UNIVERSIDADE DO VALE DO PARAÍBA – UNIVAP
INSTITUTO DE PESQUISA E DESENVOLVIMENTO – IP&D
PROGRAMA DE MESTRADO EM ENGENHARIA BIOMÉDICA – PPGEB

Ramon Varella Diniz

VALIDAÇÃO DA ESPECTROSCOPIA ATR-FTIR ASSOCIADA AO *MACHINE LEARNING* PARA IDENTIFICAÇÃO DE UM POLIMORFISMO GENÉTICO

VALIDATION OF ATR-FTIR SPECTROSCOPY ASSOCIATED WITH MACHINE LEARNING FOR THE IDENTIFICATION OF A GENETIC POLYMORPHISM

São José dos Campos – SP
2026

Ramon Varella Diniz

VALIDAÇÃO DA ESPECTROSCOPIA ATR-FTIR ASSOCIADA AO *MACHINE LEARNING* PARA IDENTIFICAÇÃO DE UM POLIMORFISMO GENÉTICO

VALIDATION OF ATR-FTIR SPECTROSCOPY ASSOCIATED WITH MACHINE LEARNING FOR THE IDENTIFICATION OF A GENETIC POLYMORPHISM

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia Biomédica, como complementação dos créditos necessários para obtenção do grau de Mestre em Engenharia Biomédica.

Orientadora: Prof. Dra. Renata de Azevedo Canevari

São José dos Campos – SP
2026

Fundação Valeparaibana de Ensino

BIBLIOTECA

TERMO DE AUTORIZAÇÃO DE DIVULGAÇÃO DA OBRA

Ficha catalográfica

Diniz, Ramon Varella
Validação da espectroscopia ATR-FTIR associada ao machine learning para identificação de um polimorfismo genético / Ramon Varella Diniz; orientadora, Renata de Azevedo Canevari. - São José dos Campos, SP, 2026.
1 CD-ROM, 118 p.

Dissertação (Mestrado Acadêmico) - Universidade do Vale do Paraíba, São José dos Campos. Programa de Pós-Graduação em Engenharia Biomédica.

Inclui referências

1. Engenharia Biomédica. 2. ATR-FTIR. 3. Machine learning. 4. Polimorfismo genético. I. Canevari, Renata de Azevedo, orient. II. Universidade do Vale do Paraíba. Programa de Pós-Graduação em Engenharia Biomédica. III. Título.

Eu, Ramon Varella Diniz, autor(a) da obra acima referenciada:

Autorizo a divulgação total ou parcial da obra impressa, digital ou fixada em outro tipo de mídia, bem como, a sua reprodução total ou parcial, devendo o usuário da reprodução atribuir os créditos ao autor da obra, citando a fonte.

Declaro, para todos os fins e efeitos de direito, que o Trabalho foi elaborado respeitando os princípios da moral e da ética e não violou qualquer direito de propriedade intelectual sob pena de responder civil, criminal, ética e profissionalmente por meus atos.

São José dos Campos, 30 de Abril de 2026.

Autor(a) da Obra

Data da defesa: 02 / 03 / 2026

UNIVERSIDADE DO VALE DO PARAÍBA

IP&D UNIVAP

RAMON VARELLA DINIZ

"VALIDAÇÃO DA ESPECTROSCOPIA ATR-FTIR ASSOCIADA AO MACHINE LEARNING PARA IDENTIFICAÇÃO DE UM POLIMORFISMO GENÉTICO."

Dissertação aprovada como requisito parcial à obtenção do grau de Mestre, do Programa de Pós-Graduação em Engenharia Biomédica, do Instituto de Pesquisa e Desenvolvimento da Universidade do Vale do Paraíba - Univap, pela seguinte banca examinadora:

Prof.ª Dr.ª Laurita do Santos – Universidade Brasil	
Prof.ª Dr.ª Maiara Lima Castilho	
Prof.ª Dr.ª Renata de Azevedo Canevari	

Prof.ª Dr.ª Juliana Ferreira Strixino

Diretora do IP&D – Univap

São José dos Campos, 02 de março de 2026.

“ O mais importante é não parar de questionar.
A curiosidade tem sua própria razão de existir. ”

- Albert Einstein

Dedico este trabalho aos meus pais, Adelval e Andreia,
por todo o apoio incondicional, incentivo e confiança
em minhas escolhas.

AGRADECIMENTOS

Agradeço primeiramente a Deus, pela força, pelo amparo nos momentos difíceis e por ter guiado cada etapa desta caminhada.

À minha família, base de tudo o que sou, pelo apoio constante, carinho e incentivo ao longo de toda essa trajetória. Em especial, aos meus pais, Adelval e Andreia, por sempre acreditarem em mim e por tornarem possíveis todas as minhas escolhas. Ao meu irmão, à minha irmã, à minha avó e a todos os meus tios e tias, que de diferentes formas estiveram presentes e contribuíram para que este momento se concretizasse. Ao meu namorado, pelo companheirismo, paciência e apoio nos dias mais desafiadores.

Aos meus amigos de longa data, especialmente à galera do Clubinho, por permanecerem ao meu lado ao longo dos anos, tornando o caminho mais leve e lembrando-me constantemente da importância das pausas, das risadas e da amizade verdadeira.

À minha orientadora, pela confiança, orientação e dedicação ao longo de toda minha jornada acadêmica, bem como a toda a equipe do Laboratório de Genética Molecular (GeneLab), pelo ambiente colaborativo, pelas discussões científicas e pelo aprendizado compartilhado. A todos os participantes que contribuíram para esta pesquisa.

Às professoras da banca examinadora, pela disponibilidade, atenção e valiosas contribuições para o aprimoramento deste trabalho.

À Universidade do Vale do Paraíba (UNIVAP), ao Instituto de Pesquisa e Desenvolvimento (IP&D) e ao Programa de Pós-Graduação em Engenharia Biomédica (PPGEB), pela estrutura e suporte acadêmico oferecidos durante o desenvolvimento desta pesquisa.

À Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), pela concessão da bolsa de mestrado (Processo nº 2024/19969-6), que fomentou a realização desta pesquisa.

IMPACTO POTENCIAL DESTA PESQUISA

1) Científica e Técnica

A pesquisa apresenta impacto científico e técnico ao propor a validação de uma abordagem complementar para triagem genotípica de polimorfismos de nucleotídeo único (SNPs) baseada na integração entre espectroscopia ATR-FTIR e *machine learinig*. O estudo contribui para o avanço de metodologias alternativas às técnicas convencionais de genotipagem e consolida um fluxo de trabalho reprodutível envolvendo PCR, aquisição espectral e modelagem computacional, com potencial para futuras aplicações.

2) Social

O impacto social da pesquisa está relacionado ao potencial de aplicação da metodologia como ferramenta de triagem para identificação de genótipos associados à predisposição genética a doenças multifatoriais. Ao possibilitar a identificação de variantes genéticas antes do estabelecimento clínico da doença, a estratégia pode contribuir para ações de atenção primária à saúde, com foco em prevenção, monitoramento e orientação individualizada, reduzindo potenciais impactos clínicos e sociais associados a diagnósticos tardios. Assim, a pesquisa se alinha a uma perspectiva de medicina preventiva, com potencial benefício coletivo.

3) Inovadora e Econômica

A proposta apresenta caráter inovador ao explorar a análise espectral de produtos amplificados de DNA associada ao *machine learning* como alternativa às técnicas tradicionais de genotipagem, especialmente para aplicações em triagem em larga escala. O método demonstra potencial para redução de custos, tempo de processamento e complexidade operacional, contribuindo para ampliar a viabilidade de estudos populacionais, triagens preliminares e aplicações em contextos com recursos limitados.

4) Educacional e Cultural

A pesquisa contribui para a formação de recursos humanos qualificados ao envolver atividades experimentais, análise espectral e modelagem computacional em um contexto interdisciplinar. O estudo favorece o desenvolvimento de competências em biologia molecular, engenharia biomédica e análise de dados, além de estimular o engajamento da comunidade e a conscientização sobre genética, prevenção de doenças e pesquisa biomédica.

5) Internacionalização e Inserção Local

A pesquisa apresenta impacto no processo de internacionalização ao envolver colaboração com pesquisador internacional na área de análise de dados e *machine learning*, fortalecendo o intercâmbio científico e ampliando a visibilidade da produção acadêmica em contexto global. Paralelamente, destaca-se a inserção local por meio da participação de voluntários da comunidade e do envolvimento institucional no âmbito da Universidade do Vale do Paraíba (Univap). Essa integração entre colaboração internacional e participação local promove desenvolvimento científico com impacto territorial.

6) Alinhamento aos Objetivos de Desenvolvimento Sustentável (ODS)

A presente pesquisa apresenta alinhamento com os seguintes Objetivos de Desenvolvimento Sustentável (ODS) estabelecidos pela Organização das Nações Unidas (ONU):

• **ODS 3 (Saúde e Bem-Estar):** A pesquisa propõe uma abordagem para identificação de genótipos associados à predisposição a doenças multifatoriais, com potencial aplicação em triagem genética, prevenção e monitoramento individualizado em saúde.

• **ODS 4 (Educação de Qualidade):** A pesquisa contribui para a formação acadêmica e capacitação técnica em contexto interdisciplinar, além de favorecer a divulgação científica e a aproximação entre universidade e sociedade.

• **ODS 9 (Indústria, Inovação e Infraestrutura):** A pesquisa promove o desenvolvimento de uma abordagem inovadora que integra espectroscopia vibracional, biologia molecular e *machine learning*. A proposta fortalece a integração multidisciplinar e apresenta potencial para futuras aplicações laboratoriais.

• **ODS 12 (Consumo e Produção Responsáveis):** A pesquisa propõe uma metodologia com potencial redução no consumo de reagentes, tempo experimental e recursos laboratoriais quando comparada a técnicas tradicionais de genotipagem. Essa otimização contribui para práticas científicas mais sustentáveis, reduzindo desperdícios e incentivando abordagens experimentais com menor impacto ambiental e maior eficiência operacional.

• **ODS 17 (Parcerias e Meios de Implementação):** A pesquisa fortalece a colaboração científica internacional e a integração entre diferentes instituições, com participação da comunidade local. Essa cooperação amplia o intercâmbio de conhecimento e o alcance científico e tecnológico do estudo.

POTENTIAL IMPACT OF THIS RESEARCH

1) Scientific and Technical

This research presents scientific and technical impact by proposing the validation of a complementary approach for the genotypic screening of single nucleotide polymorphisms (SNPs) based on the integration of ATR-FTIR spectroscopy and machine learning. The study contributes to the advancement of alternative methodologies to conventional genotyping techniques and consolidates a reproducible workflow involving PCR, spectral acquisition, and computational modeling, with potential for future applications.

2) Social

The social impact of this research is related to the potential application of the methodology as a screening tool for the identification of genotypes associated with genetic predisposition to multifactorial diseases. By enabling the identification of genetic variants before the clinical onset of disease, this strategy may contribute to primary healthcare actions focused on prevention, monitoring, and individualized guidance, reducing potential clinical and social impacts associated with late diagnoses. Therefore, the research is aligned with a preventive medicine perspective, with potential collective benefits.

3) Innovative and Economic

The proposal is innovative in character by exploring the spectral analysis of amplified DNA products associated with machine learning as an alternative to traditional genotyping techniques, especially for large-scale screening applications. The method demonstrates potential for reducing costs, processing time, and operational complexity, contributing to the expansion of feasibility for population studies, preliminary screenings, and applications in resource-limited settings.

4) Educational and Cultural

The research contributes to the training of qualified human resources by involving experimental activities, spectral analysis, and computational modeling within an interdisciplinary context. The study promotes the development of competencies in molecular biology, biomedical engineering, and data analysis, in addition to encouraging community engagement and raising awareness about genetics, disease prevention, and biomedical research.

5) Internationalization and Local Insertion

The research has an impact on the internationalization process by involving collaboration with an international researcher in the field of data analysis and machine learning, strengthening scientific exchange and increasing the visibility of academic production in a global context. At the same time, local insertion is highlighted through the participation of community volunteers and institutional involvement within the University of Vale do Paraíba (Univap). This integration between international collaboration and local participation promotes scientific development with territorial impact.

6) Alignment with the Sustainable Development Goals (SDGs)

This research is aligned with the following Sustainable Development Goals (SDGs) established by the United Nations (UN):

• **SDG 3 (Good Health and Well-Being):** The research proposes an approach for identifying genotypes associated with predisposition to multifactorial diseases, with potential application in genetic screening, prevention, and individualized health monitoring.

• **SDG 4 (Quality Education):** The research contributes to academic training and technical qualification in an interdisciplinary context, in addition to promoting scientific dissemination and strengthening the relationship between the university and society.

• **SDG 9 (Industry, Innovation and Infrastructure):** The research promotes the development of an innovative approach that integrates vibrational spectroscopy, molecular biology, and machine learning. The proposal strengthens multidisciplinary integration and presents potential for future laboratory applications.

• **SDG 12 (Responsible Consumption and Production):** The research proposes a methodology with potential reduction in reagent consumption, experimental time, and laboratory resources when compared to traditional genotyping techniques. This optimization contributes to more sustainable scientific practices by reducing waste and encouraging experimental approaches with lower environmental impact and greater operational efficiency.

• **SDG 17 (Partnerships for the Goals):** The research strengthens international scientific collaboration and integration among different institutions, with participation from the local community. This cooperation expands knowledge exchange and the scientific and technological reach of the study.

RESUMO

Polimorfismos de nucleotídeo único (SNPs) desempenham um papel central na suscetibilidade genética a distúrbios multifatoriais, como obesidade e diabetes mellitus tipo 2 (DM2), o que evidencia a necessidade de estratégias de genotipagem escaláveis e de baixo custo. Este estudo avaliou a viabilidade da espectroscopia no infravermelho com transformada de Fourier por reflexão total atenuada (ATR-FTIR) combinada a *machine learning* (ML) para discriminar os diferentes tipos de genótipos do SNP −3826A/G localizado no gene *UCP1*. A genotipagem do SNP foi realizada pela PCR quantitativa em tempo real (qPCR) utilizando ensaios TaqMan em amostras de DNA extraído do sangue de 190 participantes para a definição dos grupos genotípicos (AA, AG e GG). A PCR qualitativa foi realizada em todas as amostras e nos controles negativos de reação (NTCs). Os *amplicons* da PCR e os NTCs foram utilizados na análise espectral. As análises de componentes principais (PCA), modelos supervisionados de ML e deep learning (DL) foram aplicadas diretamente aos espectros normalizados por Variância Normal Padrão (SNV) nos intervalos espectrais de 2800–3800 cm^{-1}, 950–1200 cm^{-1} e 900–1100 cm^{-1}. A viabilidade da técnica de ATR-FTIR associada ao ML foi avaliada por meio da comparação com a qPCR e o sequenciamento de nova geração (NGS), ambas técnicas de genotipagem consideradas atualmente padrão ouro. O melhor desempenho observado na discriminação entre os genótipos AA e GG foi obtido com o modelo de DL de perceptron multicamdas com arquitetura residual simulada na região de 2800–3800 cm^{-1}, área sob a curva (AUC) de 0,654 e acurácia de 0,716. Nas regiões de *fingerprint* do DNA de 900–1100 cm^{-1} e 950–1200 cm^{-1}, o melhor desempenho foi observado com o modelo de regressão logística, com AUC de 0,635 e 0,644 e acurácia de 0,696 e 0,720, respectivamente. A técnica de ATR-FTIR associada ao ML apresentou melhor viabilidade de execução em relação as técnicas de sequenciamento e qPCR, com um tempo de processamento por amostra similar a qPCR e menor que o NGS e um custo inferior a ambas as técnicas. Este estudo é pioneiro na aplicação da espectroscopia ATR-FTIR associada ao ML na discriminação de SNPs do genoma humano, mostrando ser uma abordagem com um alto potencial de rastreamento, relativamente rápida e de baixo custo. O estabelecimento de novos critérios e modelos de ML poderão aumentar significativamente o desempenho da técnica na identificação desses polimorfismos genéticos.

