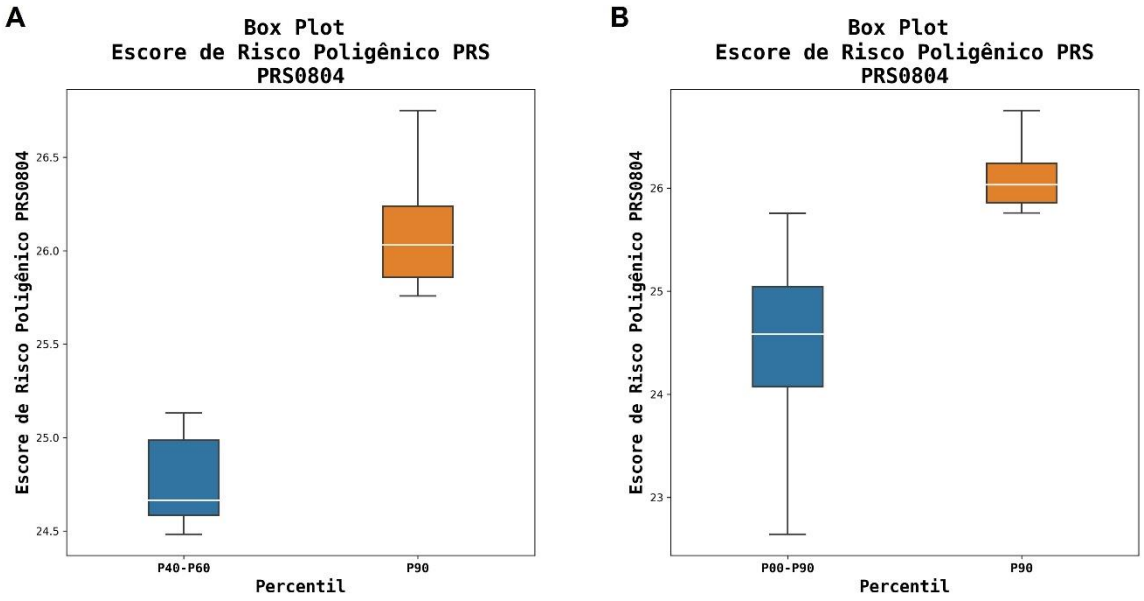

Gráfico 3.8 – Distribuição dos PRS por percentil para o PRS2308. **A)** Comparação do extremo (P90) com o centro (P40-P60) e **B)** Comparação do extremo (P90) com o restante da amostra (P00-P90).
Fonte: Elaborado pelo autor.

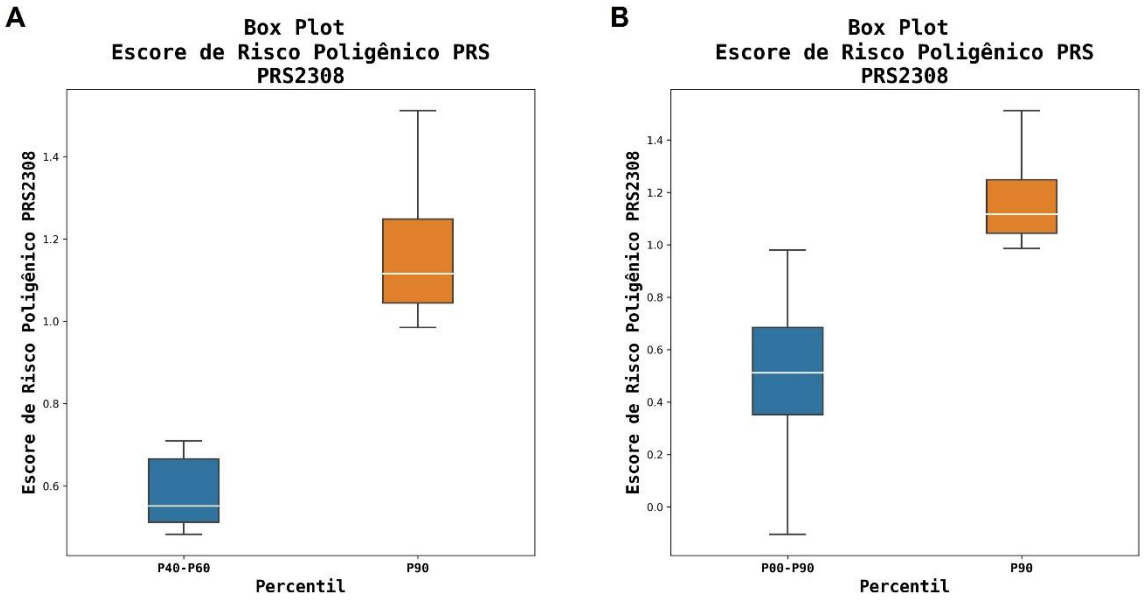

Gráfico 3.9 – Distribuição dos PRS por percentil para o PRS0804. **A)** Comparação do extremo (P90) com o centro (P40-P60) e **B)** Comparação do extremo (P90) com o restante da amostra (P00-P90).
Fonte: Elaborado pelo autor.

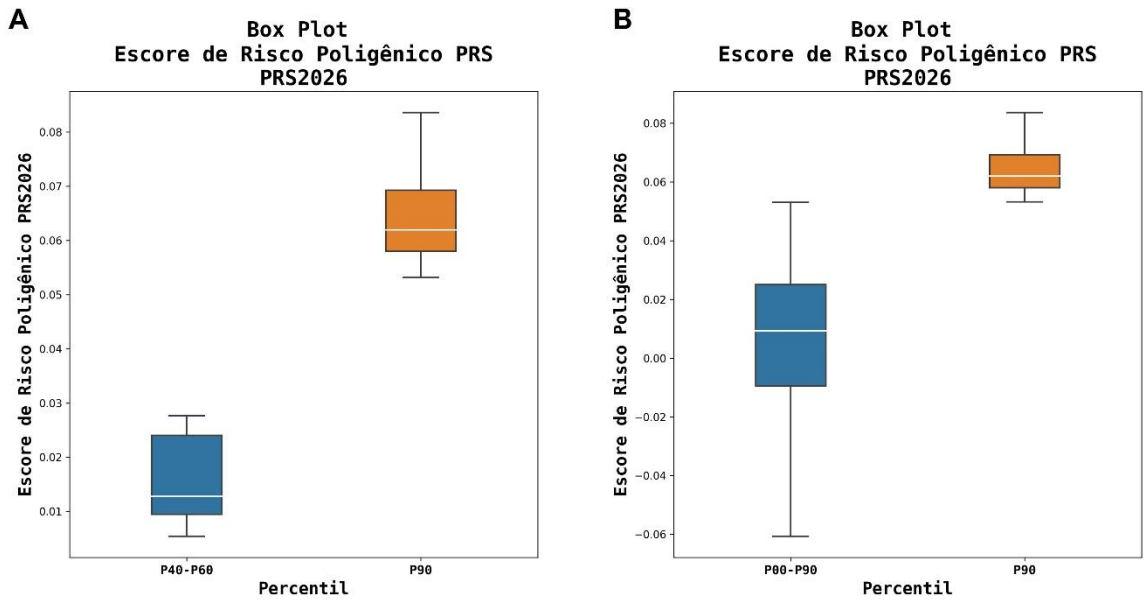

Gráfico 3.10 – Distribuição dos PRS por percentil para o PRS2026. **A)** Comparação do extremo (P90) com o centro (P40-P60) e **B)** Comparação do extremo (P90) com o restante da amostra (P00-P90).
Fonte: Elaborado pelo autor.

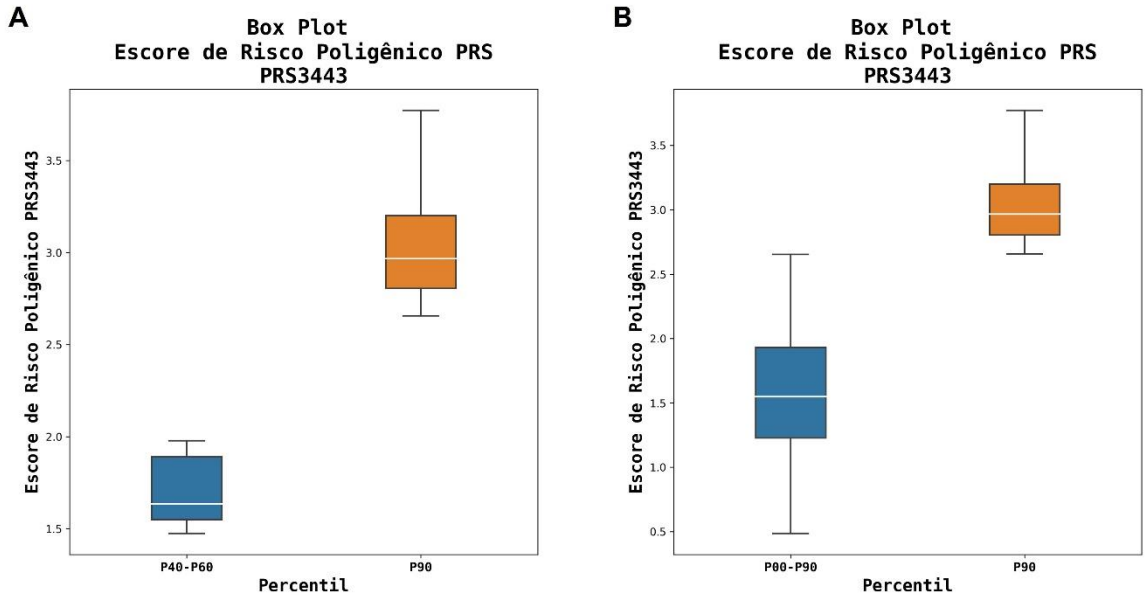

Gráfico 3.11 – Distribuição dos PRS por percentil para o PRS3443. **A)** Comparação do extremo (P90) com o centro (P40-P60) e **B)** Comparação do extremo (P90) com o restante da amostra (P00-P90).
Fonte: Elaborado pelo autor.