Palavras-chave: ATR-FTIR; *machine learning*; polimorfismo genético.

ABSTRACT

Single nucleotide polymorphisms (SNPs) play a central role in genetic susceptibility to multifactorial disorders such as obesity and type 2 diabetes mellitus (T2DM), highlighting the need for scalable and cost-effective genotyping strategies. This study evaluated the feasibility of attenuated total reflection Fourier transform infrared (ATR-FTIR) spectroscopy combined with machine learning (ML) to discriminate different genotypic variants of the −3826A/G SNP located in the UCP1 gene. SNP genotyping was performed by quantitative real-time PCR (qPCR) using TaqMan assays in DNA samples extracted from the blood of 190 participants to define the genotypic groups (AA, AG, and GG). Qualitative PCR was performed for all samples and for the negative template controls (NTCs). PCR *amplicons* and NTCs were used for spectral analysis. Principal component analysis (PCA), supervised ML models, and deep learning (DL) models were applied directly to spectra normalized by Standard Normal Variate (SNV) within the spectral ranges of 2800–3800 cm^{-1}, 950–1200 cm^{-1}, and 900–1100 cm^{-1}. The feasibility of ATR-FTIR combined with ML was evaluated through comparison with qPCR and next-generation sequencing (NGS), both currently considered gold-standard genotyping techniques. The best performance in discriminating AA and GG genotypes was achieved using a DL multilayer perceptron model with a simulated residual architecture in the 2800–3800 cm^{-1} region, yielding an area under the curve (AUC) of 0.654 and an accuracy of 0.716. In the DNA fingerprint regions of 900–1100 cm^{-1} and 950–1200 cm^{-1}, the best performance was obtained with a logistic regression model, with AUC values of 0.635 and 0.644 and accuracy values of 0.696 and 0.720, respectively. ATR-FTIR combined with ML demonstrated greater feasibility of implementation compared with sequencing and qPCR techniques, presenting a processing time per sample similar to qPCR and shorter than NGS, as well as lower cost than both techniques. This study is pioneering in applying ATR-FTIR spectroscopy combined with ML for discrimination of SNPs in the human genome, demonstrating high screening potential as a relatively rapid and cost-effective approach. The establishment of new criteria and ML models may significantly improve the performance of this technique in identifying these genetic polymorphisms.

Keywords: ATR-FTIR; machine learning; genetic polymorphism.

LISTA DE ABREVIATURAS E SIGLAS

ATR-FTIR · Espectroscopia no Infravermelho por Transformada de Fourier com Reflexão Total Atenuada (Attenuated Total Reflectance Fourier Transform Infrared Spectroscopy)

AUC · Área Sob a Curva (Area Under the Curve)

DL · Aprendizado Profundo (Deep Learning)

DL-MLP · Aprendizado Profundo com Perceptron Multicamadas (Deep Learning with Multilayer Perceptron)

DM2 · Diabetes Mellitus tipo 2

HWE · Equilíbrio de Hardy–Weinberg (Hardy–Weinberg Equilibrium)

LDA · Análise Discriminante Linear (Linear Discriminant Analysis)

ML · Aprendizado de Máquina (Machine Learning)

NGS · Sequenciamento de Nova Geração (Next-Generation Sequencing)

NTC · Controle Negativo Sem Molde (Negative Template Control)

PCR · Reação em Cadeia da Polimerase (Polymerase Chain Reaction)

PCA · Análise de Componentes Principais (Principal Component Analysis)

PC · Componente Principal (Principal Component)

ROC · Curva Característica de Operação do Receptor (Receiver Operating Characteristic Curve)

SNP · Polimorfismo de Nucleotídeo Único (Single Nucleotide Polymorphism)

SNV · Variação Normal Padrão (Standard Normal Variate)

SVM · Máquina de Vetores de Suporte (Support Vector Machine)

TEM · Tempo de Execução Manual

TOI · Tempo de Operação Instrumental

TTP · Tempo Total de Processamento

UCP1 · Proteína Desacopladora 1 (*Uncoupling Protein 1*)

LISTA DE FIGURAS

Figura 1. Esquema experimental dos passos realizados no estudo para a discriminação dos genótipos do SNP -3826A/G do gene *UCP1* pela análise de espectroscopia ATR-FTIR associada a *machine learning.* 36

Figura 2. Representação da localização genômica do SNP −3826A/G (rs1800592) no gene *UCP1* (em azul) de acordo com o NCBI (NC_000004.12, dbSNP, NCBI). Nota-se que outros polimorfismos descritos também podem ser observados na mesma região. 40

Figura 3. Espectro de absorbância no ultravioleta (UV) obtidos na quantificação de cada amostra de DNA, destacando as razões de contaminação com proteínas (260/280) e reagentes (260/230) e a concentração da amostra em ng/µL. 52

Figura 4. Perfil eletroforético de amostras de DNA extraídas de sangue em gel de agarose 1,0% (TBE 1X) corado com brometo de etídio. PM: padrão de peso molecular de 100 pb. 53

Figura 5. Curvas de amplificação obtidas por *qPCR SNP genotyping* com sondas *TaqMan* para o polimorfismo −3826A/G do gene *UCP1*. Os gráficos apresentam o número de ciclos em função da intensidade de fluorescência (ΔRn), correspondentes aos dois ensaios realizados. 54

Figura 6. Gráfico de dispersão representando a distribuição dos genótipos obtidos na reação de genotipagem por qPCR para o SNP -3826A/G (rs1800592) do gene *UCP1*. Cada ponto representando uma amostra individual. Os sinais de fluorescência relativos aos alelos foram detectados pelos fluoróforos VIC (alelo G) e FAM (alelo A). Grupo homozigoto para o alelo G (região inferior direita – em vermelho), heterozigoto (região central – em verde) e homozigoto para o alelo A (região superior esquerda – em azul. O controle negativo (NCT) é representado pela cor preta. 55

Figura 7. Eletroforese em gel de agarose a 1,6% dos produtos amplificados por PCR qualitativa da região contendo o SNP –3826A/G do gene *UCP1*. Visualizam-se bandas únicas, nítidas e bem definidas correspondentes aos *amplicons* de 153 pares de bases (pb), compatíveis com o fragmento esperado, indicando alta especificidade e eficiência da reação de amplificação após a etapa de otimização dos *primers*. O marcador de peso molecular (M) utilizado foi de 50 pb, permitindo confirmar o tamanho do produto amplificado. A ausência de bandas inespecíficas ou de amplificação no controle negativo (NTC) confirma a ausência de contaminações e a fidelidade da reação. .. 56

Figura 8. Espectro médio de absorbância no infravermelho na faixa espectral de 4000 a 550 cm^{-1} obtido por espectroscopia ATR-FTIR dos *amplicons* de PCR correspondentes à região do SNP −3826A/G do gene *UCP1.* 57

Figura 9. Espectros médios de ATR-FTIR das amostras de DNA amplificado por PCR, agrupadas de acordo com os genótipos AA (n = 88), AG (n = 80) e GG (n = 22), acompanhados do desvio padrão. O painel superior apresenta os espectros médios sem normalização, enquanto o painel inferior mostra os espectros após a aplicação da normalização por variância normal padrão (SNV). 59

Figura 10. Espectros médios normalizados de ATR-FTIR dos *amplicons* de PCR correspondentes aos diferentes grupos genotípicos AA, AG e GG, bem como aos controles negativos de reação (*negative template control*, NTC). 60

Figura 11. Espectros diferenciais obtidos pela subtração do espectro médio de ATR-FTIR dos controles negativos de reação (*negative template control*, NTC) em relação aos espectros médios dos *amplicons* de PCR para cada grupo genotípico (AA–Neg, AG–Neg e GG–Neg). A subtração evidencia regiões espectrais com menor contribuição de reagentes residuais da PCR, indicando que os intervalos de 900–1100 cm^{-1}, 950-

1200 cm⁻¹ e 2800–3800 cm⁻¹ são menos afetados por componentes não associados ao DNA. 61
Figura 12. Análise de componentes principais (*principal component analysis*, PCA) aplicada aos espectros de ATR-FTIR normalizados por SNV na região de 900–1100 cm⁻¹, selecionada com base na análise dos espectros diferenciais como um intervalo espectral com menor interferência de reagentes. (A) Gráficos de escores da PCA considerando conjuntamente todos os grupos genotípicos, com combinações pareadas entre os quatro primeiros componentes principais (PC1×PC2, PC1×PC3, PC1×PC4, PC2×PC3, PC2×PC4 e PC3×PC4), nos quais as amostras AA, AG e GG são representadas, respectivamente, por pontos pretos, vermelhos e azuis; à direita são apresentados os perfis de *loadings* correspondentes PC1–PC4 ao longo da faixa de números de onda analisada. (B–D) Gráficos de escores da PCA construídos utilizando as mesmas combinações de componentes, considerando apenas comparações pareadas entre genótipos: AA×GG (B), AA×AG (C) e AG×GG (D). 63
Figura 13. Análise de componentes principais (*principal component analysis*, PCA) aplicada aos espectros de ATR-FTIR normalizados por SNV na região de 950–1200 cm⁻¹, selecionada com base na análise dos espectros diferenciais como um intervalo espectral com menor interferência de reagentes. (A) Gráficos de escores da PCA considerando conjuntamente todos os grupos genotípicos, com combinações pareadas entre os quatro primeiros componentes principais (PC1×PC2, PC1×PC3, PC1×PC4, PC2×PC3, PC2×PC4 e PC3×PC4), nos quais as amostras AA, AG e GG são representadas, respectivamente, por pontos pretos, vermelhos e azuis; à direita são apresentados os perfis de *loadings* correspondentes PC1–PC4 ao longo da faixa de números de onda analisada. (B–D) Gráficos de escores da PCA construídos utilizando as mesmas combinações de componentes, considerando apenas comparações pareadas entre genótipos: AA×GG (B), AA×AG (C) e AG×GG (D). 64
Figura 14. Análise de componentes principais (*principal component analysis*, PCA) aplicada aos espectros de ATR-FTIR normalizados por SNV na região de 2800–3800 cm⁻¹, selecionada com base na análise dos espectros diferenciais como um intervalo espectral com menor interferência de reagentes. (A) Gráficos de escores da PCA considerando conjuntamente todos os grupos genotípicos, com combinações pareadas entre os quatro primeiros componentes principais (PC1×PC2, PC1×PC3, PC1×PC4, PC2×PC3, PC2×PC4 e PC3×PC4), nos quais as amostras AA, AG e GG são representadas, respectivamente, por pontos pretos, vermelhos e azuis; à direita são apresentados os perfis de *loadings* correspondentes PC1–PC4 ao longo da faixa de números de onda analisada. (B–D) Gráficos de escores da PCA construídos utilizando as mesmas combinações de componentes, considerando apenas comparações pareadas entre genótipos: AA×GG (B), AA×AG (C) e AG×GG (D). 65
Figura 15. Desempenho dos modelos de aprendizado de máquina na discriminação entre pares de genótipos com base nos espectros de ATR-FTIR dos *amplicons* de PCR, considerando as comparações AA×GG, AA×AG e AG×GG. Os valores de AUC são apresentados nos painéis à esquerda e os de acurácia nos painéis à direita, para todos os intervalos espectrais analisados. (A) Resultados obtidos para a região de 2800–3800 cm^{-1}. (B) Resultados obtidos para a região de 900–1100 cm⁻¹. (C) Resultados obtidos para a região de 950–1200 cm⁻¹ 67

LISTA DE TABELAS

Tabela 1. Dados técnicos sobre o SNP -3826A/G do gene *UCP1* analisado pela *qPCR SNP genotyping.* 39

Tabela 2. Condições de reação da PCR qualitativa para amplificação do SNP -3826A/G do gene *UCP1* 41

Tabela 3. Atribuição vibracional das bandas da região de *fingerprint* observadas nos espectros ATR-FTIR dos *amplicons* de PCR, com base em referências da literatura. 57

Tabela 4. Comparação do tempo de execução manual (TEM), tempo de operação instrumentação (TOI), tempo total de processamento (TTP) e custos de reagentes para as metodologias de genotipagem de SNPs analisadas 69

LISTA DE EQUAÇÕES

Equação 1. Tempo de Execução manual a nível de lote 48
Equação 2. Tempo de Operação Instrumental a nível de lote 48
Equação 3. Tempo Total de Processamento a nível de lote 48
Equação 4. Tempo de Execução manual a nível de amostra 49
Equação 5. Tempo de Operação Instrumental a nível de amostra 49
Equação 6. Tempo Total de Processamento a nível de amostra 50
Equação 7. Custo total da metodologia a nível de lote 50
Equação 8. Custo total da metodologia a nível de amostra 51

SUMÁRIO

1 INTRODUÇÃO 19
2 OBJETIVOS 22
2.1 Principal 22
2.2 Específicos 22
3 REVISÃO DE LITERATURA 23
3.1 Polimorfismos genéticos 23
3.2 Polimorfismo SNP -3826 A/G do gene *UCP1* 26
3.3 Detecção de SNPs por *qPCR SNP genotyping* 28
3.4 FTIR na análise de DNA 29
3.5 Identificação de SNPs por *machine learning* 31
4 MATERIAIS E MÉTODOS 35
4.1 Aspectos éticos, delineamento do estudo e coleta de amostras 35
4.2 Análise do SNP pela *qPCR SNP genotyping* 37
***4.2.1 Extração de DNA das amostras de sangue* 37**
***4.2.2 Quantificação e análises da integridade do DNA* 37**
4.2.3 Genotipagem pela *qPCR SNP genotyping* 38
4.3 Análise dos espectros de ATR-FTIR dos *amplicons* 40
***4.3.1 Obtenção dos amplicons por PCR qualitativa* 40**
***4.3.2 Obtenção dos espectros dos amplicons por ATR-FTIR* 41**
***4.3.3 Análise de machine learning não supervisionado* 42**
***4.3.4 Análise de machine learning supervisionado* 44**
4.4 Análise de viabilidade da espectroscopia ATR-FTIR associada ao ML 46
5 RESULTADOS 52
5.1 Qualidade das amostras de DNA 52
5.2 Genotipagem 54
5.3 Espectros obtidos dos *amplicons* 56
5.4 Espectros diferenciais em relação ao NTC 60
***5.5 Machine learning* 62**
5.6 Viabilidade da espectroscopia ATR-FTIR associado ao ML 69
6 DISCUSSÃO 71
7 CONCLUSÃO 78

REFERÊNCIAS **79**

GLOSSÁRIO **89**

APÊNDICE A: TERMO DE CONSENTIMENTO LIVRE E ESCLARECIDO – TCLE **90**

APÊNDICE B: CARACTERIZAÇÃO DEMOGRÁFICA E CLÍNICA DA POPULAÇÃO DO ESTUDO **93**

APÊNDICE C: DADOS DE CONCENTRAÇÃO E RAZÕES DE CONTAMINAÇÃO DE PROTEÍNAS (260/280) E REAGENTES (260/230) DAS AMOSTRAS DE DNA EXTRAÍDAS DE SANGUE DOS 190 PARTICIPANTES **99**

APÊNDICE D: VALORES DE FLUORESCÊNCIA (ΔRN) OBTIDOS PELA *QPCR SNP GENOTYPING* PARA O SNP RS1800592 DO GENE *UCP1*, DISCRIMINANDO OS SINAIS DOS ALELOS G (ALELO 1) E A (ALELO 2) E OS PARÂMETROS DE QUALIDADE DA REAÇÃO EM 190 AMOSTRAS ANALISADAS. **102**

APÊNDICE E: DETALHAMENTO DOS CÁLCULOS DE TEMPO E CUSTO E APLICAÇÃO DAS EQUAÇÕES EM EXEMPLOS TEÓRICOS **106**

APÊNDICE F: MEDIDAS DE DESEMPENHO DOS MODELOS DE *MACHINE LEARNING* ORGANIZADAS EM ORDEM DECRESCENTE DE ACORDO COM A FAIXA ESPECTRAL (*RANGE*) **113**

APÊNDICE G: PRINCIPAIS NÚMEROS DE ONDA COM MAIOR PODER DISCRIMINATIVO AVALIADOS PELOS VALORES DE ÁREA SOB A CURVA (AUC) EM ORDEM DECRESCENTE DE NÚMERO DE ONDA (CM^{-1}) **115**

ANEXO A: PARECER DO COMITÊ DE ÉTICA EM PESQUISA COM SERES HUMANOS – CEP **117**

1 INTRODUÇÃO

A variabilidade genética humana inclui diferentes tipos de polimorfismos, entre eles inserções e deleções, variações no número de cópias e, de forma mais frequente, os polimorfismos de nucleotídeo único (*single nucleotide polymorphisms*, SNPs). Os SNPs são variações pontuais no DNA amplamente utilizadas como marcadores genéticos, por influenciarem a regulação gênica e vias celulares (Nussbaum; Mcinnes; Willard, 2016). O mapeamento de variantes dos SNPs em conjunto com estudos de associação genética, consolidaram-se como ferramentas relevantes na biologia molecular, inclusive em estudos sobre distúrbios multifatoriais (Tam *et al*., 2019; Visscher *et al*., 2017), como a diabetes mellitus tipo 2 (DM2) e a obesidade, que se destacam pelo impacto na saúde pública mundial, com cerca de 589 milhões de adultos vivendo com DM2 em 2024 (Genitsaridi *et al*., 2026) e mais de 1 bilhão de indivíduos obesos globalmente (Janić *et al*., 2025; World Obesity Federation, 2024). Esses dados reforçam a importância do desenvolvimento de abordagens analíticas eficientes para a identificação e discriminação de variações genéticas em genes associados ao metabolismo energético em cenários de alta demanda.

Entre os genes frequentemente investigados em estudos envolvendo metabolismo energético, o gene da Proteína Desacopladora 1 (*Uncoupling Protein 1*, *UCP1*) destaca-se por seu papel na termogênese e no controle do gasto energético, especialmente no tecido adiposo marrom (Lu; Chang; Huang, 2021). O polimorfismo SNP −3826A/G (rs1800592), localizado na região promotora do gene *UCP1*, tem sido amplamente descrito na literatura por sua influência sobre a expressão gênica e por seu potencial impacto em processos metabólicos (Bouillaud; Alves-Guerra; Ricquier, 2016; Chathoth *et al*., 2018; Chaudhary; Gupta; Chauhan, 2023; Lu; Chang; Huang, 2021). Estudos prévios indicam que esse SNP pode estar relacionado a diferenças na eficiência metabólica e na regulação do metabolismo da glicose, incluindo efeitos sobre a sensibilidade à insulina (Chathoth *et al*., 2018; Gul *et al*., 2017; Pei *et al*., 2017). Este polimorfismo constitui um modelo relevante para investigações metodológicas, isoladamente, ou relacionados com distúrbios multifatoriais, por se tratar de uma variação pontual bem caracterizada, amplamente estudada e passível de validação por diferentes estratégias de genotipagem. Assim, a identificação e a diferenciação de variantes genéticas específicas, dependem de técnicas analíticas capazes de fornecer discriminação alélica confiável.

A genotipagem de SNPs foi amplamente realizada por métodos baseados em PCR seguida de análise por polimorfismo de comprimento de fragmentos de restrição (PCR-RFLP),

uma abordagem de menor custo, porém mais trabalhosa, com maior demanda de tempo e mão de obra e limitada escalabilidade para aplicações em larga escala (Mardis, 2017). Com o avanço das tecnologias moleculares, ensaios baseados em reação em cadeia da polimerase quantitativa em tempo real (qPCR) passaram a oferecer análises mais rápidas e direcionadas, com elevada acurácia e boa capacidade de discriminação alélica (Goodwin; McPherson; McCombie, 2016; Mardis, 2017). Posteriormente, métodos baseados em sequenciamento de nova geração (NGS), incluindo sequenciamento genômico completo ou direcionado, ampliaram a capacidade de obtenção de informações genéticas detalhadas e de alta resolução (Goodwin; McPherson; McCombie, 2016; Mardis, 2017). Apesar do elevado desempenho analítico, abordagens baseadas em qPCR e especialmente em NGS geralmente envolvem custos mais elevados e exigem infraestrutura especializada, sobretudo quando aplicadas a grandes volumes de amostras (Jiang *et al*., 2020; Kockum; Huang; Stridh, 2023). Essas limitações têm motivado a busca por estratégias analíticas complementares que preservem a especificidade necessária para a diferenciação alélica, ao mesmo tempo em que reduzam a complexidade operacional e os custos envolvidos.

Nesse contexto, a espectroscopia vibracional tem emergido como uma ferramenta analítica complementar em estudos biológicos. A espectroscopia no infravermelho com transformada de Fourier (FTIR) destaca-se por permitir análises rápidas, com preparo mínimo de amostras e sem o uso de reagentes químicos (Baker *et al*., 2014). A técnica fornece uma impressão digital molecular baseada nos modos vibracionais das ligações químicas, possibilitando a detecção de alterações bioquímicas e estruturais em biomoléculas complexas (Leslie *et al*., 2015; Wald *et al*., 2016). A utilização do acessório de refletância total atenuada (ATR) amplia sua aplicabilidade, ao melhorar a reprodutibilidade das análises e facilitar a avaliação de pequenos volumes de amostra (Kazarian; Chan, 2013). Nos últimos anos, a FTIR tem sido explorada na análise de ácidos nucleicos, incluindo estudos sobre a conformação do DNA, composição de bases nitrogenadas e polimorfismos SNPs (Brewer *et al*., 2002; Banyay; Sarkar; Gräslund, 2003; Emura *et al*., 2006; Mello; Vidal, 2012 ; Song *et al*., 2014; Nurdalila *et al*., 2015; Qiu *et al*., 2015; Han *et al*., 2018; Rios *et al.*, 2021; Souza *et al*., 2024). No entanto, a natureza multivariada e altamente complexa das informações espectrais geradas pelo FTIR impõe desafios significativos à extração e interpretação de informações específicas quando o interesse é avaliar alterações muito sutis, como o caso dos SNPs.

Diferentemente de abordagens analíticas tradicionais, algoritmos de aprendizado de máquina (*machine learning*, ML) têm somente agora sido utilizadas em amostras biológicas e na análise espectral de FTIR (Teklemariam *et al*., 2024; Zhang *et al*., 2024). A alta precisão do

ML a torna uma técnica mais útil na discriminação de variações sutis do genoma, como por exemplo, os diferentes tipos de SNPs, pois permite a exploração simultânea de múltiplas variáveis e a identificação de padrões associados à estrutura, composição e organização de biomoléculas em sistemas biológicos (Choi *et al*., 2020; Greener *et al*., 2022; Jiang; Gradus; Rosellini, 2020). O seu emprego poderá assim ampliar de forma significativa a capacidade interpretativa e discriminatória da espectroscopia FTIR, a fortalecendo como uma técnica complementar às análises de biologia molecular na diferenciação alélica e à análise de SNPs.

Diante desse cenário, a integração entre a espectroscopia ATR-FTIR e algoritmos de ML configura-se como uma estratégia promissora para a discriminação alélica de SNPs. A análise de sequências de DNA amplificadas por PCR de forma rápida e com menor complexidade operacional amplia a perspectiva de diferenciar variantes genéticas específicas por meio dos dados espectrais. Além disso, fatores como o tempo de análise, o custo de reagentes e simplicidade do fluxo de trabalho reforçam a relevância dessa abordagem como uma técnica complementar às metodologias convencionais de biologia molecular, especialmente em estudos exploratórios, triagens preliminares e aplicações em larga escala de amostras humanas.

2 OBJETIVOS

2.1 Principal

Avaliar a viabilidade da técnica de ATR-FTIR associada a algoritmos de ML como um método complementar quando comparado com a técnica de *qPCR SNP genotyping,* na detecção do SNP -3826A/G localizado no gene *UCP1*.

2.2 Específicos

- genotipar pela técnica de *qPCR SNP genotyping,* o SNP -3826A/G localizado no gene *UCP1* para determinação dos grupos amostrais;
- avaliar a eficiência da técnica de ATR-FTIR associada a algoritmos de ML em classificar os grupos amostrais com base nos genótipos do SNP -3826A/G no gene *UCP1*, por meio da verificação das medidas de desempenho;
- comparar a viabilidade da técnica de ATR-FTIR associada a algoritmos de ML com as técnicas de *qPCR SNP genotyping* e sequenciamento em relação ao tempo e custo necessários para a realização das metodologias.

3 REVISÃO DE LITERATURA

3.1 Polimorfismos genéticos

Polimorfismos genéticos são definidos como alterações na sequência de DNA que ocorrem dentro de genes, podendo estar localizados em regiões promotoras, codificantes (*exons*), não-codificantes (*introns*) ou em regiões intergênicas. Diferentemente das mutações, que possuem uma frequência alélica inferior a 1%, uma variante polimórfica só é considerada um polimorfismo se a sua frequência alélica for superior a 1% em toda a população mundial, independentemente do tipo de alteração que a originou, do tamanho do segmento de DNA envolvido ou de sua associação ao desenvolvimento de uma doença (Nussbaum; Mcinnes; Willard, 2016).

A identificação de polimorfismos genéticos pode ter implicações significativas para a compreensão e predição de várias doenças. O tipo e a localização do polimorfismo associado aos fatores ambientais poderão ou não influenciar no desenvolvimento de uma doença genética. Entre os polimorfismos, o tipo mais comum é o polimorfismo de nucleotídeo único (*single nucleotide polymorphism*, SNP), onde apenas um único nucleotídeo da sequência do DNA está alterado (Chiarella; Capone; Sisto, 2023). As consequências mediante a presença dos SNPs podem variar, sendo classificadas em silenciosas, quando ocorre em um códon que codifica o mesmo aminoácido, de sentido trocado (*missense*), quando há mudança no códon, codificando um aminoácido diferente do original e sem sentido (*nonsense*), quando a alteração do nucleotídeo resulta em um códon de término, encerrando a tradução do mRNA precocemente (Nussbaum; Mcinnes; Willard, 2016).

A análise de SNPs possui relevância em diversas aplicações biomédicas, destacando-se a identificação de variantes genéticas associadas à predisposição ao desenvolvimento de determinadas doenças. A detecção dessas variantes possibilita a implementação de estratégias preventivas e contribui para a definição de abordagens terapêuticas mais direcionadas e potencialmente mais eficazes (Chiarella; Capone; Sisto, 2023).

Os SNPs têm sido amplamente associados ao desenvolvimento de distúrbios multifatoriais, que constituem a forma mais comum de doenças genéticas, também denominadas de herança multifatorial ou complexa. Essas condições resultam da interação entre múltiplas variantes genéticas, fatores ambientais e estilo de vida. Entre os principais distúrbios multifatoriais destacam-se a diabetes mellitus tipo 2 (DM2) e a obesidade,

reconhecidos como importantes problemas de saúde pública devido à sua elevada prevalência e ao impacto significativo na população mundial (Nussbaum; Mcinnes; Willard, 2016).

A Diabetes Mellitus (DM) é uma patologia caracterizada por hiperglicemia crônica, resultante de secreção insuficiente ou ausente de insulina, ou pela resistência à ação desse hormônio no organismo (Wu *et al*., 2014). A doença é classificada principalmente em diabetes tipo 1 e tipo 2 (DM2), sendo a DM2 responsável por aproximadamente 90% dos casos no mundo (Ong *et al*., 2023). Embora geralmente se manifeste na fase adulta, a DM2 também pode ocorrer na adolescência, e seus fatores de risco incluem obesidade, sedentarismo e predisposições genéticas (Nelson; Cox; Hoskins, 2022; Oktavianthi *et al*., 2012). Segundo a *International Diabetes Federation* (IDF) (Genitsaridi *et al*., 2026) em 2024, estimou-se que 589 milhões de indivíduos entre 20 e 79 anos viviam com DM2 no mundo. No Brasil, 15,7 milhões de pessoas já foram diagnosticadas, com projeção de aumento para 23,2 milhões até 2045 (Genitsaridi *et al*., 2026; Ong *et al*., 2023). Além do impacto epidemiológico, a DM2 pode levar à disfunção ou falência de órgãos e tecidos incluindo olhos, rins, nervos, coração e vasos sanguíneos e apresentando sintomas como queda no desempenho físico, perda de peso e maior suscetibilidade a infecções; em quadros graves, pode evoluir para cetoacidose ou síndrome hiperosmolar não cetoacidótica, com risco de coma (Umpierrez *et al.*, 2024)

Nesse cenário, SNPs associados à DM2 podem modificar a expressão ou função de proteínas envolvidas em processos metabólicos e de sinalização celular, favorecendo condições propensas ao desenvolvimento da doença (Yanasegaran *et al*., 2024). Variantes no gene *TCF7L2* têm sido consistentemente associadas ao aumento de predisposição por influenciarem secreção e sensibilidade à insulina (Chaudhuri *et al*., 2021; Kumar *et al*., 2024a). Além disso, SNPs no gene *PPARG* têm sido relacionados a alterações na atividade do receptor, contribuindo para menor sensibilidade à insulina e maior suscetibilidade à DM2 (Chaudhuri *et al*., 2021; Hashemian *et al*., 2021). A relevância desses achados também aparece em estudos populacionais como os trabalhos que avaliaram o genótipo GG do SNP rs1801278 no gene *IRS1* e associaram a menor eficácia da resposta à insulina em populações indiana (Rasool *et al*., 2022) e paquistanesa (Albegali *et al*., 2019) , enquanto o alelo T do SNP rs7903146 no gene *TCF7L2* foi relacionado à diminuição da secreção de insulina em populações indiana (Kumar *et al*., 2024a) e mianmarense (Phu *et al*., 2023). Da mesma forma, o genótipo CC do SNP rs5219 no gene *KCNJ11* foi associado a alterações na liberação de insulina em células β pancreáticas e à DM2 na população mexicana (Díaz-García *et al*., 2024) e o alelo G do SNP rs10830963 no gene *MTNR1B*, que codifica o receptor de melatonina, foi relacionado à regulação da glicose,

resistência à insulina e DM2 em populações saudita (Kaabi, 2024) e chinesa (Li; Wang; Zhang, 2022)

A obesidade, por sua vez, é definida pelo acúmulo excessivo de gordura corporal e pela elevação do índice de massa corporal (IMC) além de valores considerados normais, podendo desencadear diversas outras patologias. A maior ocorrência de obesidade entre membros de uma mesma família decorre tanto da herança de perfis genéticos semelhantes quanto do compartilhamento de hábitos alimentares e de estilo de vida. Sedentarismo e dieta desequilibrada constituem fatores predisponentes, e quando combinados a fatores genéticos aumentam significativamente o risco dessa condição (Kapoor et al., 2020; Longo *et al*., 2019). Estimativas da World Obesity Federation indicam que, em 2024, 1 bilhão de pessoas no mundo foram classificados com obesidade grau I, correspondendo aproximadamente a uma em cada cinco mulheres e um em cada sete homens (Janić *et al*., 2025; World Obesity Federation, 2024), com projeções para alcançar 1,53 bilhões até 2035. No Brasil, dados do Sistema de Vigilância de Fatores de Risco e Proteção para Doenças Crônicas por Inquérito Telefônico (Vigitel) indicam que, em 2019, 19,8% da população apresentava IMC $\geq$ 30 kg/m², sendo classificada como obeso grau I, e evidências epidemiológicas na última década associam a obesidade ao aumento de morbidade e mortalidade (Brasil, 2023; Lobstein *et al.*, 2022).