Para verificar os aumento do risco de DM2 considerando os extremos dos percentis de risco, foi calculado o OR para cada PRS, para duas comparações diferentes. A primeira comparando o extremo (P90) com os valores centrais da amostra (P40-P60), e a segunda comparando o mesmo extremo com o restante total da amostra (P00-P90) (Tabela 3.15).

Tabela 3.15 – *Odds Ratio* indicando o aumento de risco de DM2 em relação ao percentil extremo (P90) e o centro (P40-P60), e o extremo e o restante da amostra.

PRS	P90 x P40-P60	P-Valor	P90 x P00-P90	P-Valor
PGS0804	1,3796	3,74 e-02	1,8897	7,76 e-07
PGS2308	1,6520	1,91 e-03	1,9630	6,93 e-08
PGS2026	1,4823	2,26 e-02	1,7994	4,56 e-06
PGS3443	1,5840	8,00 e-03	1,9630	6,93 e-08

Fonte: Elaborada pelo autor

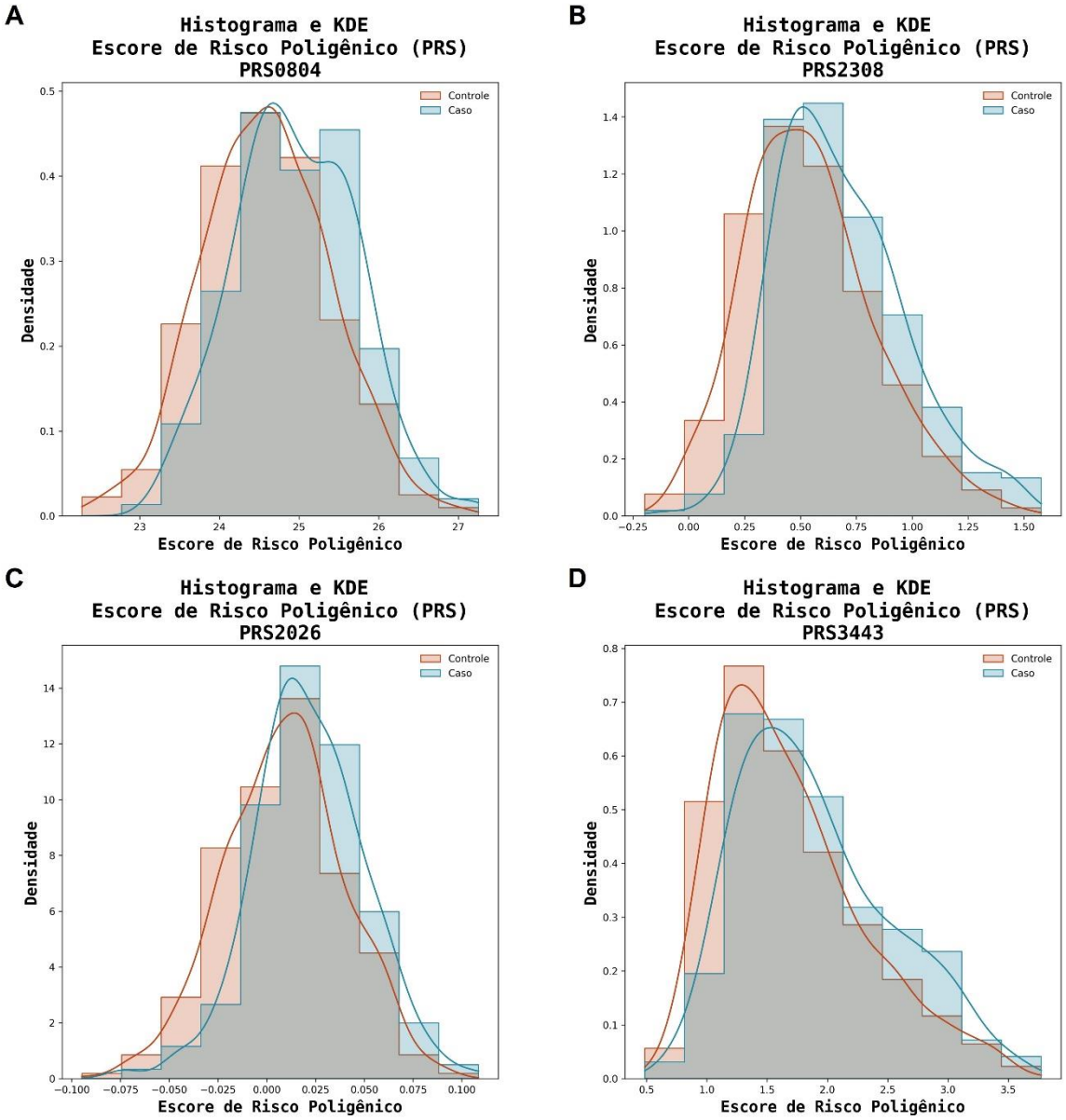

Gráfico 3.12 – Histograma e KDE com a densidade da distribuição dos dois grupos, caso, em azul e controle, em laranja. **A)** PRS0804, **B)** PRS2308, **C)** PRS2026 e **D)** PRS3443.

Fonte: Elaborado pelo autor.

Para cada PRS aplicado, as amostras foram divididas em caso e controle e foi verificada a diferença na distribuição dos escores para esses dois grupos (Gráfico 3.12).

É possível observar um desvio da distribuição de casos para escores maiores, um padrão observado para todos os PRS aplicados. Há uma diferença significativa entre a distribuição dos dois grupos em todos os PRS (Tabela 3.16).

Tabela 3.16 – Valores de P para os testes estatísticos de Mann-Whitney para os grupos caso e controle de cada PRS aplicado.

PRS	**Mann-Whitney P-Valor**
PRS0804	9,50 e-11
PRS2308	1,03 e-12
PRS2026	1,77 e-09
PRS3443	2,59 e-07

Fonte: Elaborada pelo autor

Com a divisão das amostras em caso e controle, foi verificada também a relação entre os percentis de risco de PRS e a prevalência do DM2, para cada PRS analisado (Gráfico 3.13). A prevalência em cada percentil foi calculada como a razão entre o número de casos e a contagem total de indivíduos em cada percentil do escore de risco.

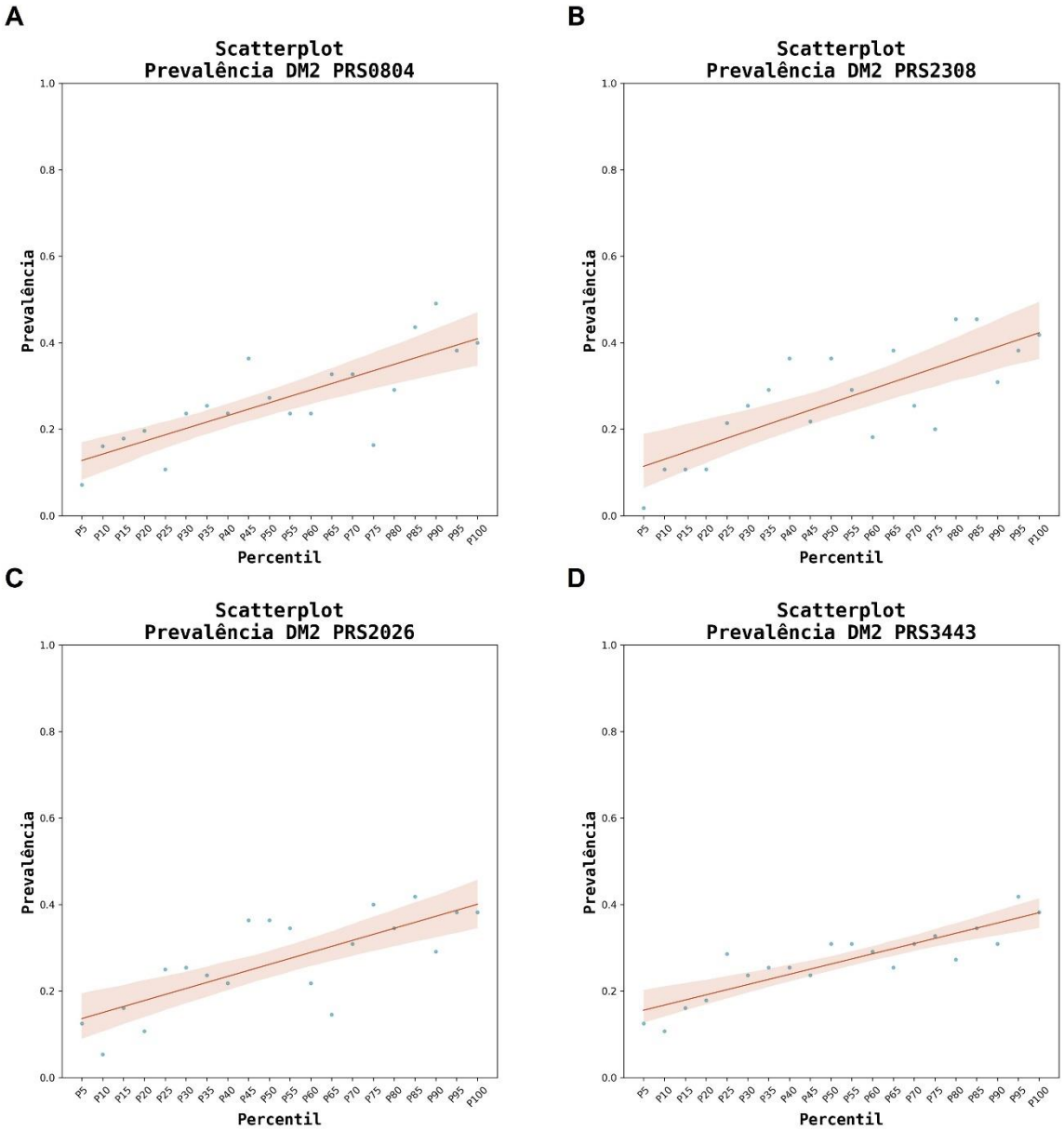

Gráfico 3.13 – Prevalência de DM2 de acordo com 20 percentis de acordo com os escores de risco, com cada percentil representando 5% da amostra total. **A)** PRS0804, **B)** PRS2308, **C)** PRS2026 e **D)** PRS3443.
Fonte: Elaborado pelo autor.

Para verificar a validação desses PRS aplicados no SABE, foi realizada uma análise de curva ROC (*Receiver Operating Characteristic Curve – ROC Curve*) para cada um deles, e foram calculadas as áreas sob a curva (*"area under the curve"* – AUC), indicando quais modelos apresentam um melhor desempenho de PRS de DM2 para a população do SABE (Tabela 3.17).

Tabela 3.17 – Desempenho de cada PRS na população do SABE, medido através de AUC.

PRS	AUC ROC
PRS0804	0,636
PRS2308	0,638
PRS2026	0,640
PRS3443	0,592

Fonte: Elaborada pelo autor

Foi realizada uma análise de curva ROC para verificar o desempenho da aplicação dos estudos de PRS na população do SABE. Os resultados de AUC estão apresentados no Gráfico 3.14.

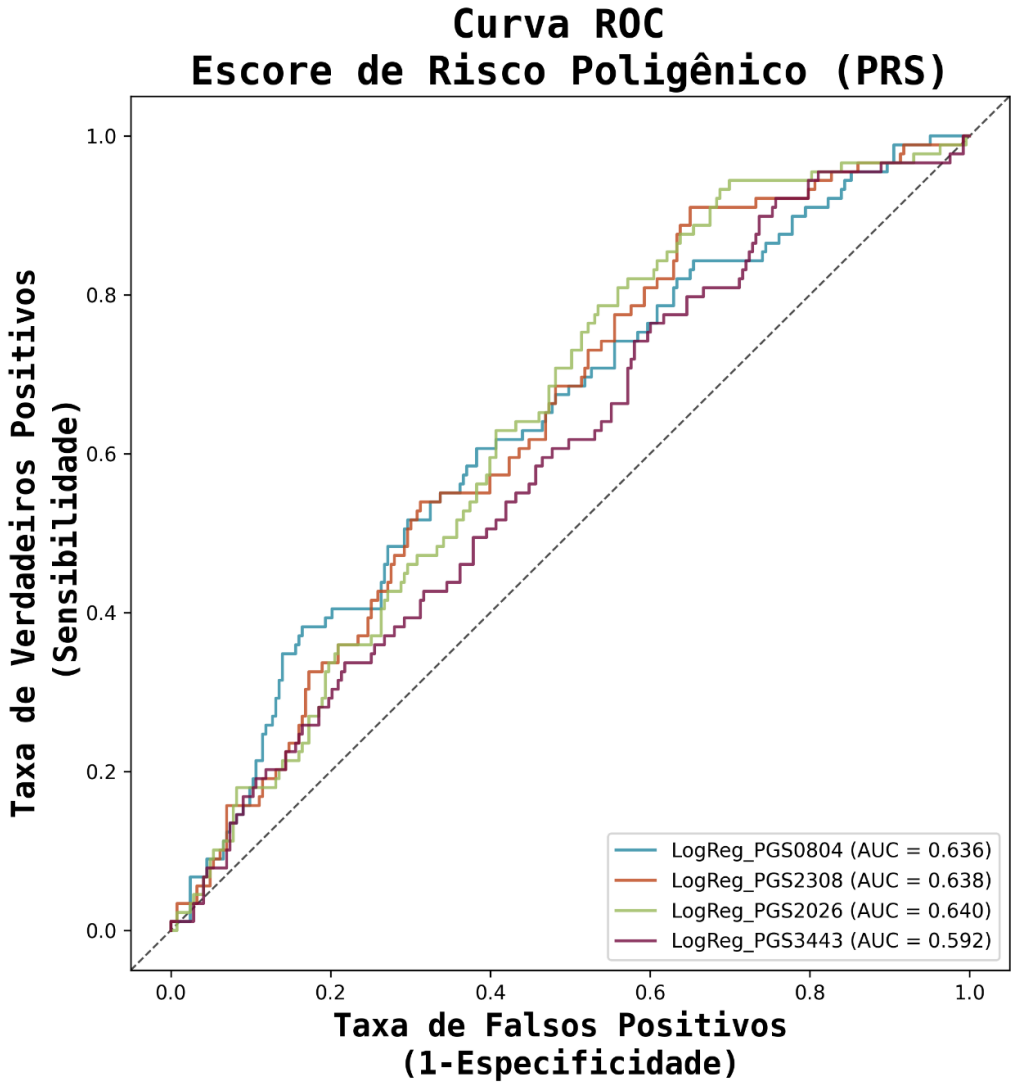

Gráfico 3.14 – Curvas ROC de cada PRS aplicado no SABE e os respectivos valores de AUC.

Fonte: Elaborado pelo autor.

3.4 Discussão e Perspectivas

Os resultados provenientes da análise de variantes raras, apresentados na seção 3.3.3, indicam 41 genes enriquecidos com variantes raras. Desses genes, 22 apresentam variantes sinônimas raras para o grupo GENES_GWAS, associados ao DM2, e 2 genes para o grupo GENES_MODY. Ao considerar as variantes não-sinônimas, relacionadas a perda de função e com potencial deletério (CADD > 20), foram indicados 14 genes para GENES_GWAS e 3 para GENES_MODY. É importante destacar que todos os genes indicados na análise apresentaram um p-valor nominal para esses resultados. Essa significância é perdida ao expandir essa análise para correção de múltiplos testes.

O estudo de variantes raras pode ser valioso. Embora essas variantes não expliquem a maior parte do risco das doenças, elas podem ser importantes para prever o risco individual de quem as possui (145). Foi identificada uma carga significativa de variantes associadas ao MODY entre indivíduos com DM2, podendo implicar em mudanças substanciais no tratamento da doença (146).

O pequeno tamanho da amostra do SABE (1.171 indivíduos) não favorece a identificação, nem a significância estatística da análise. Mesmo um estudo de larga escala para DM2, envolvendo 127.145 indivíduos, encontrou empiricamente apenas 6 variantes raras. Esse número aumenta ao considerar outros modelos simulados, com frações de até 7,1% (76). Em um estudo de altura, foi encontrada uma fração maior de variantes raras (9,7%), considerando uma amostra também maior (711.428) (145). Este fato apoia a hipótese de que variantes raras associadas ao DM2 ou a outras doenças com características poligênicas possam ser descobertas à medida que o tamanho da amostra também aumente.

Para a análise de variantes comuns, descrita na seção 3.3.4, foram validados 63 SNPs (7,96%) descritos em outros GWAS. Esses SNPs apresentaram OR estatisticamente significativo para a população do SABE. Dos SNPs validados, 11 apresentam um direcionamento oposto ao relatado no GWAS original, possuindo caráter protetivo na população brasileira, de acordo com os dados encontrados no SABE. A descoberta desses dados é bastante importante para identificar diferentes suscetibilidades genéticas ao DM2 que podem variar entre populações de ancestralidades diferentes.

Embora a suscetibilidade genética ao DM2 capturada pela variação genética comum é, principalmente, compartilhada entre ancestrais (60), GWAS recentes de DM2 com populações do leste asiático, anteriormente sub-representadas, identificaram 89 novos *loci* de risco ao DM2 (120,147).