Do ponto de vista genético, SNPs em genes específicos podem favorecer predisposição à obesidade ao influenciarem a regulação do apetite, armazenamento e metabolismo de gorduras, além da resposta hormonal, contribuindo para maior acúmulo de gordura corporal (Kim; Lee; Kim, 2018; Vourdoumpa; Paltoglou; Charmandari, 2023). Entre os genes mais investigados, destacam-se SNPs no *FTO*, associados ao maior risco de obesidade por influenciarem apetite e preferência por alimentos mais calóricos (Resende *et al*., 2021; Yin *et al*., 2023) e variantes no *MC4R*, envolvidas na regulação do apetite e do gasto energético, contribuindo para alterações na sinalização neural relacionada à saciedade e comportamento alimentar (Resende *et al*., 2021; Yu *et al*., 2020). Em diferentes populações, associações específicas têm sido relatadas. O alelo A do SNP rs9939609 no gene *FTO* foi relacionado ao aumento do apetite e à menor saciedade em populações mexicana (Martínez-López *et al*., 2024) e coreana (Park; Choi, 2023). O alelo C do SNP rs4994 no gene *ADRB3* foi associado a menor gasto energético em repouso e à obesidade em populações do leste asiático (Xie *et al*., 2020). O alelo T do SNP rs6265 no gene *BDNF* foi relacionado à obesidade na população italiana (Ricci *et al*., 2021). Além destes, o alelo G do SNP rs4846914 no gene *GALNT2* foi associado à obesidade e a baixos níveis de colesterol HDL em populações europeias (Qaddoumi *et al*., 2022).

Embora a literatura demonstre associações consistentes entre SNPs e distúrbios multifatoriais, a interpretação dessas relações permanece desafiadora devido à natureza complexa dessas condições. A contribuição de cada gene tende a ser pequena, de modo que a quantificação de efeitos requer comparação com muitos genomas e avaliação integrada de vias metabólicas, processos celulares e alterações específicas nas sequências. Assim, apesar do avanço na identificação de marcadores genéticos, permanece fundamental a investigação de metodologias que viabilizem a detecção e análise de variantes com eficiência, escalabilidade e potencial integração a rotinas laboratoriais e clínicas. Nesse contexto, genes envolvidos na regulação do metabolismo energético, como o gene da Proteína Desacopladora 1 (*Uncoupling Protein 1*, *UCP1*) ganham destaque, uma vez que variações genéticas nesse gene podem influenciar processos metabólicos fundamentais, tornando-o um candidato relevante para investigação em distúrbios multifatoriais.

3.2 Polimorfismo SNP -3826 A/G do gene *UCP1*

O gene *Uncoupling Protein 1* (*UCP1*) pertence à família de genes *UCPs* (proteínas desacopladoras, "*uncoupling proteins*") que é composta por cinco membros principais: *UCP1, UCP2, UCP3, UCP4 e UCP5*. Estes genes codificam proteínas encontradas principalmente na membrana interna das mitocôndrias e desempenham um papel essencial na regulação do metabolismo energético, dissipando o gradiente de prótons gerado pela cadeia respiratória como calor, ao invés de converter a energia em adenosina trifosfato (ATP). Além disso, exercem efeitos no controle do metabolismo de ácidos graxos, na proteção contra o acúmulo de espécies reativas de oxigênio e na termogênese, onde cada membro possui funções específicas dependendo do tecido (Bouillaud; Alves-Guerra; Ricquier, 2016).

O gene *UCP1*, localizado no cromossomo 4q28-31, está presente predominantemente no tecido adiposo marrom e é responsável por exercer função essencial na regulação da termogênese adaptativa em mamíferos, permitindo a geração de calor ao dissipar o potencial de prótons na mitocôndria, crucial para a manutenção da temperatura corporal em ambientes frios. A proteína codificada por este gene é predominantemente expressa no tecido adiposo marrom e é estimulada por hormônios como a norepinefrina, frio e ingestão de alimentos, estando envolvida na regulação do gasto energético e na queima de gordura, o que contribui para a manutenção da temperatura corporal e para o equilíbrio energético do organismo. A ativação desta proteína envolve a oxidação de ácidos graxos, liberando energia na forma de calor (Chathoth *et al*., 2018).

O mecanismo de ação da proteína é mediado pela sua capacidade de transportar prótons através da membrana mitocondrial interna, dissociando a força próton-motriz gerada pela cadeia transportadora de elétrons (Lu; Chang; Huang, 2021). Ela atua como um regulador, permitindo que a energia gerada durante a respiração celular seja desviada para a produção de calor em vez de ser utilizada exclusivamente para a síntese de ATP, fundamental para a regulação da termogênese e a queima de calorias, além de oferecer proteção contra a formação excessiva de espécies reativas de oxigênio (Chaudhary; Gupta; Chauhan, 2023).

O SNP −3826 A/G, localizado na região promotora do gene *UCP1*, corresponde à substituição de adenina (A) por guanina (G), podendo influenciar a regulação transcricional do gene e reduzir a expressão da proteína no tecido adiposo (Nicoletti *et al*., 2016). A diminuição da expressão da proteína do gene *UCP1*, que está envolvida na termogênese e na dissipação de energia na forma de calor, pode comprometer o gasto energético basal, favorecendo o acúmulo de tecido adiposo e aumentando o risco de obesidade. Considerando a estreita relação entre adiposidade excessiva, resistência à insulina e desregulação da homeostase glicêmica, alterações nesse eixo metabólico também podem contribuir para maior susceptibilidade ao desenvolvimento da DM2 (Brondani *et al*., 2012).

Na literatura, apenas um número limitado de estudos (Chathoth *et al*., 2018; Brondani *et al*., 2014; Souza *et al*., 2013; Forga *et al*., 2003; Gul *et al*., 2017; Kieć-Wilk *et al*., 2002; Lee *et al*., 2015; Nicoletti *et al*., 2016; Pei *et al*., 2017; Sramkova *et al*., 2007; Vimaleswaran *et al*., 2010) em todo o mundo investigou a relação do polimorfismo -3826A/G do gene *UCP1* com o desenvolvimento de DM2 e obesidade, abrangendo condições metabólicas distintas.

Na população chinesa, o alelo G, quando combinado a diferentes haplótipos envolvendo outros SNPs nos genes *PPARGC1A* e *UCP1*, foi relacionado ao desenvolvimento de DM2 na mesma população (Pei *et al*., 2017). Em contraste, na população brasileira, um estudo avaliando diretamente a possível associação entre o SNP −3826A/G e a predisposição à DM2 não observou relação significativa entre suas variantes e o desenvolvimento da doença (Souza *et al*., 2013). Ainda no contexto brasileiro, o alelo G foi associado a menor peso e menor gordura corporal, bem como à redução do risco de desenvolvimento de DM2 em pacientes obesos (Nicoletti *et al*., 2016). Adicionalmente, Vimaleswaran *et al*. (2010) identificaram que o alelo A, quando combinado com o alelo C do SNP 5' UTR A/C localizado no éxon 1 e com o alelo A do SNP Met299Leu (A→T) localizado no éxon 5, ambos no gene *UCP1*, esteve associado ao aumento do risco de desenvolvimento de DM2 na população indiana.

No que se refere à obesidade e aos parâmetros antropométricos e metabólicos associados, a presença do genótipo GG desse SNP foi relacionada ao desenvolvimento de

obesidade na população saudita (Chathoth *et al*., 2018), à obesidade infantil e à hipertrigliceridemia na população turca (Gul *et al*., 2017) e à hipertrigliceridemia associada a baixos níveis de colesterol HDL na população polonesa (Kieć-Wilk *et al*., 2002). A presença isolada do alelo G também foi associada ao aumento do índice de massa corporal (IMC) e da gordura corporal em indivíduos diabéticos na população tcheca (Sramkova *et al*., 2007) e em indivíduos obesos na população espanhola (Forga *et al*., 2003). Em contrapartida, na população brasileira, o genótipo GG foi associado a valores normais de IMC e a níveis elevados de colesterol HDL (Brondani *et al*., 2014). Resultados discordantes foram descritos por Lee *et al*. (2015), que identificaram que a presença do genótipo AA desse SNP, quando combinada ao genótipo CC do SNP −55C/T no gene *UCP3*, esteve associada ao aumento do IMC e da gordura corporal na população malasiana.

A identificação precisa de SNPs, como o −3826A/G no gene *UCP1*, é essencial para aprofundar a compreensão da base genética de distúrbios multifatoriais e de suas interações com fatores ambientais. A correlação entre genótipos e fenótipos depende de métodos capazes de detectar variantes com alta sensibilidade, especificidade e reprodutibilidade. Nesse contexto, a escolha da metodologia de genotipagem torna-se um fator crítico, especialmente diante das limitações relacionadas a custo, tempo de execução e exigências de infraestrutura presentes nas abordagens tradicionalmente empregadas.

3.3 Detecção de SNPs por *qPCR SNP genotyping*

A genotipagem de polimorfismos do tipo SNP tem sido realizada principalmente pela técnica de reação em cadeia da polimerase quantitativa em tempo real (*qPCR SNP genotyping*) com o objetivo de identificar e analisar essas variantes genéticas com alta sensibilidade e especificidade (Jiang *et al*., 2020). Essa metodologia permite a detecção rápida e quantitativa de pequenas alterações nucleotídicas em amostras de DNA, o que a torna essencial para a realização de associações genéticas com doenças multifatoriais e para o avanço do tratamento personalizado (Jiang *et al*., 2020). Além disso, a precisão na detecção de SNPs é importante não somente para estudos de correlação genotípica (Klusek *et al*., 2024; Bakay *et al*., 2022), mas também para diagnósticos moleculares (Ali *et al*., 2023; Gaiolla; Moraes; Oliveira, 2021; Nishita *et al*., 2009).

A técnica de *qPCR SNP genotyping* baseia-se na amplificação do DNA e na detecção de variantes específicas em tempo real, utilizando sondas fluorescentes que se ligam especificamente aos alelos de interesse. Durante a reação de PCR, a sonda fluorescente é clivada

pela atividade da Taq DNA polimerase, liberando um sinal que é detectado em tempo real, permitindo assim a distinção entre os diferentes genótipos de forma precisa e eficiente (Schleinitz; Distefano; Kovacs, 2011; Hui; DelMonte; Ranade, 2008). Embora o sequenciamento genético seja reconhecido como o padrão-ouro para a genotipagem de SNPs (Cheng; Fei; Xiao, 2023; Chan, 2009), a genotipagem pela *qPCR SNP genotyping* se destaca como uma das metodologias mais utilizadas. Isso se deve à sua elevada sensibilidade e especificidade, além da capacidade de processar muitas amostras simultaneamente. Essa abordagem permite a detecção rápida e precisa dos SNPs, o que é essencial para o acompanhamento clínico e para a tomada de decisões terapêuticas fundamentadas no perfil genético do paciente (Yu *et al*., 2018).

A genotipagem pela *qPCR SNP genotyping* tem sido amplamente utilizada em estudos voltados à identificação de SNPs associados à predisposição genética para diversas doenças multifatoriais. Por exemplo, alguns SNPs relacionados a DM2 foram identificados, como os SNPs rs2989924 e rs3758269 do gene *AQP7* na população chinesa (Wang *et al*., 2018), o rs7799039 do gene *LEP* na população egípcia (Mohamed *et al*., 2023) e os SNPs rs12778366 e rs3758391 do gene *SIRT1* na população saudita (Kaabi, 2024). Em relação à obesidade, destacam-se o SNP rs3748024 do gene *CHCHD5* na população chinesa (Wu *et al*., 2017), o rs1421085 do gene *FTO* na população mexicana (González-Herrera *et al*., 2019) e o rs228729 do gene *PER3* na população brasileira (Azevedo *et al*., 2021).

A busca por métodos mais acessíveis continua sendo essencial para ampliar o acesso à genotipagem em larga escala, especialmente em estudos populacionais e em países em desenvolvimento. Apesar de sua eficácia, a *qPCR SNP genotyping*, bem como o sequenciamento genético, são técnicas relativamente caras, o que pode limitar o uso dependendo do número de amostras e recursos financeiros obtidos. Como alternativa, a metodologia que combina a técnica de espectroscopia do infravermelho por transformada de Fourier (FTIR) com algoritmos de aprendizado de máquina (*machine learning*, ML) tem se mostrado promissora para análises de material biológico, e mais recentemente, de ácidos nucleicos para discriminação genotípica.

3.4 FTIR na análise de DNA

A espectroscopia de infravermelho por transformada de Fourier (*Fourier Transform Infrared Spectroscopy*, FTIR) é uma técnica não invasiva baseada na interação da radiação infravermelha com a matéria, analisando as vibrações moleculares que ocorrem em diferentes

ligações químicas. Essas vibrações geram um espectro de absorção característico na faixa do infravermelho médio (4000–400 cm^{-1}), que funciona como uma "impressão digital" (*fingerprint*) molecular, capaz de distinguir compostos pela composição e pela estrutura química. Por exigir preparo mínimo de amostra e oferecer tempo reduzido de análise, a técnica minimiza erros experimentais e se destaca pela sensibilidade e reprodutibilidade (Wald *et al*., 2016; Leslie *et al*., 2015).

A técnica de Refletância Total Atenuada (*Attenuated Total Reflectance*, ATR) representa uma das abordagens de amostragem mais difundidas na espectroscopia FTIR devido à sua versatilidade e simplicidade. Seu princípio baseia-se na reflexão interna total, em que a radiação infravermelha, ao incidir sobre um cristal de alto índice de refração, gera uma onda evanescente que penetra apenas superficialmente na amostra, em uma profundidade determinada pelo comprimento de onda e pelos índices de refração do cristal e da própria amostra. Essa característica assegura espectros consistentes, mesmo diante de variações físicas entre diferentes materiais sólidos. Assim, a combinação entre ATR e FTIR possibilita a análise direta de amostras líquidas, sólidas ou pastosas, reduzindo a necessidade de preparo prévio e ampliando a aplicabilidade da técnica para uma ampla gama de sistemas biológicos e não biológicos (Kazarian; Chan, 2013).

No campo biológico, o FTIR tem sido amplamente empregado no estudo de materiais orgânicos, como células e tecidos, uma vez que permite identificar modificações nos grupos funcionais de biomoléculas. Alterações em proteínas, lipídios, ácidos nucleicos e carboidratos podem ser detectadas diretamente a partir dos espectros de absorção, refletindo processos fisiológicos ou patológicos (Zhang *et al*., 2015; Lewis *et al*., 2010). Entre suas principais vantagens em relação a técnicas tradicionais de diagnóstico estão a obtenção de resultados em tempo real, a possibilidade de uso em análises não invasivas ou minimamente invasivas e a especificidade na detecção de mudanças bioquímicas que precedem alterações morfológicas ainda não manifestadas em nível micro ou macroscópico. Assim, a técnica consolidou-se como uma ferramenta promissora para aplicações biomédicas, incluindo o diagnóstico, a diferenciação e o monitoramento de doenças (Baker *et al*., 2014; Movasaghi; Rehman; Ur Rehman, 2008)

Por ser uma metodologia simples, rápida e não destrutiva (Zhang *et al*., 2015; Kahn *et al*., 2009), a espectroscopia FTIR tem sido amplamente utilizada para diferenciação bioquímica de vários tipos de amostras biológicas e separação entre diferentes grupos, devido à sua alta sensibilidade na detecção de alterações moleculares (Martinez-Marin *et al*., 2017). Por exemplo, Li *et al*. (2005) foram capazes de distinguir amostras de tecidos de câncer colorretal

de amostras de tecidos saudáveis. Rymsza *et al*. (2018) avaliaram a aplicabilidade das técnicas de PCR e da espectroscopia ATR-FTIR de forma complementar, ao analisar amostras de fluido cervical para diagnóstico da infecção genital por papiloma vírus humano (HPV). Caixeta *et al*. (2023) obtiverem sucesso em segregar amostras de saliva de indivíduos diabéticos e não diabéticos. Kino *et al*. (2024) por sua vez, destacaram alterações relevantes em amostras de plasma sanguíneo de pacientes com glioma em relação as amostras de pacientes saudáveis.