Esses estudos mostram que 8,4% das variantes de maior efeito apresentam heterogeneidade significativa no tamanho do efeito entre os resultados do GWAS da Ásia e da Europa, enquanto são comuns ou de baixa frequência em asiáticos orientais, são raras em europeus (60).

Ao analisar as frequências dos alelos de risco em diferentes ancestralidades no SABE e no GNOMAD, foi identificada uma alta correlação entre as distribuições de frequência, fortalecendo a ideia de que a maioria dos SNPs associados ao DM2 apresentam prevalências parecidas entres os grupos populacionais.

Os resultados indicam uma heterogeneidade de 10,42% das variantes, apresentando frequências relativas entre as frequências do SABE e GNOMAD acima de 1,5. Foram destacados 15 SNPs que possuem frequência no SABE, pelo menos, duas vezes maior que no GNOMAD, podendo indicar um maior efeito no risco do DM2.

Com a falta de um GWAS brasileiro, ficou claro que é essencial uma amostra de dados muito maior para obtermos resultados bem mais precisos e completos sobre as variantes analisadas. Essa necessidade e essencialidade de uma amostra populacional maior foi evidenciada quando foi calculado o poder estatístico da amostra.

Foi identificado que, apesar de alguns SNPs apresentarem um p-valor estatisticamente significativo em seu cálculo de OR, eles não apresentaram um alto poder estatístico, sendo vulneráveis a outro tipo de erro ao considerar a relevância desses dados. Com isso, foi calculado o tamanho ideal da amostra para que fosse alcançado um poder estatístico de 0,8 (80%), que é o limiar utilizado em diversos estudos populacionais. Nossos resultados indicam que a grande maioria dos SNPs deveriam ser analisados em amostras, pelo menos, 100 ou 1000 vezes maiores. Considerando o tamanho atual do SABE, seria desejável uma amostra entre, aproximadamente, 100 mil e 1 milhão de indivíduos, respeitando também a proporção entre casos e controles.

Ao aplicar os estudos de PRS na população do SABE, foi verificado o desempenho de cada um deles. Ao verificar os resultados obtidos no estudo original do PRS0804 (142), é possível observar que foi obtido uma AUC de 0,568 quando validado em africanos, 0,825 em europeus, 0,625 em hispânicos e latino-americanos e 0,626 em asiáticos. Para esse mesmo PRS, foi obtido para o SABE um desempenho de 63,6% (AUC 0,636) utilizando 576 SNPs (98,97%). Este resultado ficou abaixo apenas do modelo validado na população europeia.

O modelo utilizado no PRS2308 (129) inclui as covariáveis de idade, sexo e os percentis extremos. Os resultados apresentam uma AUC que varia de 0,631 para alguns grupos com ancestralidade africana, até 0,851 para um grupo de ancestralidades hispânicas e latino-

americanas. O desempenho do modelo para esse PRS foi de 63,8% (AUC 0,638), analisando 1.258.907 SNPs (99,93%). É possível comparar o desempenho deste modelo com os validados em algumas amostras de ancestralidade africana. Porém, ao comparar com outras ancestralidades, foi apresentado um desempenho menor. É válido ressaltar que não utilizamos covariáveis no modelo desenvolvido.

O PRS2026 (143) apresentou seus resultados utilizando o método de correlação parcial, com covariáveis de sexo, idade, data de nascimento, e índice de privação, não apresentando valores de desempenho baseados em AUC. Embora tenha sido o PRS que apresentou o melhor desempenho na análise, utilizando 830.507 SNPs (99,97%), não é possível comparar de maneira direta com os resultados apresentados no estudo original.

A validação do PRS3443 foi realizada em uma população com ancestralidade hispânica e latino-americana, apresentando uma AUC de 0,747, ao considerar covariáveis de sexo, idade e outros pesos específicos. Foram utilizados 1.085.339 SNPs (99,34%), o modelo apresentou um desempenho de 59,2% (AUC 0,592), abaixo dos resultados encontrados no estudo original.

Embora alguns desempenhos dos modelos de aplicação de PRS possuam resultados consideravelmente abaixo dos estudos originais, os resultados representam um avanço positivo para estudos na população brasileira. É preciso considerar, primeiramente, o tamanho e o perfil da amostra avaliada. A amostra do SABE, além de possuir um tamanho limitado, trata-se de um grupo de indivíduos idosos, com idade média de 71,86 anos.

Os PRS de DM2 foram mais amplamente desenvolvidos e validados em indivíduos de ascendência europeia. Considerando que o desempenho preditivo do PRS muitas vezes é atenuado nessa população (141; 148), e que os outros grupos de ancestralidade estão experimentando taxas crescentes contínuas de DM2 (129), é extremamente importante validar a aplicação de um PRS de DM2 em outras populações antes que possam ser implementados em ambientes clínicos.

4 CONCLUSÕES

Os estudos realizados, nesta tese, sobre as variantes genéticas associadas ao DM2 revelaram padrões de características dessas variantes a respeito de diversos parâmetros avaliados.

Ao explorar a essencialidade dos éxons próximos aos SNPs associados ao DM2, observa-se tendências sugestivas de que esses éxons podem estar localizados em regiões genômicas mais permissivas. Entretanto, deve-se reconhecer algumas limitações nos métodos adotados, como a quantidade de espécies analisadas e a necessidade de aprimorar a abordagem para uma categorização mais precisa dos éxons em relação à localização dos SNPs.

Além disso, a análise das variantes raras ressaltou a importância de amostras populacionais significativas para identificar esses tipos de variantes associadas ao DM2, uma vez que o tamanho da amostra do SABE pode não ser suficiente para obter resultados estatisticamente significativos. Embora as variantes raras não possam explicar a maior parte do risco do DM2, elas desempenham um papel crucial na previsão do risco individual.

Para mais, a análise das variantes comuns revelou a existência de SNPs associados ao DM2 com características diferentes das relatadas em estudos anteriores, indicando a importância de considerar variações genéticas em diferentes populações. As descobertas também destacam a necessidade de uma amostra populacional brasileira mais robusta para obter resultados mais precisos e completos sobre as variantes analisadas.

Por fim, ao aplicar PRS na população do SABE, foram obtidos resultados que indicam um progresso significativo na adaptação desses modelos para a população brasileira. É importante notar que, embora os resultados possam estar abaixo dos estudos originais, representam um passo positivo em direção à validação e aplicação de PRS em diferentes grupos populacionais.

Estas descobertas enfatizam a complexidade da genética do DM2 e a necessidade de considerar as particularidades das populações ao desenvolver estratégias de prevenção e tratamento. Além disso, ressaltam a importância de expandir as amostras e abordagens metodológicas para obter uma compreensão mais abrangente sobre a etiologia do DM2. À medida que a base genética dessa doença continua a ser estudada, novas descobertas podem abrir caminho para abordagens mais personalizadas e eficazes no prognóstico, diagnóstico e tratamento do DM2 em diferentes grupos étnicos e populacionais.

REFERÊNCIAS

1 FOROUHI, N. G.; WAREHAM, N. J.. Epidemiology of diabetes. **Medicine**, v. 47, n. 1, p. 22-27, Jan. 2019. DOI: 10.1016/j.mpmed.2018.10.004.

2 ELSAYED, N. A. *et al*. 2. Classification and diagnosis of diabetes: standards of care in diabetes: 2023. **Diabetes Care**, v. 46, n. 1, p. 19-40, 12 dez. 2022. DOI: 10.2337/dc23-s002.

3 SUN, H. *et al*. IDF Diabetes Atlas: global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. **Diabetes Research and Clinical Practice**, v. 183, p. 109119, Jan. 2022. DOI: 10.1016/j.diabres.2021.109119.

4 DEFRONZO, R. A. *et al*. Type 2 diabetes mellitus. **Nature Reviews Disease Primers**, v. 1, n. 1, p. 1-22, 23 July 2015. DOI: 10.1038/nrdp.2015.19.

5 GALICIA-GARCIA, U. *et al*. Pathophysiology of type 2 diabetes mellitus. **International Journal of Molecular Sciences**, v. 21, n. 17, p. 6275, 30 Aug. 2020. DOI: 10.3390/ijms21176275.

6 HARDING, J. L. *et al*. Global trends in diabetes complications: a review of current evidence. **Diabetologia**, v. 62, n. 1, p. 3-16, 31 Aug. 2018. DOI: 10.1007/s00125-018-4711-2.

7 MANSOUR, A. *et al*. Microvascular and macrovascular complications of type 2 diabetes mellitus: exome wide association analyses. **Frontiers in Endocrinology**, v. 14, n. 1, p. 1-11, 23 mar. 2023. DOI: 10.3389/fendo.2023.1143067.

8 SZMUILOWICZ, E. D.; JOSEFSON, J. L.; METZGER, B. E. Gestational diabetes mellitus. **Endocrinology and Metabolism Clinics of North America**, v. 48, n. 3, p. 479-493, Sept. 2019. DOI: 10.1016/j.ecl.2019.05.001.

9 LYNAM, A. L *et al*. Logistic regression has similar performance to optimised machine learning algorithms in a clinical setting: application to the discrimination between type 1 and type 2 diabetes in young adults. **Diagnostic and Prognostic Research**, v. 4, n. 1, p. 4-6, 4 June 2020. DOI: 10.1186/s41512-020-00075-2.

10 SHERWANI, S. I. *et al*. Significance of HbA1c test in diagnosis and prognosis of diabetic patients. **Biomarker Insights**, v. 11, p. 38440, Jan. 2016. DOI: 10.4137/bmi.s38440.

11 KAIAFA, G. *et al*. Is HbA1c an ideal biomarker of well-controlled diabetes? **Postgraduate Medical Journal**, v. 97, n. 1148, p. 380-383, 10 Sept. 2020. DOI: 10.1136/postgradmedj-2020-138756.

12 CHOBOT, A. *et al*. Obesity and diabetes-not only a simple link between two epidemics. **Diabetes/Metabolism Research and Reviews**, v. 34, n. 7, p. 3042, 17 July 2018. DOI: 10.1002/dmrr.3042.

13 TINAJERO, M. G. *et al*. An update on the epidemiology of type 2 diabetes. **Endocrinology and Metabolism Clinics of North America**, v. 50, n. 3, p. 337-355, Sept. 2021. DOI: 10.1016/j.ecl.2021.05.013.

14 ASCHNER, P. *et al*. The international diabetes federation's guide for diabetes epidemiological studies. **Diabetes Research and Clinical Practice**, v. 172, p. 108630, Feb. 2021. DOI: 10.1016/j.diabres.2020.108630.

15 JING, X *et al*. Related factors of quality of life of type 2 diabetes patients: a systematic review and meta-analysis. **Health and Quality of Life Outcomes**, v. 16, n. 1, p. 1-14, 19 Sept. 2018. DOI: 10.1186/s12955-018-1021-9.

16 PARK, J. J. Epidemiology, pathophysiology, diagnosis and treatment of heart failure in diabetes. **Diabetes & Metabolism Journal**, v. 45, n. 2, p. 146-157, 31 Mar. 2021. DOI: 10.4093/dmj.2020.0282.

17 CORRER, C. J. *et al*. Prevalence of people at risk of developing type 2 diabetes mellitus and the involvement of community pharmacies in a national screening campaign: a pioneer action in Brazil. **Diabetology & Metabolic Syndrome**, v. 12, n. 1, p. 1-11, 8 out. 2020. Springer Science and Business Media LLC. DOI: 10.1186/s13098-020-00593-5.

18 DUNCAN, B. B. *et al*. The burden of diabetes and hyperglycemia in Brazil: a global burden of disease study 2017. **Population Health Metrics**, v. 18, n. 1, p. 1-11, Sept. 2020. DOI: 10.1186/s12963-020-00209-0.

19 MALTA, D. C. *et al*. Prevalência de diabetes mellitus determinada pela hemoglobina glicada na população adulta brasileira, Pesquisa Nacional de Saúde. **Revista Brasileira de Epidemiologia**, v. 22, n. 2, p. 1-13, 2019. DOI: 10.1590/1980-549720190006.supl.2.

20 SCHMIDT, M. I. *et al*. High prevalence of diabetes and intermediate hyperglycemia – The Brazilian longitudinal study of adult health (ELSA-Brasil). **Diabetology & Metabolic Syndrome**, v. 6, n. 1, p. 123-131, 18 Nov. 2014. DOI: 10.1186/1758-5996-6-123.

21 MENDES, A. B. V. *et al*. Prevalence and correlates of inadequate glycaemic control: results from a nationwide survey in 6,671 adults with diabetes in Brazil. **Acta Diabetologica**, v. 47, n. 2, p. 137-145, 5 Aug. 2009. DOI: 10.1007/s00592-009-0138-z.

22 VAN TILBURG, J. *et al*. Defining the genetic contribution of type 2 diabetes mellitus. **Journal of Medical Genetics**, v. 38, n. 9, p. 569-578, 1 Sept. 2001. DOI: 10.1136/jmg.38.9.569.

23 SPEAKMAN, J R. Thrifty genes for obesity, an attractive but flawed idea, and an alternative perspective: the 'drifty gene' hypothesis. **International Journal of Obesity**, v. 32, n. 11, p.1611-1617, 14 Oct. 2008. DOI: 10.1038/ijo.2008.161.

24 SÉGUREL, L. *et al*. Positive selection of protective variants for type 2 diabetes from the Neolithic onward: a case study in central Asia. **European Journal of Human Genetics**, v. 21, n. 10, p. 1146-1151, 23 Jan. 2013. DOI: 10.1038/ejhg.2012.295.

25 WISE, P. H. Positive selection of type 2 diabetes genotypes – the glycaemic threshold hypothesis. **Medical Hypotheses**, v. 127, p. 150-153, June 2019. DOI: 10.1016/j.mehy.2019.04.014.

26 PADILLA-MARTÍNEZ, F. *et al*. Systematic review of polygenic risk scores for type 1 and type 2 diabetes. **International Journal of Molecular Sciences**, v. 21, n. 5, p. 1703, 2 Mar. 2020. DOI: 10.3390/ijms21051703.

27 MAI, T.T.; TURNER, P.; CORANDER, J. Boosting heritability: estimating the genetic component of phenotypic variation with multiple sample splitting. **BMC Bioinformatics**, v. 22, n. 1, p. 1-16, 27 Mar. 2021. DOI: 10.1186/s12859-021-04079-7.

28 WILLEMSEN, G. *et al*. The concordance and heritability of type 2 diabetes in 34,166 twin pairs from international twin registers: the discordant twin (discotwin) consortium. **Twin Research and Human Genetics**, v. 18, n. 6, p. 762-771, Dec. 2015. DOI: 10.1017/thg.2015.83.

29 AVERY, A. R.; DUNCAN, G. E. Heritability of Type 2 Diabetes in the Washington State Twin Registry. **Twin Research and Human Genetics**, v. 22, n. 2, p. 95-98, Apr. 2019. DOI: 10.1017/thg.2019.11.

30 HEJASE, H. A.; DUKLER, N.; SIEPEL, A.. From Summary Statistics to Gene Trees: methods for inferring positive selection. **Trends in Genetics**, v. 36, n. 4, p. 243-258, Apr. 2020. DOI: 10.1016/j.tig.2019.12.008.

31 MEEKS, K. A C *et al*. Evolutionary forces in diabetes and hypertension pathogenesis in Africans. **Human Molecular Genetics**, v. 30, n. 1, p. 110-118, 1 Mar. 2021. DOI: 10.1093/hmg/ddaa238.

32 NEEL, J. V. Diabetes Mellitus: a “Thrifty” genotype rendered detrimental by “progress”? **American Journal of Human Genetics**, v. 14, n. 4, p. 353-362, Dec. 1962.

33 BINDON, J. R.; BAKER, P. T. Bergmann's rule and the thrifty genotype. **American Journal of Physical Anthropology**, v. 104, n. 2, p. 201-210, Oct. 1997. DOI: 10.1002/(sici)1096-8644(199710)104:23.0.co;2-0.

34 JOFFE, B.; ZIMMET, P. The thrifty genotype in type 2 diabetes: an unfinished symphony moving to its finale? **Endocrine**, v. 9, n. 2, p. 139-142, 1998. DOI: 10.1385/endo:9:2:139.

35 PRENTICE, A. Early influences on human energy regulation: thrifty genotypes and thrifty phenotypes. **Physiology & Behavior**, v. 86, n. 5, p.640-645, 15 Dec. 2005. DOI: 10.1016/j.physbeh.2005.08.055.

36 HALES, C.; BARKER, D. Type 2 (non-insulin-dependent) diabetes mellitus: the thrifty phenotype hypothesis. **International Journal of Epidemiology**, v. 42, n. 5, p. 1215-1222, 1 Oct. 2013. DOI: 10.1093/ije/dyt133.

37 AYUB, Q. *et al*. Revisiting the thrifty gene hypothesis via 65 loci associated with susceptibility to type 2 diabetes. **American Journal of Human Genetics**, v. 94, n. 2, p. 176-185, Feb. 2014. DOI: 10.1016/j.ajhg.2013.12.010.

38 VOS, T. *et al*. Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990–2013: a

systematic analysis for the global burden of disease study 2013. **Lancet**, v. 386, n. 9995, p. 743-800, Aug. 2015. DOI: 10.1016/s0140-6736(15)60692-4.

39 ZIMMET, P.; ALBERTI, K. G. M. M.; SHAW, J. Global and societal implications of the diabetes epidemic. **Nature**, v. 414, n. 6865, p.782-787, 13 Dec. 2001. DOI: 10.1038/414782a.

40 BOWDIN, S. *et al*. The genome clinic: a multidisciplinary approach to assessing the opportunities and challenges of integrating genomic analysis into clinical care. **Human Mutation**, v. 35, n. 5, p.513-519, 7 Apr. 2014. DOI: 10.1002/humu.22536.

41 HOWRIGAN, D. J.; WILLIAM, N.; DARLINGTON, T. M. Complex multifactorial genetic diseases. **Els**, p. 1-11, 22 Jan. 2018. DOI: 10.1002/9780470015902.a0001881.pub3.