Embora a espectroscopia FTIR permite a obtenção de perfis vibracionais detalhados das amostras biológicas, a complexidade e a alta dimensionalidade dos dados espectrais frequentemente dificultam a identificação direta de diferenças sutis associadas a variantes genéticas específicas. Nesse contexto, algoritmos ML têm sido recentemente incorporados às análises espectroscópicas, possibilitando a extração de padrões não lineares e relações multivariadas que não são facilmente detectáveis pelos espectros de FTIR ou por métodos estatísticos convencionais.

3.5 Identificação de SNPs por *machine learning*

O aprendizado de máquina (*machine learning*, ML) constitui um subcampo da inteligência artificial voltado ao desenvolvimento de modelos computacionais capazes de identificar padrões em dados e utilizar essas informações para realizar classificações ou previsões. Diferentemente de abordagens baseadas em regras fixas previamente definidas, os algoritmos de ML são treinados a partir de conjuntos de dados, ajustando seus parâmetros internos com o objetivo de otimizar o desempenho em tarefas específicas (Choi *et al*., 2020). O crescimento exponencial do volume de dados biológicos e o avanço na capacidade de processamento computacional impulsionaram a evolução dessa área nas últimas décadas, possibilitando a aplicação de modelos cada vez mais sofisticados e eficientes em contextos biomédicos (Greener *et al*., 2022).

Na área da saúde, a aplicação de ML tem permitido diagnósticos mais precisos, previsões sobre a progressão de doenças e até mesmo a personalização de tratamentos (Mondal et al., 2023). Técnicas de ML, combinadas com base de dados clínicos, tem auxiliado identificar padrões associados a doenças complexas e multifatoriais, como a DM2 (Mizani *et al*., 2024; Deberneh; Kim, 2021) e a obesidade (Gutiérrez-Gallego *et al*., 2024; Safaei *et al*., 2021). Na área da biologia molecular, o uso do ML tem se mostrado promissor, especialmente na identificação de SNPs (Kumar *et al*., 2024b; Bonakdari *et al*., 2022; Cho *et al*., 2022; Aguiar-Pulido *et al*., 2010; Gaudillo *et al*., 2019; Hwa *et al*., 2019; Lu *et al*., 2012; González-Recio *et*

al., 2009), e mais recentemente, diferentes genótipos de um SNP em bovinos, foram classificados a partir de DNA amplificado e espectroscopia FTIR combinado com algoritmos de *machine learning* (Rios *et al*., 2021).

O primeiro estudo a utilizar algoritmos de ML para análise de SNPs foi o de González-Recio *et al*. (2009), que investigaram a relação de SNPs associados à artrite reumatoide em caucasianos nos Estados Unidos. O algoritmo foi capaz de pré-selecionar 1.500 SNPs na região do antígeno leucocitário humano ao fornecerem dados que continham as sequências de nucleotídeos de DNA extraído de indivíduos portadores da doença e dos indivíduos controles.

A aplicação do ML em um estudo realizado com a população espanhola foi capaz de reconhecer sequências de DNA associadas a esquizofrenia e classificar corretamente entre 78,3 e 93,8% dos indivíduos com esquizofrenia ao usar os conjuntos de dados fornecidos que continham SNPs dos genes *HTR2A* e *DRD3* previamente genotipados por *MassARRAY SNP genotyping* (Aguiar-Pulido *et al*., 2010). Em outro estudo, um modelo de predição obteve 92,4% de precisão ao classificar indivíduos esquizofrênicos por meio dos genótipos do 5-HTTLPR, um SNP do gene *SLC6A4* associado a esquizofrenia nas populações holandesa e alemã (Lu *et al*., 2012).

Algoritmos de ML também foram, foram utilizados para identificar SNPs associados a predisposição à asma, com 65,3% de precisão (Gaudillo *et al*., 2019) e à osteoartrite de joelho, com 90,5% e 85,7% de precisão em dois modelos propostos (Bonakdari *et al*., 2022). Além disso, um painel de 220 SNPs relacionados à ancestralidade foi desenvolvimento por meio de ML para identificação individual e diferenciação de origem étnica entre caucasianos e populações do leste e sudeste asiático, obtendo 88,9% de precisão (Hwa *et al*., 2019).

Estudos de associação em todo o genoma (*Genome-Wide Association Studies*, GWAS) e Análise de Componentes Principais (*Principal Component Analysis*, PCA) associados ao ML supervisionado foram empregados para refinar um total 500 SNPs da raça bovina *Tharparkar*, no Paquistão, identificando grupos de 23 a 48 SNPs ideais para o desenvolvimento do gado em termos de qualidade da carne, com taxas de precisão de 95,2 a 98,4% (Kumar *et al*., 2024b). O ML também foi aplicado para separar diferentes raças de frango na Coreia do Sul, obtendo 100% de precisão ao processar a base de dados de GWAS no DNA extraído desses espécimes (Cho *et al*., 2022).

A integração da espectroscopia FTIR com métodos de ML também tem se mostrado uma estratégia poderosa para diagnósticos biomédicos e detecção de padrões moleculares complexos. Elmi *et al*. (2017) aplicaram ATR-FTIR combinada à PCA e Análise Discriminante Linear (Linear Discriminant Analysis, LDA) para diferenciar amostras séricas de pacientes com

câncer de mama e controles, obtendo clara separação baseada em alterações proteicas. Siqueira *et al*. (2018) ampliaram esse potencial ao empregar FTIR associada a algoritmos como o Algoritmo de Projeções Sucessivas (*Successive Projections Algorithm,* SPA), Algoritmo Genético (*Genetic Algorithm,* GA) e Máquina de Vetores de Suporte (*Support Vector Machine,* SVM) na análise de tecidos prostáticos, alcançando sensibilidade de 100% e especificidade de 80%, com destaque para biomarcadores nas regiões proteicas e de ácidos nucleicos. Nogueira *et al*. (2022) demonstraram que a combinação de FTIR e o algoritmo weighted KNN (*weighted K-Nearest Neighbors*) permite classificar com alta precisão amostras de saliva de pacientes com diabetes mellitus e periodontite, reforçando o caráter não invasivo da técnica. Já Dong *et al*. (2023) utilizaram PCA e Análise Discriminante de Fisher (*Fisher's Discriminant Analysis*, FDA) aplicadas a FTIR para distinguir tecidos de linfonodos metastáticos e não metastáticos em câncer gástrico.

Até o momento, o estudo de Rios *et al*. (2021) representa a única aplicação da espectroscopia ATR-FTIR com ML como ferramenta para a estratificação de diferentes genótipos a partir de produtos amplificados por PCR de DNA humano. Os autores avaliaram o SNP C>T na posição 1547 do gene *bGH*, utilizando material genético bovino. A identificação inicial dos genótipos foi realizada por PCR associada à Análise de Polimorfismo de Comprimento de Fragmentos de Restrição (PCR-RFLP) em um conjunto de 60 amostras, composto por 20 homozigotas CC, 20 heterozigotas CT e 20 homozigotas TT. O produto amplificado foi aplicado ao cristal de ATR por sobreposição, seguido de 10 varreduras espectrais em espectrômetro FTIR. O pré-processamento dos espectros incluiu o uso de normalização por variância normal padrão (*Standard Normal Variate*, SNV) para tratamento inicial dos dados. Os espectros foram analisados por PCA para separação dos grupos e, posteriormente, classificados por algoritmos supervisionados, incluindo análises de SVM e KNN. A validação cruzada *leave-one-out* (LOO-CV) foi empregada para avaliar a precisão dos modelos. A acurácia global obtida na distinção entre os três genótipos (CC, CT e TT) foi de 75%. Os modelos KNN e SVM apresentaram os melhores desempenhos, com acurácia de até 95% na discriminação entre CD e DD, 90% entre CC e CD e 82,5% entre CC e DD.

A aplicação de algoritmos de ML na análise de SNPs tem demonstrado grande potencial para otimizar a interpretação de dados genéticos complexos, permitindo a identificação e classificação de variantes genéticas associadas a doenças multifatoriais. Ao integrar métodos de ML com dados espectrais, torna-se possível automatizar a análise, reduzir erros e aumentar a sensibilidade das análises. Essa perspectiva abre caminho para a utilização de abordagens complementares, como a espectroscopia ATR-FTIR combinada com ML, apresentando uma

alternativa promissora, rápida e de baixo custo para análises que utilizam DNA genômico para genotipagem e triagem genética em larga escala, oferecendo uma alternativa complementar para as técnicas tradicionais, tais como a *qPCR SNP genotyping* e o sequenciamento genético. Esta abordagem, ainda recente e pouco explorada para genotipagem, mostra-se promissora como uma alternativa mais acessível e integrativa para a detecção de SNPs, especialmente aqueles associados a distúrbios multifatoriais.

4 MATERIAIS E MÉTODOS

4.1 Aspectos éticos, delineamento do estudo e coleta de amostras

Este estudo foi aprovado pelo Comitê de Ética em Pesquisa (CEP) em seres humanos da Universidade do Vale do Paraíba (UNIVAP) (CAAE nº 62978022.0.0000.5503, parecer nº 7.294.487; Anexo A), estando em conformidade com a Resolução nº 466/2012 do Conselho Nacional de Saúde. Todos os participantes foram previamente informados sobre os objetivos da pesquisa e autorizaram sua participação mediante a assinatura do Termo de Consentimento Livre e Esclarecido (TCLE) (Apêndice A).

A população amostral foi composta por 190 indivíduos de ambos os sexos, com predominância do sexo feminino (69,5%) e idade média de 50,7 anos. Entre os participantes, 18,4% apresentavam diagnóstico de DM2 (n = 35), enquanto 34,7% foram classificados como obesos com base no índice de massa corporal (IMC > 30 kg/m²) (n = 66). Além disso, 12,6% dos indivíduos apresentavam histórico de câncer (n = 24), considerado neste estudo apenas quanto à presença ou ausência da condição, independentemente do tipo. Considerando a sobreposição entre condições (por exemplo, indivíduos com câncer concomitante à obesidade e/ou DM2), em que as proporções não são aditivas, no total, 53,7% (n = 102) da população apresentou pelo menos um distúrbio multifatorial relevante. Como critérios de exclusão, não foram inclusas pessoas com diagnóstico de diabetes gestacional, gestantes no momento da coleta, bem como indivíduos com idade inferior a 18 anos de idade. Com base nos dados obtidos a partir de um questionário aplicado aos participantes no momento da coleta, foi possível caracterizar a população amostral quanto às variáveis demográficas (sexo e idade) e clínicas que foram relevantes para caracterizar a amostra estudada (Apêndice B).

Amostras de sangue periférico foram coletadas no Centro de Práticas Supervisionadas (CPS) da Universidade do Vale do Paraíba (UNIVAP), seguindo procedimento operacional padrão para coleta de sangue venoso conforme as recomendações da Sociedade Brasileira de Patologia Clínica/Medicina Laboratorial (SBPC/ML, 2009). Foram coletados aproximadamente 3 a 4 mL de sangue por punção venosa em tubos a vácuo contendo ácido etilenodiamino tetra-acético (EDTA, tampa roxa). As amostras foram devidamente identificadas e armazenadas sob refrigeração a 6 °C (±4 °C) até o processamento molecular.

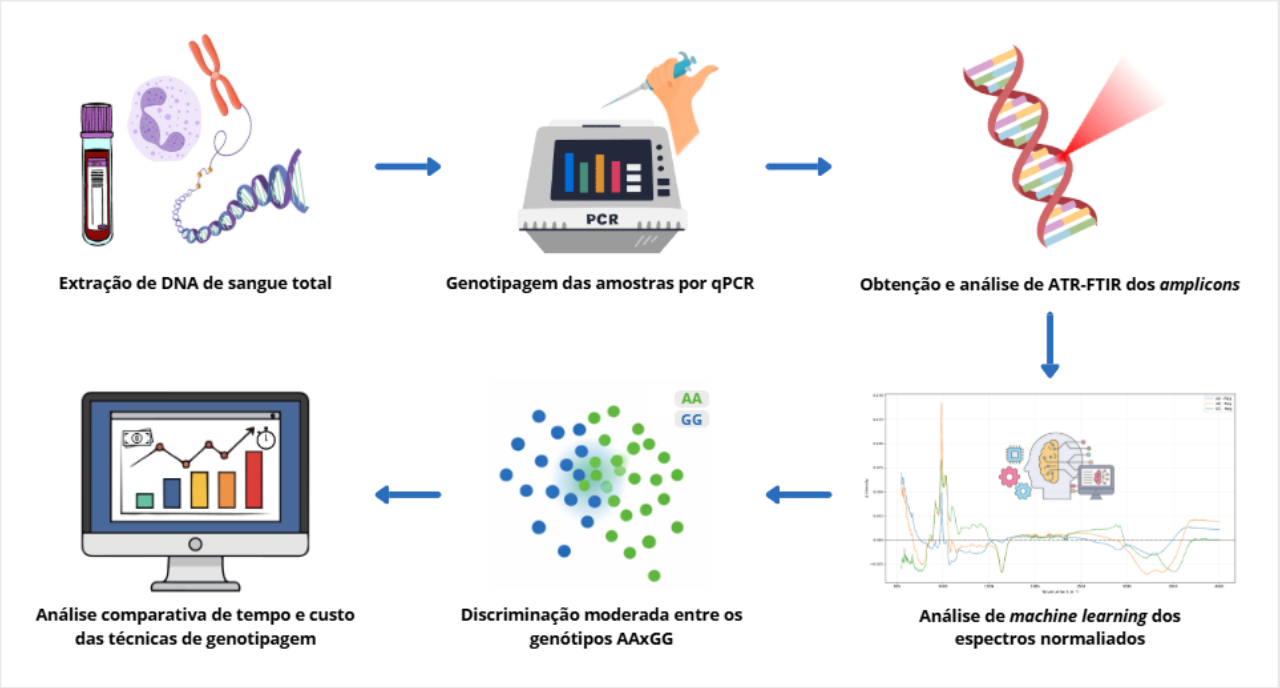
As etapas experimentais foram conduzidas no Laboratório de Genética Molecular (GeneLab) e na Central Multiusuário, ambos vinculados ao Instituto de Pesquisa e

Desenvolvimento (IP&D) da UNIVAP. A partir das amostras de sangue total coletadas em tubos contendo EDTA, procedeu-se à extração de DNA genômico de todos os 190 participantes. O DNA extraído foi utilizado para as duas técnicas de PCR: (i) genotipagem do polimorfismo −3826A/G no gene *UCP1* por PCR quantitativa em tempo real (*qPCR SNP genotyping*) e (ii) amplificação qualitativa por PCR convencional, visando à obtenção de *amplicons* para posterior aquisição espectral por espectroscopia ATR-FTIR.

É importante destacar que o mesmo conjunto de 190 amostras foi empregado de forma consistente em todas as etapas analíticas, desde a genotipagem e a análise espectral até a análise de ML. A classificação dos grupos genotípicos do SNP −3826A/G do gene *UCP1* para a posterior análise de ML foram determinadas após a análise de *qPCR SNP genotyping* e consistiram em três grupos de acordo com os diferentes genótipos obtidos, sendo AA, AG e GG.

Na Figura 1 está ilustrado o esquema experimental contendo todas os grupos analisados e os passos realizados no estudo, incluindo a extração de DNA das amostras de sangue coletadas, genotipagem pela qPCR, PCR qualitativa, aquisição dos espectros dos *amplicons* por espectroscopia ATR-FTIR, aplicação de modelos de ML e a análise comparativa entre as técnicas em relação ao tempo de processamento e custo da análise.

Figura 1. Esquema experimental dos passos realizados no estudo para a discriminação dos genótipos do SNP -3826A/G do gene *UCP1* pela análise de espectroscopia ATR-FTIR associada a *machine learning*.



Fonte: Autor.

4.2 Análise do SNP pela *qPCR SNP genotyping*

4.2.1 Extração de DNA das amostras de sangue

A extração do DNA das amostras de sangue foi realizada utilizando-se em um total de 190 amostras utilizando-se duas metodologias distintas, sendo o método de fenol:clorofórmio em 84 amostras e o método *QIAamp DNA blood Mini Kit* (*Qiagen*) em 106 amostras (Apêndice C).