42 EDWARDS, S. L. *et al*. Beyond GWASs: illuminating the dark road from association to function. **American Journal of Human Genetics**, v. 93, n. 5, p. 779-797, Nov. 2013. DOI: 10.1016/j.ajhg.2013.10.012.

43 GROTZ, A. K.; GLOYN, A. L.; THOMSEN, S. K. Prioritising causal genes at type 2 diabetes risk loci. current **Diabetes Reports**, v. 17, n. 9, p.76-84, 31 July 2017. DOI: 10.1007/s11892-017-0907-y.

44 UFFELMANN, E. *et al*. Genome-wide association studies. **Nature Reviews Methods Primers**, v. 1, n. 1, p. 1-21, 26 Aug. 2021. DOI: 10.1038/s43586-021-00056-9.

45 BUSH, W. S.; MOORE, J. H. Chapter 11: genome-wide association studies. **PLoS Computational Biology**, v. 8, n. 12, p. 1002822, 27 Dec. 2012. DOI: 10.1371/journal.pcbi.1002822.

46 PAL, L. R *et al*. Insights from GWAS: emerging landscape of mechanisms underlying complex trait disease. **BMC Genomics**, v. 16, n. 8, p. 1-15, 18 June 2015. DOI: 10.1186/1471-2164-16-S8-S4.

47 SCHRIML, L. M. *et al*. Modeling the enigma of complex disease etiology. **Journal of Translational Medicine**, v. 21, n. 1, p. 1-14, 2023. DOI: 10.1186/s12967-023-03987-x.

48 WELTER, D. *et al*. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. **Nucleic Acids Research**, v. 42, n. 1, p. 1001-1006, 6 Dec. 2013. DOI: 10.1093/nar/gkt1229.

49 SOLLIS, E. *et al*. The NHGRI-EBI GWAS catalog: knowledgebase and deposition resource. **Nucleic Acids Research**, v. 51, n. 1, p. 977-985, 9 Nov. 2022. DOI: 10.1093/nar/gkac1010.

50 SAKAUE, S. *et al*. A cross-population atlas of genetic associations for 220 human phenotypes. **Nature Genetics**, v. 53, n. 10, p. 1415-1424, 30 Sept. 2021. DOI: /10.1038/s41588-021-00931-x.

51 HIRSCHHORN, J. N.; DALY, M. J. Genome-wide association studies for common diseases and complex traits. **Nature Reviews Genetics**, v. 6, n. 2, p. 95-108, Feb. 2005. DOI: 10.1038/nrg1521.

52 LAAKSO, M.; SILVA, L. F.. Genetics of type 2 diabetes: past, present, and future. **Nutrients,** v. 14, n. 15, p. 3201, 4 Aug. 2022. DOI: 10.3390/nu14153201.

53 SAXENA, R. *et al*. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. **Science**, v. 316, n. 5829, p. 1331-1336, June 2007. DOI: 10.1126/science.1142358.

54 SLADEK, R. *et al*. A genome-wide association study identifies novel risk loci for type 2 diabetes. **Nature**, v. 445, n. 7130, p. 881-885, Feb. 2007. DOI: 10.1038/nature05616.

55 ZEGGINI, E. *et al*. Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. **Science**, v. 316, n. 5829, p. 1336-1341, June 2007. DOI: 10.1126/science.1142364.

56 MORRIS, A. P *et al*. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. **Nature Genetics**, v. 44, n. 9, p. 981-990, 12 Aug. 2012. DOI: 10.1038/ng.2383.

57 INGELSSON, E. *et al*. Detailed physiologic characterization reveals diverse mechanisms for novel genetic loci regulating glucose and insulin metabolism in humans. **Diabetes**, v. 59, n. 5, p. 1266-1275, 25 Feb. 2010. DOI: 10.2337/db09-1568.

58 VOIGHT, B. F *et al*. Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. **Nature Genetics**, v. 42, n. 7, p. 579-589, 27 June 2010. DOI: 10.1038/ng.609.

59 SCOTT, R. A. *et al*. An expanded genome-wide association study of type 2 diabetes in europeans. **Diabetes**, v. 66, n. 11, p. 2888-2902, 31 May 2017. DOI: 10.2337/db16-1253.

60 DEFOREST, N.; MAJITHIA, A. R. Genetics of Type 2 Diabetes: implications from large-scale studies. **Current Diabetes Reports**, v. 22, n. 5, p. 227-235, 19 Mar. 2022. DOI: 10.1007/s11892-022-01462-3.

61 PIERCE, S. E. *et al*. Post-GWAS knowledge gap: the how, where, and when. **Npj Parkinson'S Disease**, v. 6, n. 1, p. 1-5, 9 Sept. 2020. DOI: 10.1038/s41531-020-00125-y.

62 BARTONICEK, N. *et al*. Intergenic disease-associated regions are abundant in novel transcripts. **Genome Biology**, v. 18, n. 1, p. 1-16, Dec. 2017. DOI: 10.1186/s13059-017-1363-3.

63 MEIGS, J. B. The genetic epidemiology of type 2 diabetes: opportunities for health translation. **Current Diabetes Reports**, v. 19, n. 8, p. 1-8, 22 July 2019. DOI: 10.1007/s11892-019-1173-y.

64 PAGANI, F.; BARALLE, F. E. Opinion: genomic variants in exons and introns. **Nature Reviews Genetics**, v. 5, n. 5, p.389-396, May 2004. DOI: 10.1038/nrg1327.

65 STROEVE, J. H. M. *et al*. Phenotypic flexibility as a measure of health: the optimal nutritional stress response test. **Genes & Nutrition**, v. 10, n. 3, p. 1-21, 21 Apr. 2015. DOI: 10.1007/s12263-015-0459-1.

66 DLAMINI, Z.; MOKOENA, F.; HULL, R. Abnormalities in alternative splicing in diabetes: therapeutic targets. **Journal of Molecular Endocrinology**, v. 59, n. 2, p. 93-107, Aug. 2017. DOI: 10.1530/jme-17-0049.

67 MERCADER, J. M. *et al*. A Loss-of-function splice acceptor variant in IGF2 Is protective for type 2 diabetes. **Diabetes**, v. 66, n. 11, p. 2903-2914, 24 Aug. 2017. DOI: 10.2337/db17-0187.

68 LEE, S. C.; ABDEL-WAHAB, O. Therapeutic targeting of splicing in cancer. **Nature Medicine**, v. 22, n. 9, p. 976-986, Sept. 2016. DOI: 10.1038/nm.4165.

69 GALLEGO-PAEZ, L. M. *et al*. Alternative splicing: the pledge, the turn, and the prestige. **Human Genetics**, v. 136, n. 9, p. 1015-1042, 3 Apr. 2017. DOI: 10.1007/s00439-017-1790-y.

70 DELGADO-LISTA, J. *et al*. CORonary diet intervention with olive oil and cardiovascular PREVention study (the CORDIOPREV study): rationale, methods, and baseline characteristics. **American Heart Journal**, v. 177, p. 42-50, July 2016. DOI: 10.1016/j.ahj.2016.04.011.

71 GAHETE, M. D. *et al*. Changes in splicing machinery components influence, precede, and early predict the development of type 2 diabetes: from the cordioprev study. **Ebiomedicine**, v. 37, p. 356-365, Nov. 2018. DOI: 10.1016/j.ebiom.2018.10.056.

72 SAXENA, A. *et al*. Whole transcriptome RNA-seq reveals key regulatory factors involved in type 2 diabetes pathology in peripheral fat of Asian Indians. **Scientific Reports**, v. 11, n. 1, p. 1-10, 20 May 2021. DOI: 10.1038/s41598-021-90148-z.

73 JENKINSON, C. P. *et al*. Transcriptomics in type 2 diabetes: bridging the gap between genotype and phenotype. **Genomics Data**, v. 8, p. 25-36, June 2016. DOI: 10.1016/j.gdata.2015.12.001.

74 VIÑUELA, A. *et al*. Genetic variant effects on gene expression in human pancreatic islets and their implications for T2D. **Nature Communications**, v. 11, n. 1, p. 1-14, 30 Sept. 2020. DOI: 10.1038/s41467-020-18581-8.

75 GAMAZON, E. R. *et al*. Using an atlas of gene regulation across 44 human tissues to inform complex disease- and trait-associated variation. **Nature Genetics**, v. 50, n. 7, p. 956-967, 28 June 2018. DOI: 10.1038/s41588-018-0154-4.

76 FUCHSBERGER, C. *et al*. The genetic architecture of type 2 diabetes. **Nature**, v. 536, n. 7614, p. 41-47, 11 July 2016. DOI: 10.1038/nature18642.

77 WANG, Y. *et al*. A crowdsourcing open platform for literature curation in UniProt. **PLoS Biology**, v. 19, n. 12, p. 3001464, 6 Dec. 2021. DOI: 10.1371/journal.pbio.3001464.

78 BUCHFINK, B.; REUTER, K.; DROST, H. Sensitive protein alignments at tree-of-life scale using DIAMOND. **Nature Methods**, v. 18, n. 4, p. 366-368, Apr. 2021. DOI: 10.1038/s41592-021-01101-x.