As amostras de sangue extraídas pelo método de fenol:clorofórmio foram inicialmente centrifugadas para separar as fases, descartando plasma e hemácias, e mantendo-se a papa leucocitária, que foi tratada com tampão Hemálise-RBC 1X para remover hemácias residuais, com agitação e incubação à temperatura ambiente, repetindo o procedimento até que o pellet estivesse praticamente livre de hemácias. Em seguida, o pellet foi submetido à digestão com acetato de sódio 0,2 M, SDS 10% e proteinase K a 37ºC durante a noite, seguida de extrações sucessivas com fenol:clorofórmio:álcool isoamílico (25:24:1) para remover componentes celulares indesejáveis e reagentes residuais. O DNA foi então precipitado com álcool etílico absoluto gelado, armazenado a -20°C por mais uma noite, seco e ressuspendido em água livre de nucleases.

As amostras de sangue extraídas com o *QIAamp DNA blood Mini Kit* (*Qiagen*) foram processadas seguindo o protocolo do fabricante. Inicialmente, as células foram lisadas com tampão de lise contendo proteinase K para degradação de proteínas e liberação do DNA. Em seguida, a solução foi aplicada em colunas de sílica, onde o DNA se ligou à matriz, enquanto impurezas foram removidas por lavagens sucessivas com tampões específicos. Por fim, o DNA foi eluído em tampão TE obtendo-se um material genético puro e estável.

Todas as 190 amostras extraídas foram armazenadas a -20ºC para posteriores análises de *qPCR SNP genotyping* e PCR qualitativa para aquisição dos espectros de ATR-FTIR.

4.2.2 Quantificação e análises da integridade do DNA

A quantificação e pureza e a integridade das 190 amostras de DNA foram avaliadas pela espectroscopia de absorção ultravioleta no equipamento *NanoDrop* (*ND-1000 Spectrophotometer v.3.0.1, Labtrade*) e pela eletroforese em gel de agarose a 1% na *Wide Mini-Sub Cell* (*Bio-Rad*), respectivamente.

A quantificação e análise de pureza do DNA, foram realizadas com 1,5 µl de cada amostra aplicadas diretamente na célula de medição do *NanoDrop*, permitindo a determinação

da concentração de DNA pela absorbância em 260nm. A pureza das amostras foi avaliada pelas razões A260/A280 e A260/A230, indicando respectivamente, possíveis contaminações por proteínas e reagentes de extração, como fenol ou sais residuais. Valores ideais de 1,8–2,0 para A260/A280 e acima de 2,0 para A260/A230 foram utilizados como referência para confirmar a integridade e a qualidade do DNA, garantindo que as amostras fossem adequadas para análises subsequentes (Apêndice C).

O gel utilizado na eletroforese foi preparado dissolvendo 1g de agarose de grau molecular em 100 mL de tampão TBE 1X (Tris base, ácido bórico e EDTA), ao qual foram adicionados aproximadamente 4 µL de brometo de etídio para visualização das bandas. Em cada poço foram carregados 200 ng de DNA previamente misturados com volume adequado de corante de carga, enquanto os poços extremos, direito e esquerdo, continham o marcador de peso molecular de 100bp igualmente misturado com o tampão de carregamento. A corrida eletroforética foi conduzida sob tensão constante de 120V durante 1 hora e 20 minutos, permitindo a migração das moléculas de DNA e a posterior avaliação de sua integridade, para a identificação de possível fragmentação ou degradação do DNA por meio do padrão observado em comparação com o marcador. A incidência da luz ultravioleta (UV), que permite a visualização das bandas no gel, foi controlada pelo equipamento *Transilluminator-D Pro MiniBIS* (*DNR Bio-Imaging Systems Ltd*) e a captura da imagem foi obtida utilizando-se o software *GelCapture* para posterior análise.

4.2.3 Genotipagem pela *qPCR SNP genotyping*

A genotipagem do SNP -3826A/G (rs1800592) do *UCP1* foi realizada utilizando-se o ensaio *TaqMan SNP Genotyping Assays* de discriminação alélica usando conjuntos de *primers*-sondas (*Taqman SNP assays MTO Human SM 10, Thermo Fisher Scientific*).

O ensaio incluiu iniciadores não marcados com duas sondas de oligonucleotídeos *TaqMan* fluorescentes (sonda específica para alelo 1 marcada com fluoróforo VIC e sonda específica para o alelo 2 marcada com FAM). Os corantes *reporters* VIC e FAM são ligados covalentemente à base 5' terminal das duas sondas e o componente *quencher* não fluorescente é ligado próximo a extremidade 3′. As sondas foram capazes de ligarem-se diferencialmente aos *amplicons* gerados durante a PCR selecionando os seus respectivos alelos. As informações respectivas ao SNP estudado estão descritas na Tabela 1.

Tabela 1. Dados técnicos sobre o SNP -3826A/G do gene *UCP1* analisado pela *qPCR SNP genotyping.*

Gene	Gene ID	*SNP* ID	Polimorfismo	Localização do polimorfismo no gene	Localização do polimorfismo na sequência de DNA
Uncoupling Protein 1 (UCP1)	7350	rs1800592 (C_8866368_20)	T/C, Transição, Substituição	-3826 A/G Cr 4 Intragênica	TGTAGAACACATTAACAAATGC ACT**[T/C]**GATCAAACTGTGGTC AATCAGAAAT

Todas as reações de *qPCR SNP genotyping* foram realizadas em placas de 96 poços e constituídas por 20 ng de DNA de cada participante contido em 1 µl de água ultrapura, 10 µl de PCR *master mix TaqMan* universal, 1 µl de mix de primers-sondas específicos e 8 µl de água ultrapura para reações, totalizando 20 µl de volume final para cada reação. Os controles negativos (*negative template control,* NTC) também foram analisados. A PCR em tempo real foi realizada no equipamento *ABI Prism 7500* (*Thermo Scientific*) nas seguintes condições: 50°C por 2 minutos, 95°C por 10 minutos e então 40 ciclos de amplificação (92°C de desnaturação por 15 segundos, 62°C de anelamento/extensão por 60 segundos). A temperatura de anelamento foi determinada empiricamente para promover ligação altamente específica das sondas sem a perda de sensibilidade do ensaio. Para cada ciclo, o software SDS determinou o ΔRn, que é o sinal fluorescente normalizado (isto é, em comparação com um fluoróforo de referência passivo) da sonda marcada com VIC ou FAM. Para esta análise foi utilizado o valor de ΔRn após o último ciclo, pois é mais confiável que o ciclo *threshold* (valor CT). A razão do sinal do fluoróforo ΔRnFAM/ΔRnVIC (FAM/VIC) foi calculado para cada amostra para determinação final dos genótipos (Apêndice D).

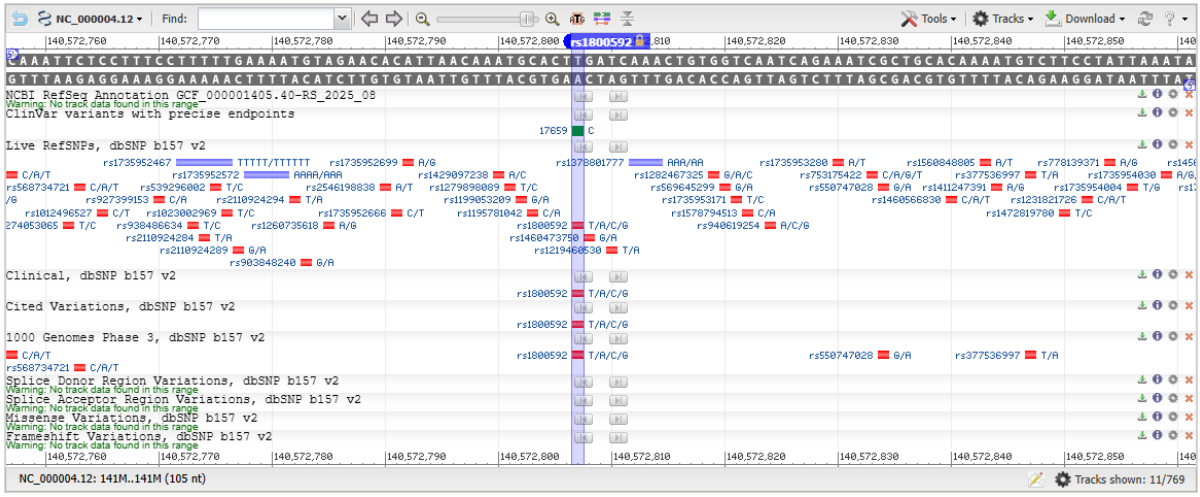
Os cálculos de frequências alélicas e a avaliação do equilíbrio de *Hardy-Weinberg* (HWE) também foram realizados, com o objetivo de caracterizar a distribuição genética das variantes estudadas na população analisada. Essa abordagem permitiu verificar a consistência interna dos dados e detectar possíveis desvios de frequências esperadas, garantindo a confiabilidade das análises. Além disso, as frequências alélicas observadas foram comparadas com os valores de referência disponíveis no banco de dados dbSNP do Centro Nacional de Informações de Biotecnologia (*Nacional Center for Biotechnology Information,* NCBI) (https://www.ncbi.nlm.nih.gov/snp/rs1800592), permitindo avaliar a concordância dos resultados obtidos com os padrões populacionais previamente reportados na literatura.

4.3 Análise dos espectros de ATR-FTIR dos *amplicons*

4.3.1 Obtenção dos amplicons por PCR qualitativa

Após a verificação da integridade e pureza do DNA extraído de cada amostra, foi realizada a PCR qualitativa para a amplificação da região de interesse (*amplicons*) do SNP -3826A/G do gene *UCP1*. A Figura 2 apresenta a localização genômica do SNP −3826A/G (rs1800592) no gene *UCP1*, destacando sua posição na região promotora do gene (NC_000004.12, dbSNP, NCBI). A representação gráfica evidencia o posicionamento do nucleotídeo variante (A/G) dentro da sequência genômica, bem como sua localização em relação a outras variantes descritas na mesma região.

Figura 2. Representação da localização genômica do SNP −3826A/G (rs1800592) no gene *UCP1* (em azul) de acordo com o NCBI (NC_000004.12, dbSNP, NCBI). Nota-se que outros polimorfismos descritos também podem ser observados na mesma região.



Fonte: SNP(rs)Report, 2024.

Os oligonucleotídeos (*primers*) utilizados para a amplificação da região contendo o SNP −3826A/G foram desenhados utilizando a ferramenta *Primer-BLAST* disponível no NCBI, considerando como referência a sequência genômica humana disponível no banco de dados público. O desenho dos *primers* foi realizado de modo a flanquear especificamente o SNP alvo −3826A/G (rs1800592). No desenho dos *primers* foram considerados parâmetros como especificidade, ausência de formação significativa de dímeros e estruturas secundárias, além de temperatura de *melting* (Tm) compatível entre os *primers forward* e *reverse*. O par de *primers* selecionado permitiu a obtenção de um *amplicon* com comprimento de 153 pares de bases (pb), que flanqueia o SNP de interesse.

A PCR foi realizada em um termociclador (*Veriti™ 96-Well Thermal Cycler, Thermo Fisher Scientific*) com um volume final de 25 µl, utilizando 100 ng de DNA, composto por 50 mM de KCl, 10 mM de Tris-HCl (pH 8,3), 1,5 mM de MgCl2, 200 µM de desoxinucleotídeotrifosfatados (dNTP), 0,6 µM de cada um dos *primers* (*foward* e *reverse*) e 2 U de Taq DNA polimerase. Os NTCs foram incluídos em cada lote de PCR, totalizando 19 reações de NTC, sendo uma para cada lote de amplificação realizado. As sequências dos *primers* e as condições de amplificação estão descritas na Tabela 2.

Tabela 2. Condições de reação da PCR qualitativa para amplificação do SNP -3826A/G do gene *UCP1*

Região de interesse e tamanho do *amplicon*	**Primer *foward* (*sense*)**	**Primer *reverse* (*antisense*)**	**Desnaturação inicial**	**Etapas de 35 ciclos**			**Extensão Final**
				Desnaturação	**Anelamento**	**Extensão**	
SNP -3826A/G *UCP1* (153np)	5′-GTCAGTATGAGCAAGGGCAAC-3′	5′-GTGCAGCGATTTCTGATTGACC-3′	94ºC – 5min.	94ºC – 1min.	58ºC – 45s	72ºC – 1min.	72ºC – 6min.

Os produtos amplificados obtidos pela PCR foram analisados em eletroforese em gel de agarose a 1,6%. O gel foi preparado dissolvendo 1,6g de agarose de grau molecular em 100 mL de tampão TBE 1X (Tris base, ácido bórico e EDTA), ao qual foram adicionados 4 µL de brometo de etídio para visualização das bandas. Em cada poço foram carregados 5 µL de cada produto amplificado pela PCR, previamente misturados com o corante de carga, enquanto o primeiro poço, continha o marcador de peso molecular de 50bp misturado com o corante de carga. A corrida eletroforética foi conduzida sob tensão constante de 120V durante 1 hora e 20 minutos. A incidência da luz ultravioleta (UV), que permite a visualização das bandas no gel, foi controlada pelo equipamento *Transilluminator-D Pro MiniBIS* (*DNR Bio-Imaging Systems Ltd*) e a captura da imagem foi obtida utilizando-se o software *GelCapture* para posterior análise.

4.3.2 Obtenção dos espectros dos amplicons por ATR-FTIR

Os *amplicons* de PCR obtidos a partir das 190 amostras, previamente genotipadas por *qPCR SNP genotyping*, foram posteriormente analisados por espectroscopia ATR-FTIR. Os espectros de ATR-FTIR também foram obtidos a partir de todos os produtos de PCR dos NTCs, constituindo as referências espectrais negativas. Essa estratégia possibilitou a identificação de características espectrais provenientes exclusivamente dos reagentes da PCR e após a subtração

do espectro médio dos NTCs, facilitou a distinção das contribuições espectrais específicas do DNA nas análises subsequentes.

A aquisição espectral foi realizada utilizando um espectrômetro (*Spectrum Two FT-IR, PerkinElmer*), equipado com acessório de ATR de diamante e controlado pelo software *Spectrum IR* (*PerkinElmer*) (versão 10.7.2). Para garantir a reprodutibilidade, um protocolo de medição idêntico e padronizado foi aplicado a todas as amostras. Espectros de fundo (*background*) foram coletados com 32 varreduras (*scans*) no início das análises e repetidos periodicamente ao longo das medições, com o objetivo de minimizar a influência de gases atmosféricos e da umidade ambiente.

Para cada medição, 10 µL da solução contendo o *amplicon* de PCR foram depositados sobre o cristal de diamante do acessório ATR e secos à temperatura ambiente utilizando um fluxo suave de ar pressurizado. Após a evaporação completa da solução, foi aplicada pressão controlada (80–100 *newtons*, N) por meio do pistão de torque do equipamento, com a finalidade de melhorar o contato da amostra com o cristal, promover a formação uniforme do filme seco e eliminar microgotículas residuais que poderiam comprometer a qualidade espectral. Esse procedimento assegurou interação otimizada entre o feixe de radiação infravermelha e o filme seco de DNA. A aquisição espectral foi iniciada somente após a secagem completa da amostra, verificada pela definição clara das bandas de absorção características e pela ausência de sinais associados à água residual. Em particular, a redução das bandas largas de estiramento O–H na região aproximada de 3200–3600 cm^{-1} e das vibrações de deformação próximas a 1600–1650 cm^{-1} foi utilizada como indicador da remoção eficaz da água.

Os espectros de absorbância ATR-FTIR foram adquiridos utilizando 16 *scans* na região do infravermelho médio (4000–550 cm^{-1}), com resolução espectral de 4 cm^{-1}. Todos os espectros foram coletados sob condições experimentais idênticas e posteriormente utilizados nas análises multivariadas e nos modelos de ML.