79 PERSSON, E.; SONNHAMMER, E. L L. InParanoid-DIAMOND: faster orthology analysis with the inparanoid algorithm. **Bioinformatics**, v. 38, n. 10, p. 2918-2919, 31 Mar. 2022. DOI: 10.1093/bioinformatics/btac194.

80 PERSSON, E.; SONNHAMMER, E. L.L. InParanoiDB 9: ortholog groups for protein domains and full-length proteins. **Journal Of Molecular Biology**, v. 435, n. 14, p. 168001, July 2023. DOI: 10.1016/j.jmb.2023.168001.

81 ALTSCHUL, S. F. *et al*. Basic local alignment search tool. **Journal of Molecular Biology**, v. 215, n. 3, p. 403-410, Oct. 1990. DOI: 10.1016/s0022-2836(05)80360-2.

82 ALTSCHUL, S. *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. **Nucleic Acids Research**, v. 25, n. 17, p. 3389-3402, 1 Sept. 1997. DOI: 10.1093/nar/25.17.3389.

83 FRANKISH, A. *et al*. GENCODE reference annotation for the human and mouse genomes. **Nucleic Acids Research**, v. 47, n. 1, p. 766-773, 24 Oct. 2018. DOI: 10.1093/nar/gky955.

84 MÉSZÁROS, B.; ERDOS, G.; DOSZTÁNYI, Z. IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. **Nucleic Acids Research**, v. 46, n. 1, p.329-337, 1 June 2018. DOI: 10.1093/nar/gky384.

85 ERDOS, G.; DOSZTÁNYI, Z. Analyzing protein disorder with IUPred2A. **Current Protocols in Bioinformatics**, v. 70, n. 1, p. 1-15, Apr. 2020. DOI: 10.1002/cpbi.99.

86 ERDOS, G.; PAJKOS, M.; DOSZTÁNYI, Z.. IUPred3: prediction of protein disorder enhanced with unambiguous experimental annotation and visualization of evolutionary conservation. **Nucleic Acids Research**, v. 49, n. 1, p. 297-303, 28 May 2021. DOI: 10.1093/nar/gkab408.

87 MÉSZÁROS, B.; SIMON, I.; DOSZTÁNYI, Z. Prediction of protein binding regions in disordered proteins. **PLoS Computational Biology**, v. 5, n. 5, p. 1000376, 1 May 2009. DOI: 10.1371/journal.pcbi.1000376.

88 DOSZTÁNYI, Z. *et al*. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. **Journal of Molecular Biology**, v. 347, n. 4, p. 827-839, abr. 2005. DOI: 10.1016/j.jmb.2005.01.071.

89 AGUET, F. *et al*. The GTEx consortium atlas of genetic regulatory effects across human tissues. **Science**, v. 369, n. 6509, p. 1318-1330, 11 Sept. 2020. DOI: 10.1126/science.aaz1776.

90 YANAI, I. et al. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. **Bioinformatics**, v. 21, n. 5, p. 650-659, 2005. DOI: 10.1093/bioinformatics/bti042.

91 KRYUCHKOVA-MOSTACCI, N.; ROBINSON-RECHAVI, M.. A benchmark of gene expression tissue-specificity metrics. **Briefings In Bioinformatics**, p. 008, 18 Feb. 2016. DOI: 10.1093/bib/bbw008.

92 JOSHI, Chintan J. *et al*. What are housekeeping genes? **Plos Computational Biology**, v. 18, n. 7, p. 1010295, 13 July 2022. DOI: 10.1371/journal.pcbi.1010295.

93 SCHOBER, P; BOER, C.; SCHWARTE, L. A. Correlation coefficients: appropriate use and interpretation. **Anesthesia & Analgesia**, v. 126, n. 5, p. 1763-1768, May 2018. DOI: 10.1213/ane.0000000000002864.

94 SHEATHER, S. J. Density estimation. **Statistical Science**, v. 19, n. 4, p. 588-597, 1 Nov. 2004. DOI: 10.1214/088342304000000297.

95 BARTON, H. J; ZENG, K. New methods for inferring the distribution of fitness effects for INDELs and SNPs. **Molecular Biology and Evolution**, v. 35, n. 6, p. 1536-1546, 4 Apr. 2018. DOI: 10.1093/molbev/msy054.

96 SCOTTI, M. M.; SWANSON, M. S..RNA mis-splicing in disease. **Nature Reviews Genetics**, v. 17, n. 1, p. 19-32, 23 Nov. 2015. DOI: 10.1038/nrg.2015.3.

97 SALIH, M. H.; AL-AZZAWIE, A. F; AL-ASSIE, A. H. A. Intronic SNPs and genetic diseases: a review. **International Journal for Research in Applied Sciences and Biotechnology**, v. 8, n. 2, p. 267-274, 20 Apr. 2021. DOI: 10.31033/ijrasb.8.2.36.

98 LIGHT, S.; ELOFSSON, A. The impact of splicing on protein domain architecture. **Current Opinion in Structural Biology**, v. 23, n. 3, p. 451-458, June 2013. DOI: 10.1016/j.sbi.2013.02.013.

99 WILHELMI, I. *et al*. Enriched alternative splicing in islets of diabetes-susceptible mice. **International Journal of Molecular Sciences**, v. 22, n. 16, p. 8597, 10 Aug. 2021. DOI: 10.3390/ijms22168597.

100 ANNA, A.; MONIKA, G. Splicing mutations in human genetic disorders: examples, detection, and confirmation. **Journal of Applied Genetics**, v. 59, n. 3, p. 253-268, 21 Apr. 2018. DOI: 10.1007/s13353-018-0444-7.

101 TU, Z. *et al*. Further understanding human disease genes by comparing with housekeeping genes and other genes. **BMC Genomics**, v. 7, n. 1, p. 1-16, 21 Feb. 2006. DOI: 10.1186/1471-2164-7-31.

102 EISENBERG, E.; LEVANON, E. Y. Human housekeeping genes, revisited. **Trends in Genetics**, v. 29, n. 10, p. 569-574, Oct. 2013. DOI: 10.1016/j.tig.2013.05.010.

103 HANSON, R. L. *et al*. Assessment of the potential role of natural selection in type 2 diabetes and related traits across human continental ancestry groups: comparison of phenotypic with genotypic divergence. **Diabetologia**, v. 63, n. 12, p. 2616-2627, 4 Sept. 2020. DOI: 10.1007/s00125-020-05272-8.

104 WELLS, J. C K. Ethnic variability in adiposity and cardiovascular risk: the variable disease selection hypothesis. **International Journal of Epidemiology**, v. 38, n. 1, p. 63-71, 27 Sept. 2008. DOI: 10.1093/ije/dyn183.

105 WANG, Y. *et al*. Theoretical and empirical quantification of the accuracy of polygenic scores in ancestry divergent populations. **Nature Communications**, v. 11, n. 1, p. 1-9, 31 July 2020. DOI: 10.1038/s41467-020-17719-y.

106 SOUTHAM, L. *et al*. Is the thrifty genotype hypothesis supported by evidence based on confirmed type 2 diabetes- and obesity-susceptibility variants? **Diabetologia**, v. 52, n. 9, p. 1846-1851, 13 June 2009. DOI: 10.1007/s00125-009-1419-3.

107 LEE, S. *et al*. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. **American Journal of Human Genetics**, v. 91, n. 2, p. 224-237, Aug. 2012. DOI: 10.1016/j.ajhg.2012.06.007.

108 MOUTSIANAS, L. *et al*. The power of gene-based rare variant methods to detect disease-associated variation and test hypotheses about complex disease. **PLoS Genetics**, v. 11, n. 4, p. 1005165, 23 Apr. 2015. DOI: 10.1371/journal.pgen.1005165.

109 CIRULLI, E. T. *et al*. Exome sequencing in amyotrophic lateral sclerosis identifies risk genes and pathways. **Science**, v. 347, n. 6229, p. 1436-1441, 27 Mar. 2015. DOI: 10.1126/science.aaa3650.

110 BOCHER, O. *et al*. Rare variant association testing for multicategory phenotype. **Genetic Epidemiology**, v. 43, n. 6, p. 646-656, 13 May 2019. DOI: 10.1002/gepi.22210.

111 FLANNICK, J. The contribution of low-frequency and rare coding variation to susceptibility to type 2 diabetes. **Current Diabetes Reports**, v. 19, n. 5, p. 1-10, 8 abr. 2019. DOI:10.1007/s11892-019-1142-5.

112 HUERTA-CHAGOYA, A. *et al*. The power of TOPMed imputation for the discovery of Latino-enriched rare variants associated with type 2 diabetes. **Diabetologia**, v. 66, n. 7, p. 1273-1288, 6 May 2023. DOI: 10.1007/s00125-023-05912-9.

113 MORGENTHALER, S.; THILLY, W. G. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (cast). **Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis**, v. 615, n. 1-2, p. 28-56, Feb. 2007. DOI: 10.1016/j.mrfmmm.2006.09.003.

114 BARTON, A. R. *et al*. Whole-exome imputation within UK Biobank powers rare coding variant association and fine-mapping analyses. **Nature Genetics**, v. 53, n. 8, p. 1260-1269, 5 July 2021. DOI: 10.1038/s41588-021-00892-1.