4.3.3 Análise de machine learning não supervisionado

Os espectros de ATR-FTIR dos *amplicons* de PCR das 190 amostras foram inicialmente analisados em sua forma bruta, obtendo-se um único espectro médio representativo de todas amostras sem tratamento prévio para analisar o perfil espectral de maneira qualitativa e realizar a atribuição preliminar das bandas de acordo com estudos da literatura sobre espectroscopia vibracional de ácidos nucleicos (Souza *et al*., 2024; Mello; Vidal, 2012; Banyay; Sarkar;

Gräslund, 2003; Brewer *et al*., 2002). Essa etapa permitiu identificar bandas na região de *fingerprint* que consiste na extensão de 1800 a 800 cm^{-1}, e associadas a modos vibracionais característicos do DNA, tais como o esqueleto fosfato, bases nitrogenadas e vibrações relacionadas à conformação da molécula. Os espectros médios de absorbância para cada grupo genotípico, definidos após a análise de *qPCR SNP genotyping,* e que consistiram em 88 amostras com o genótipo AA, 80 amostras com o genótipo AG e 22 amostras com o genótipo GG), foram calculados utilizando-se inicialmente os espectros brutos. Em seguida, o desvio padrão de cada número de onda foi calculado para cada um dos grupos, permitindo a construção das curvas médias acompanhadas de área sombreada, representando a variabilidade intergrupo. Essa abordagem possibilitou a comparação visual mais detalhada entre os grupos genotípicos, tanto em termos de tendência central quanto de dispersão espectral.

Os espectros médios foram submetidos à normalização pelo método de Variância Normal Padrão (*Standard Normal Variate*, SNV) para avaliar a ocorrência de diferenças sutis que podem ser mascaradas pelas variações de intensidade global ou efeitos multiplicativos. Após a normalização, o procedimento de cálculo de médias e desvios padrão foi repetido, permitindo-se comparar os perfis espectrais antes e depois do pré-processamento e verificar o impacto da correção sobre a evidência de possíveis diferenças associadas aos genótipos. Após a etapa de normalização dos espectros por SNV, os espectros médios de absorbância foram calculados para cada grupo genotípico (AA, AG e GG), bem como para os NTCs. Em seguida, os valores dos espectros médios foram utilizados para a visualização qualitativa das tendências espectrais e as próximas etapas.

O espectro médio do grupo NTC foi subtraído dos espectros médios de cada grupo genotípico com o objetivo de identificar as características espectrais que estão associadas a sequência de DNA do *amplicon* que contêm o SNP e também para reduzir a contribuição dos reagentes utilizados na reação de PCR. Os espectros diferenciais resultantes foram analisados para a identificação de regiões espectrais que apresentam maiores variações associadas ao grupo genotípico. Com base nessa análise de subtração, intervalos espectrais específicos que demonstraram diferenças pronunciadas foram selecionados para as análises de ML não supervisionada e supervisionada.

A análise dos componentes principais (PCA) foi utilizada como uma ferramenta multivariada exploratória utilizando-se os valores dos espectros normalizados pela SNV, com o objetivo de avaliar a estrutura intrínseca dos dados espectrais, a distribuição da variância e as possíveis tendências de agrupamento entre os grupos genotípicos. A PCA foi aplicada aos intervalos espectrais pré-selecionados, o que permite a visualização da distribuição das

amostras em espaço de dimensionalidade reduzida e fornece informações sobre a variabilidade espectral associada aos genótipos. O número de componentes principais (*principal* componente, PC) potencialmente relevante foi inicialmente avaliado por meio do critério de Kaiser (autovalores maiores que 1), algoritmo empregado exclusivamente como referência metodológica para interpretação da variância explicada.

Todo o pré-processamento espectral, os cálculos estatísticos descritivos, a geração dos espectros médios e diferenciais, bem como as análises de ML não supervisionado, foram realizados por meio de *scripts* personalizados desenvolvidos na linguagem *Python* (versão 3.13.5). Para a execução das rotinas computacionais, foram empregadas bibliotecas amplamente utilizadas em ciência de dados e análise espectral, incluindo *NumPy* para operações numéricas matriciais e processamento dos espectros, *pandas* para organização e manipulação dos dados, *scikit-learn* para a implementação da análise por componentes principais (PCA) e cálculo da variância explicada, além de *Matplotlib* para a construção dos gráficos espectrais, curvas médias com área sombreada e visualização dos agrupamentos multivariados.

4.3.4 Análise de machine learning supervisionado

Modelos supervisionados de aprendizado de máquina (ML) e aprendizado profundo (deep learning, DL) foram aplicados para a classificação dos grupos genotípicos utilizando os espectros normalizados por SNV. Os modelos foram avaliados por meio de estratégia de reamostragem aleatória repetida, correspondente à validação cruzada de Monte Carlo, adotada para minimizar possíveis vieses decorrentes do desbalanceamento entre os três grupos genotípicos (AA: 88 amostras; AG: 80 amostras; GG: 22 amostras). Em cada uma das 50 iterações independentes, o conjunto de dados foi inicialmente particionado aleatoriamente em 80% das amostras para desenvolvimento do modelo e 20% para teste. Posteriormente, dentro dos 80% destinados ao desenvolvimento, os dados foram subdivididos em 60% para treinamento e 20% para validação, utilizados para ajuste e seleção dos modelos. O conjunto de teste, correspondente aos 20% restantes, foi mantido independente e empregado exclusivamente para a avaliação final do desempenho, garantindo que as métricas fossem obtidas a partir de dados não utilizados no treinamento ou na validação.

Quatro algoritmos clássicos de ML foram implementados: Máquina de Vetores de Suporte linear (*Linear Support Machine Vector*, SVM Linear), Máquina de Vetores de Suporte com Kernel de base radial (*Support Vector Machine with Radial Basis Function Kernel*, SVM

RBF Kernel), Análise Discriminante Linear (*Linear Discriminant Analysis*, LDA) e Regressão Logística. O SVM Linear foi empregado para investigar separabilidade linear entre genótipos em espaço espectral de alta dimensionalidade, enquanto o SVM RBF Kernel abordou relações não lineares por meio da projeção implícita dos dados em espaço de maior dimensionalidade. O LDA foi aplicado com o objetivo de maximizar a variância entre classes em relação à variância intraclasse, sob a suposição de distribuição aproximadamente normal das variáveis e a Regressão Logística foi utilizada como modelo linear probabilístico de referência.

Dois modelos de DL baseados em arquiteturas de Perceptron Multicamadas (*Multilayer Perceptron*, MLP) foram avaliados. A primeira configuração, denominada Aprendizado Profundo com Perceptron Multicamadas em Arquitetura Funil (*Deep Learning with Multilayer Perceptron Funnel Architecture*, DL-MLP Funnel), empregou arquitetura em formato de funil, com camadas ocultas progressivamente decrescentes (128, 64 e 32 neurônios), projetada para comprimir a informação espectral e enfatizar características biológicas dominantes, reduzindo simultaneamente ruído de alta frequência. A segunda configuração, denominada Aprendizado Profundo com Perceptron Multicamadas com Conexões Residuais Simuladas (*Deep Learning with Multilayer Perceptron with Simulated Residual Connections*, DL-MLP Residual Simulated) utilizou arquitetura de profundidade uniforme com três camadas ocultas de igual tamanho (100, 100 e 100 neurônios), simulando comportamento semelhante a conexões residuais e preservando elevada capacidade de representação ao longo das camadas, com o objetivo de capturar correlações não lineares sutis entre regiões espectrais distantes. Funções de ativação do tipo *Rectified Linear Unit* (ReLU) foram empregadas em todas as camadas ocultas.

O desempenho dos modelos foi avaliado por meio de curvas Características de Operação do Receptor *(Receiver Operating Characteristic*, ROC), com cálculo da área sob a curva (*Area Under the Curve*, AUC) para cada combinação de intervalo espectral e método de classificação ao longo das iterações. Os limiares ótimos de classificação foram determinados utilizando o índice de Youden (J), de forma a equilibrar sensibilidade e especificidade. Nos limiares definidos, foram calculados a sensibilidade, especificidade, acurácia e F1-score, denominadas medidas de desempenho e foram expressas como valores médios obtidos ao longo das 50 iterações para cada modelo e intervalo espectral analisado. A adoção de 50 iterações teve como objetivo garantir maior robustez estatística à avaliação, minimizando o efeito de variações decorrentes de particionamentos aleatórios dos conjuntos de treinamento e teste. Esse número de repetições foi considerado suficiente para obter estimativas médias estáveis das métricas de

desempenho, reduzindo a possibilidade de viés associado a um modelo excessivamente ajustado a uma única amostragem dos dados.

A importância das variáveis e a relevância espectral foram avaliadas por meio de abordagens baseadas em permutação, adaptadas a cada estrutura de modelagem. Para modelos lineares, incluindo Regressão Logística, a importância das variáveis foi inferida a partir da sensibilidade do desempenho do modelo a variações de intensidade em números de onda individuais, evidenciando regiões espectrais capazes de promover separação linear entre genótipos. Para os modelos de DL, a importância por permutação foi empregada para identificar regiões espectrais que contribuíssem para fronteiras de decisão não lineares, refletindo interações complexas entre múltiplos modos vibracionais associados a características conformacionais e estruturais do DNA.

Todo o pré-processamento espectral e as implementações dos modelos de ML foram realizados por meio de *scripts* personalizados desenvolvidos na linguagem *Python* (versão 3.13.5). Para a execução das rotinas computacionais, foram empregadas bibliotecas amplamente utilizadas em ciência de dados e ML, incluindo *NumPy* para operações numéricas matriciais, *pandas* para organização e manipulação dos dados, *scikit-learn* para implementação dos algoritmos clássicos de ML, particionamento dos conjuntos amostrais e cálculo das medidas de desempenho, bem como *TensorFlow/Keras* para a construção e treinamento das arquiteturas de DL baseadas em perceptron multicamadas. As bibliotecas *SciPy* e *Matplotlib* foram utilizadas para análises complementares e visualização gráfica dos resultados.

4.4 Análise de viabilidade da espectroscopia ATR-FTIR associada ao ML

Além da avaliação da eficiência operacional, a análise de viabilidade de uma técnica também se fundamenta no tempo requerido e em todos os custos envolvidos para a sua execução, permitindo-se assim a determinação da sua aplicabilidade e relação custo-benefício.

A viabilidade da aplicação da técnica de ATR-FTIR associada ao ML utilizada neste estudo para a classificação genotípica do SNP −3826A/G do gene UCP1 foi avaliada por meio da comparação do tempo e custo de sua aplicação em relação ao sequenciamento de nova geração (NGS) e à qPCR SNP genotyping com sondas TaqMan, metodologias de biologia molecular consideradas padrão ouro para a genotipagem de SNPs. Nesta análise, todos os fluxos de trabalho foram considerados a partir do DNA genômico previamente extraído.

Os tempos empregados para a análise comparativa entre a qPCR SNP genotyping e a espectroscopia ATR-FTIR associada ao ML corresponderam aos tempos experimentais observados neste estudo para a execução de todas as etapas de ambas as técnicas. No caso da abordagem de ATR-FTIR associada ao ML, as estimativas foram realizadas considerando um algoritmo previamente desenvolvido, treinado e validado, em condição operacional, de modo que o tempo computacional incluído na análise correspondeu à etapa de inferência e classificação das amostras, e não ao processo inicial de desenvolvimento, treinamento e otimização do modelo. Dessa forma, a estimativa foi construída com base em um cenário de aplicação prática em rotina laboratorial.

Os custos, foram definidos com base nos valores dos reagentes de diferentes empresas utilizados em cada uma das etapas de ambas as metodologias. As estimativas de tempo e custo necessários para a técnica de NGS foram obtidas de estudos prévios da literatura (Cheng; Fei; Xiao, 2023; Kockum; Huang; Stridh, 2023; Rios *et al*., 2021; Green; Sambrook, 2019; Mardis, 2017; Goodwin; McPherso; McCombie, 2016; Lorenz, 2012; Chan, 2009) e das empresas de biotecnologia consolidadas na área (*Thermo Fisher Scientific*, *Illumina*, *Qiagen*, *Sigma-Aldrich* e *Promega*), possibilitando assim uma comparação padronizada entre as etapas experimentais de cada metodologia que possui requisitos técnicos e infraestruturais distintos.

Os cálculos de tempo e custo foram realizados utilizando tamanhos de lote idênticos para todos os fluxos de trabalho avaliados, permitindo comparação direta da eficiência operacional e da escalabilidade. Foi adotado um tamanho de lote de 96 amostras, refletindo o uso padronizado de placas de 96 poços em laboratórios de biologia molecular e em rotinas de genotipagem. Custos relacionados a depreciação de equipamentos, infraestrutura laboratorial e mão de obra não foram incluídos na análise, devido à elevada variabilidade institucional desses parâmetros.

O presente estudo propõe duas métricas complementares para a avaliação temporal das técnicas analisadas, com o objetivo de quantificar tanto o esforço operacional do analista quanto o tempo total necessário para a conclusão do fluxo experimental. Inicialmente, foi definido o tempo de execução manual (TEM), que corresponde ao tempo efetivamente despendido pelo operador na realização das etapas experimentais, incluindo atividades como preparo e processamento das amostras, montagem das reações e configuração dos equipamentos, conforme descrito na Equação 1.

Equação 1. Tempo de Execução manual a nível de lote

$$TEM_{lote(N)} = \sum_{k=1}^{m} t_k^{manual} \quad (1)$$

- $TEM_{lote(N)}$: tempo de execução manual do lote contendo N amostras;
- N: número total de amostras processadas no lote;
- k: índice correspondente à etapa experimental;
- m: número total de etapas do fluxo de trabalho;
- t_k^{manual}: tempo de execução manual referente à etapa k.

Em seguida, foi considerado o tempo de operação instrumental (TOI), definido como o período em que os equipamentos permanecem em funcionamento para a execução das análises, sem a necessidade de intervenção direta do operador, conforme apresentado na Equação 2.

Equação 2. Tempo de Operação Instrumental a nível de lote

$$TOI_{lote(N)} = \sum_{k=1}^{m} t_k^{instrumental} \quad (2)$$

- $TOI_{lote(N)}$: tempo de operação instrumental contendo N amostras;
- N: número total de amostras processadas no lote;
- k: índice correspondente à etapa experimental;
- m: número total de etapas do fluxo de trabalho;
- t_k^{manual}: tempo de execução manual referente à etapa k.

Assim, o tempo total de processamento (TTP) foi estabelecido como a soma do tempo de execução manual e do tempo de operação instrumental, representando o tempo total decorrido desde o início até a conclusão do processamento do lote de amostras, conforme demonstrado na Equação 3.

Equação 3. Tempo Total de Processamento a nível de lote

$$TTP_{lote(N)} = TEM_{lote(N)} + TOI_{lote(N)} \quad (3)$$

- $TTP_{lote(N)}$: tempo total de processamento contendo *N* amostras;
- $TEM_{lote(N)}$: tempo de execução manual do lote contendo *N* amostras;
- $TOI_{lote(N)}$: tempo de operação instrumental contendo *N* amostras;

Para um lote contendo *N* amostras e um fluxo experimental composto por m etapas, as métricas foram calculadas pela soma dos tempos específicos de cada etapa, em que t_k^{manual} representa o tempo de execução manual referente à etapa k, e $t_k^{instrumental}$ corresponde ao tempo de operação instrumental da mesma etapa.

Com o objetivo de permitir a comparação padronizada entre os diferentes fluxos de trabalho e os distintos tamanhos de lote empregados nas três metodologias analisadas, as métricas temporais previamente definidas em nível de lote foram também expressas em nível de amostra. Para isso, os valores obtidos para o tempo de execução manual (TEM), tempo de operação instrumental (TOI) e tempo total de processamento (TTP) foram normalizados pelo número total de amostras processadas no lote (*N*), conforme apresentado nas Equações 4, 5 e 6, respectivamente. Essa abordagem permite estimar o tempo médio demandado por amostra, favorecendo uma comparação mais direta entre as metodologias, independentemente da quantidade total de amostras analisadas em cada execução experimental.