115 NEWTON-CHEH, C. What can we learn from common genetic variants with weak effects on cardiovascular disease risk? **Journal of the American College of Cardiology**, v. 73, n. 23, p. 2943-2945, 2019. DOI: 10.1016/j.jacc.2019.05.002.

116 PALLA, L.; DUDBRIDGE, F. A fast method that uses polygenic scores to estimate the variance explained by genome-wide marker panels and the proportion of variants affecting a trait. **American Journal of Human Genetics**, v. 97, n. 2, p. 250-259, Aug. 2015. DOI: 10.1016/j.ajhg.2015.06.005.

117 LOCKE, A. E. *et al*. Genetic studies of body mass index yield new insights for obesity biology. **Nature**, v. 518, n. 7538, p. 197-206, 11 Feb. 2015. DOI: 10.1038/nature14177.

118 KUNKLE, B. W. *et al*. Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing. **Nature Genetics**, v. 51, n. 3, p. 414-430, 28 Feb. 2019. DOI: 10.1038/s41588-019-0358-2.

119 MAHAJAN, A. *et al*. Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. **Nature Genetics**, v. 50, n. 11, p. 1505-1513, 8 Oct. 2018. DOI: 10.1038/s41588-018-0241-6.

120 SUZUKI, K. *et al*. Identification of 28 new susceptibility loci for type 2 diabetes in the Japanese population. **Nature Genetics**, v. 51, n. 3, p. 379-386, 4 Feb. 2019. DOI: 10.1038/s41588-018-0332-4.

121 VUJKOVIC, M. *et al*. Discovery of 318 new risk loci for type 2 diabetes and related vascular outcomes among 1.4 million participants in a multi-ancestry meta-analysis. **Nature Genetics**, v. 52, n. 7, p. 680-691, 15 June 2020. DOI: 10.1038/s41588-020-0637-y.

122 MAHAJAN, A. *et al*. Multi-ancestry genetic study of type 2 diabetes highlights the power of diverse populations for discovery and translation. **Nature Genetics**, v. 54, n. 5, p. 560-572, maio 2022. DOI: 10.1038/s41588-022-01058-3.

123 YANG, J. *et al*. Common SNPs explain a large proportion of the heritability for human height. **Nature Genetics**, v. 42, n. 7, p. 565-569, 20 June 2010. DOI: 10.1038/ng.608.

124 DUDBRIDGE, F. Power and predictive accuracy of polygenic risk scores. **PLoS Genetics**, v. 9, n. 3, p. 1003348, 21 Mar. 2013. DOI: 10.1371/journal.pgen.1003348.

125 DUDBRIDGE, F. Polygenic epidemiology. **Genetic Epidemiology**, v. 40, n. 4, p. 268-272, 7 Apr. 2016. DOI: 10.1002/gepi.21966.

126 MEFFORD, J. *et al*. Efficient estimation and applications of cross-validated genetic predictions to polygenic risk scores and linear mixed models. **Journal of Computational Biology**, v. 27, n. 4, p. 599-612, 1 Apr. 2020. DOI: 10.1089/cmb.2019.0325.

127 CHOI, S. W.; MAK, T. S.; O'REILLY, P. F. Tutorial: a guide to performing polygenic risk score analyses. **Nature Protocols**, v. 15, n. 9, p. 2759-2772, 24 July 2020. DOI: 10.1038/s41596-020-0353-1.

128 BITARELLO, B. D; MATHIESON, I. Polygenic scores for height in admixed populations. **Genomes Genetics**, v. 10, n. 11, p. 4027-4036, 1 Nov. 2020. DOI: 10.1534/g3.120.401658.

129 GE, T. *et al*. Development and validation of a trans-ancestry polygenic risk score for type 2 diabetes in diverse populations. **Genome Medicine**, v. 14, n. 1, p. 1-16, 29 June 2022. DOI: 10.1186/s13073-022-01074-2.

130 MARS, N. *et al*. Genome-wide risk prediction of common diseases across ancestries in one million people. **Cell Genomics**, v. 2, n. 4, p. 100118, Apr. 2022. DOI: 10.1016/j.xgen.2022.100118.

131 TORKAMANI, A.; WINEINGER, N. E.; TOPOL, E. J. The personal and clinical utility of polygenic risk scores. **Nature Reviews Genetics**, v. 19, n. 9, p. 581-590, 22 May 2018.. DOI: 10.1038/s41576-018-0018-x.

132 MAYER-DAVIS, E. J. *et al*. Incidence Trends of type 1 and Type 2 diabetes among youths, 2002–2012. **New England Journal of Medicine**, v. 376, n. 15, p. 1419-1429, 13 Apr. 2017. DOI: 10.1056/nejmoa1610187.

133 BELLOU, V. *et al*. Risk factors for type 2 diabetes mellitus: an exposure-wide umbrella review of meta-analyses. **PLoS One**, v. 13, n. 3, p. 0194127, 20 Mar. 2018. DOI: 10.1371/journal.pone.0194127.

134 UDLER, M. S *et al*. Genetic Risk Scores for Diabetes Diagnosis and Precision Medicine. **Endocrine Reviews**, v. 40, n. 6, p. 1500-1520, 19 jul. 2019. DOI: 10.1210/er.2019-00088.

135 NASLAVSKY, M. S.; SCLIAR, M. O.; YAMAMOTO, G. L. *et al*. Whole-genome sequencing of 1,171 elderly admixed individuals from Brazil. **Nature Communications**, v. 13, n. 1, p. 1-11, 4 Mar. 2022. DOI: 10.1038/s41467-022-28648-3.

136 RENTZSCH, P. *et al*. CADD-Splice—improving genome-wide variant effect prediction using deep learning-derived splice scores. **Genome Medicine**, v. 13, n. 1, p. 1-12, 22 Feb. 2021. DOI: 10.1186/s13073-021-00835-9.

137 FLÓRIO, F. M. *et al*. Size effect in observational studies in public oral health: importance, calculation and interpretation. **Ciência & Saúde Coletiva**, v. 28, n. 2, p. 599-608, Feb. 2023. DOI: 10.1590/1413-81232023282.09822022en.

138 CHEN, H.; COHEN, P.; CHEN, S. How big is a big odds ratio? interpreting the magnitudes of odds ratios in epidemiological studies. **Communications in Statistics - simulation and computation**, v. 39, n. 4, p. 860-864, 31 Mar. 2010. DOI: 10.1080/03610911003650383.

139 KARCZEWSKI, K. J. *et al*. The mutational constraint spectrum quantified from variation in 141,456 humans. **Nature**, v. 581, n. 7809, p. 434-443, 27 May 2020. DOI: 10.1038/s41586-020-2308-7.

140 LAMBERT, S. A. *et al*. The polygenic score catalog as an open database for reproducibility and systematic evaluation. **Nature Genetics**, v. 53, n. 4, p. 420-425, 10 Mar. 2021. DOI: 10.1038/s41588-021-00783-5.

141 MARTIN, A. R. *et al*. Clinical use of current polygenic risk scores may exacerbate health disparities. **Nature Genetics**, v. 51, n. 4, p. 584-591, 29 Mar. 2019. DOI: 10.1038/s41588-019-0379-x.

142 POLFUS, L. M. *et al*. Genetic discovery and risk characterization in type 2 diabetes across diverse populations. **Human Genetics and Genomics Advances**, v. 2, n. 2, p. 100029, Apr. 2021. DOI: 10.1016/j.xhgg.2021.100029.

143 PRIVÉ, F. *et al*. Portability of 245 polygenic scores when derived from the UK Biobank and applied to 9 ancestry groups from the same cohort. **American Journal of Human Genetics**, v. 109, n. 1, p. 12-23, Jan. 2022. DOI: 10.1016/j.ajhg.2021.11.008.

144 SERDAR, C. C. *et al*. Sample size, power and effect size revisited: simplified and practical approaches in pre-clinical, clinical and laboratory studies. **Biochemia Medica**, v. 31, n. 1, p. 27-53, 15 Feb. 2021. DOI:10.11613/bm.2021.010502.

145 MAROULI, E. *et al*. Rare and low-frequency coding variants alter human adult height. **Nature**, v. 542, n. 7640, p. 186-190, 1 Feb. 2017. DOI:10.1038/nature21039.

146 BONNEFOND, A. *et al*. Pathogenic variants in actionable MODY genes are associated with type 2 diabetes. **Nature Metabolism**, v. 2, n. 10, p. 1126-1134, 12 Oct. 2020. DOI:10.1038/s42255-020-00294-3.

147 SPRACKLEN, C. N. *et al*. Identification of type 2 diabetes loci in 433,540 East Asian individuals. **Nature**, v. 582, n. 7811, p. 240-245, 6 May 2020. DOI: 10.1038/s41586-020-2263-3.

148 WANG, L. *et al*. Trends in Prevalence of Diabetes and Control of Risk Factors in Diabetes Among US Adults, 1999-2018. **Jama**, v. 326, n. 8, p. 704, 24 Aug. 2021. DOI: 10.1001/jama.2021.9883.