Equação 4. Tempo de Execução manual a nível de amostra

$$TEM_{por\ amostra\,(N)} = \frac{1}{N} \sum_{k=1}^{m} t_k^{manual} \tag{4}$$

- $TEM_{por\ amostra(N)}$: tempo de execução manual do lote contendo *N* amostras;
- *N*: número total de amostras processadas no lote;
- *k*: índice correspondente à etapa experimental;
- *m*: número total de etapas do fluxo de trabalho;
- t_k^{manual}: tempo de execução manual referente à etapa *k*.

Equação 5. Tempo de Operação Instrumental a nível de amostra

$$TOI_{por\ amostra\,(N)} = \frac{1}{N} \sum_{k=1}^{m} t_k^{instrumental} \tag{5}$$

- $TOI_{por\ amostra(N)}$: tempo de operação instrumental contendo *N* amostras;
- *N*: número total de amostras processadas no lote;
- *k*: índice correspondente à etapa experimental;
- *m*: número total de etapas do fluxo de trabalho;
- t_k^{manual}: tempo de execução manual referente à etapa *k*.

Equação 6. Tempo Total de Processamento a nível de amostra

$$TTP_{por\ amostra(N)} = TEM_{por\ amostra(N)} + TOI_{por\ amostra(N)} \quad (6)$$

- $TTP_{por\ amostra(N)}$: tempo total de processamento contendo *N* amostras;
- $TEM_{por\ amostra(N)}$: tempo de execução manual do lote contendo *N* amostras;
- $TOI_{por\ amostra(N)}$: tempo de operação instrumental contendo *N* amostras;

O custo dos insumos foi estimado por meio de uma abordagem de custeio baseado em atividades, na qual o custo total de cada metodologia foi determinado a partir da soma dos custos de todos os reagentes e materiais consumíveis empregados ao longo das etapas experimentais. Para essa estimativa, os preços unitários ($P_{k,j}$) foram definidos com base nos valores médios praticados pelas empresas consultadas, sendo todos os custos expressos em dólares americanos (USD).

Para um fluxo de trabalho composto por *m* etapas experimentais, o custo total em nível de lote foi calculado conforme a Equação 7, na qual $q_{k,j}$ representa a quantidade do reagente ou consumível *j* utilizado na etapa k, e $P_{k,j}$ corresponde ao seu respectivo preço unitário. Essa formulação também contempla o custo dos controles experimentais negativos, uma vez que esses componentes estão incluídos na quantidade total de insumos consumidos por lote.

Equação 7. Custo total da metodologia a nível de lote

$$Custo_{lote(N)} = \sum_{k=1}^{m} \sum_{j=1}^{p_k} q_{k,j} \cdot P_{k,j} \quad (7)$$

- $Custo_{lote(N)}$: custo total dos insumos utilizados no processamento de um lote contendo *N* amostras;
- *N*: número total de amostras processadas no lote;
- *k*: índice correspondente à etapa experimental;

- m: número total de etapas do fluxo de trabalho;
- j: índice correspondente ao reagente ou consumível utilizado em cada etapa;
- p_k: número total de reagentes ou consumíveis empregados na etapa k;
- $q_{k,j}$: quantidade do reagente ou consumível j utilizada na etapa k;
- $P_{k,j}$: preço unitário do reagente ou consumível j na etapa k, expresso em USD.

Com o objetivo de permitir a comparação entre metodologias com diferentes tamanhos de lote, o custo também foi expresso em nível de amostra, por meio da normalização do custo total do lote pelo número total de amostras processadas (N), conforme descrito na Equação 8.

Equação 8. Custo total da metodologia a nível de amostra

$$Custo_{por\ amostra(N)} = \frac{1}{\mathrm{N}} \sum_{k=1}^{m} \sum_{j=1}^{p_k} q_{k,j} \cdot P_{k,j} \tag{8}$$

- $Custo_{por\ amostra(N)}$: custo médio dos insumos por amostra em um lote contendo N amostras;
- N: número total de amostras processadas no lote;
- k: índice correspondente à etapa experimental;
- m: número total de etapas do fluxo de trabalho;
- j: índice correspondente ao reagente ou consumível utilizado em cada etapa;
- p_k: número total de reagentes ou consumíveis empregados na etapa k;
- $q_{k,j}$: quantidade do reagente ou consumível j utilizada na etapa k;
- $P_{k,j}$: preço unitário do reagente ou consumível j na etapa k, expresso em USD.

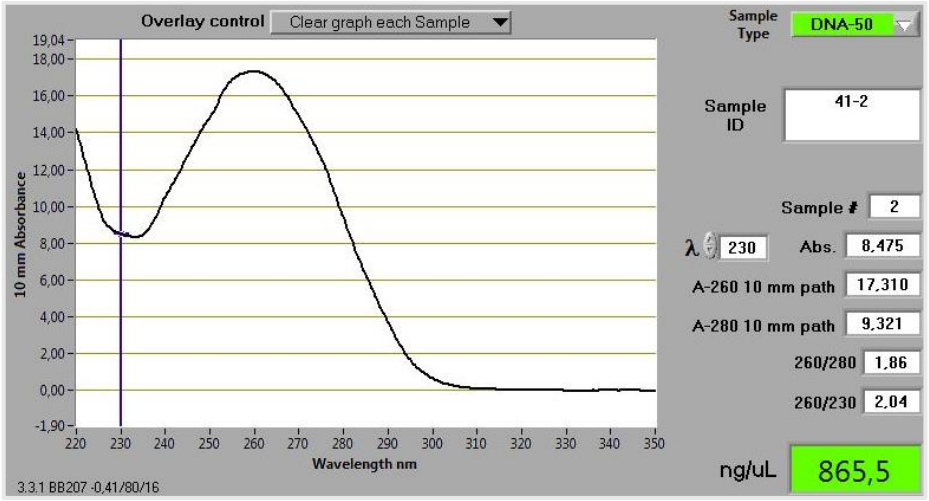
A descrição detalhada dos elementos presentes nas equações desenvolvidas, bem como a demonstração de sua aplicação em exemplos teóricos, encontra-se apresentada no Apêndice E.

5 RESULTADOS

5.1 Qualidade das amostras de DNA

As análises de qualidade das 190 amostras de DNA extraídas permitiram avaliar tanto a eficiência dos métodos utilizados quanto à adequação do material para aplicações posteriores em genotipagem. Das 190 amostras, as 84 amostras extraídas pelo método clássico de fenol-clorofórmio, apresentaram uma média de concentração de 659,41 ng/µL, razão de absorbância 260/280 média de 1,88 e 260/230 média de 1,90. As 106 amostras obtidas utilizando o *QIAamp DNA blood Mini Kit* (*Qiagen*) apresentaram média de concentração de 240,86 ng/µL, razão 260/280 de 1,85 e 260/230 de 1,76. Considerando o conjunto total de 190 amostras, a média geral foi de 336,00 ng/µL para a concentração, 1,86 para 260/280 e 1,82 para 260/230. Na Figura 3 está exemplificado a visualização dos resultados obtidos por meio da quantificação das amostras de DNA.

Figura 3. Espectro de absorbância no ultravioleta (UV) obtidos na quantificação de cada amostra de DNA, destacando as razões de contaminação com proteínas (260/280) e reagentes (260/230) e a concentração da amostra em ng/µL.



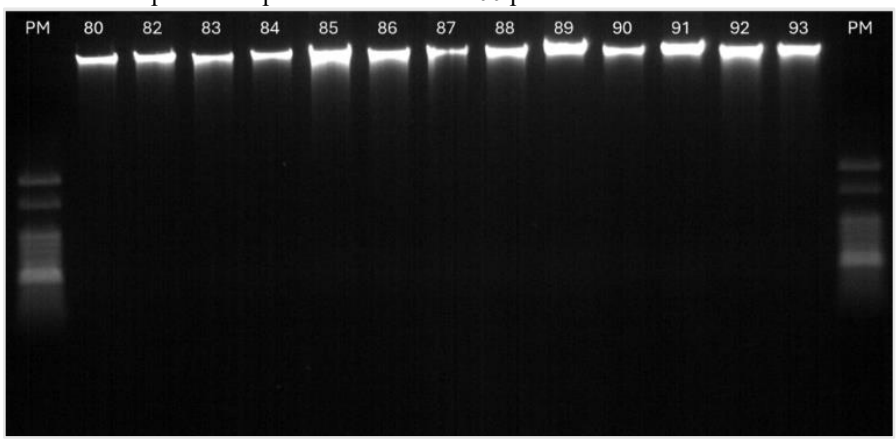
Fonte: Autor.

Esses valores indicam que, em média, as amostras apresentaram boa qualidade de DNA, adequada para aplicações das análises subsequentes. A razão 260/280 próxima de 1,8 sugere baixo nível de contaminação proteica, enquanto a razão 260/230, embora ligeiramente abaixo do ideal em algumas amostras do KIT *QIAamp DNA blood Mini Kit*, permaneceu em níveis aceitáveis, indicando mínima presença de contaminantes orgânicos ou salinos. Observa-se que a extração realizada pelo método fenol-clorofórmio resultou em amostras com concentrações de DNA significativamente maiores, porém com razões 260/230 e 260/280 semelhantes quando

comparadas às obtidas pelo kit *QIAamp DNA blood Mini Kit*, refletindo a maior eficiência do método *Mini Kit* em relação a obtenção de DNA de melhor qualidade mesmo que em concentrações menores que o método fenol-clorofórmio. Os dados de quantificação e razões de contaminação para cada uma das 190 amostras estão exibidos no Apêndice C.

A avaliação da integridade do DNA genômico total das 190 amostras extraídas, realizada por eletroforese em gel de agarose a 1%, mostrou que todas as amostras analisadas apresentaram alta integridade evidenciada pela formação e uma única banda definida no início do gel (Figura 4).

Figura 4. Perfil eletroforético de amostras de DNA extraídas de sangue em gel de agarose 1,0% (TBE 1X) corado com brometo de etídio. PM: padrão de peso molecular de 100 pb.



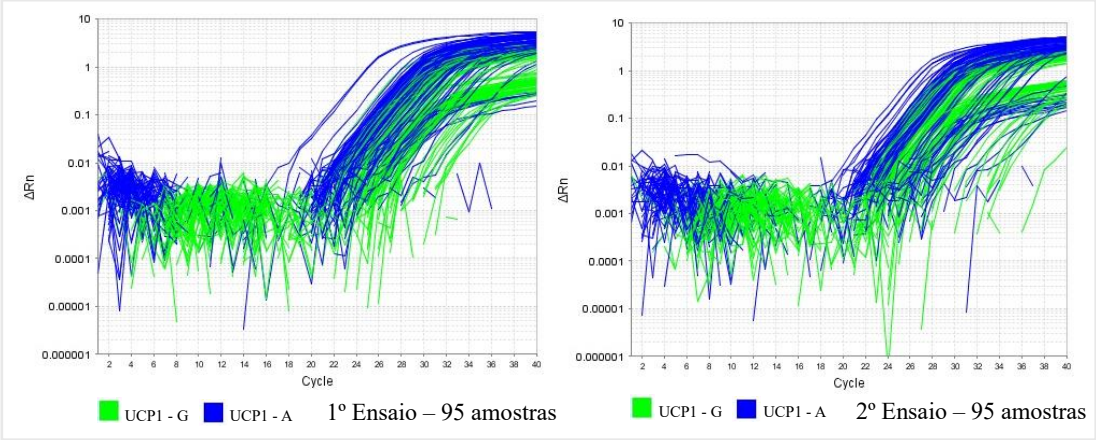
Fonte: Autor.

Esse padrão é característico de DNA de alto peso molecular, indicando que o material extraído manteve sua integridade estrutural e não apresentou sinais de fragmentação ou degradação significativa. A ausência de produto fragmentado ao longo do gel (*smeares*) ou múltiplas bandas reforça que os métodos de extração utilizados, tanto o clássico de fenol-clorofórmio quanto o kit comercial da *Qiagen*, foram eficazes na obtenção de DNA genômico de alta qualidade, adequado para aplicações subsequentes em análises moleculares, como PCR e genotipagem de SNPs. Todas as amostras extraídas apresentaram parâmetros compatíveis com requisitos de integridade e pureza para análises moleculares subsequentes, garantindo a confiabilidade dos experimentos posteriores.

5.2 Genotipagem

As curvas de amplificação obtidas nos ensaios realizados em duas etapas foram avaliadas quanto ao comportamento da fluorescência ao longo dos ciclos de reação, contemplando tanto as amostras alvo quanto os NTCs. De modo geral, as amostras positivas apresentaram aumento progressivo da intensidade de fluorescência após os ciclos iniciais, refletindo a amplificação do material genético e a detecção dos alelos pelas sondas *TaqMan*. Em contrapartida, os NTCs não apresentaram sinais de amplificação detectáveis, mantendo-se próximos à linha de base durante todos os ciclos analisados. A ausência de amplificação nos controles negativos foi considerada um critério essencial para a validação dos ensaios, indicando ausência de contaminação e adequada especificidade da reação (Figura 5).

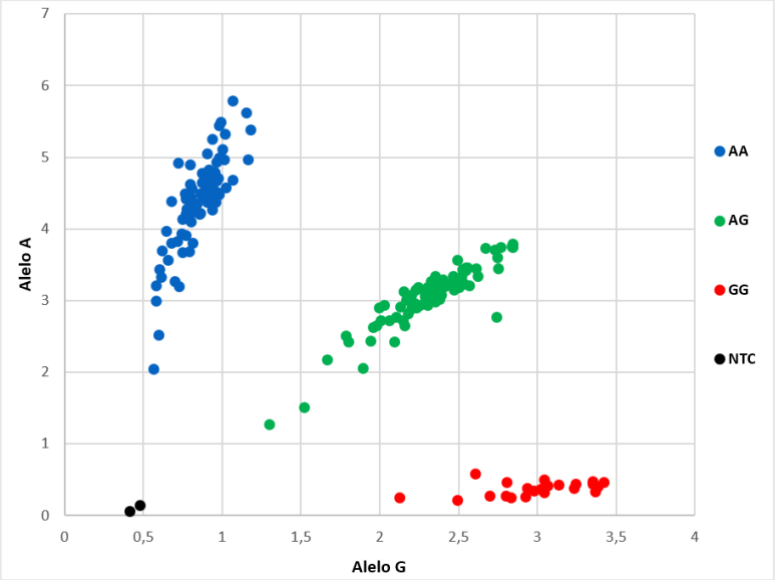
Figura 5. Curvas de amplificação obtidas por *qPCR SNP genotyping* com sondas *TaqMan* para o polimorfismo −3826A/G do gene *UCP1*. Os gráficos apresentam o número de ciclos em função da intensidade de fluorescência (ΔRn), correspondentes aos dois ensaios realizados.



Fonte: Autor.

A genotipagem do SNP –3826A/G do gene *UCP1*, realizada por meio da *qPCR SNP genotyping* com sondas alelo-específicas marcadas com fluoróforos distintos (FAM e VIC), permitiu a identificação clara dos três genótipos possíveis (Figura 6). A determinação dos genótipos para o SNP analisado foi realizada pela análise dos valores de ΔRn após o último ciclo de amplificação para cada sonda, bem como os valores da referência passiva (ROX) e o índice de qualidade para cada reação (%).

Figura 6. Gráfico de dispersão representando a distribuição dos genótipos obtidos na reação de genotipagem por qPCR para o SNP -3826A/G (rs1800592) do gene *UCP1*. Cada ponto representando uma amostra individual. Os sinais de fluorescência relativos aos alelos foram detectados pelos fluoróforos VIC (alelo G) e FAM (alelo A). Grupo homozigoto para o alelo G (região inferior direita – em vermelho), heterozigoto (região central – em verde) e homozigoto para o alelo A (região superior esquerda – em azul. O controle negativo (NCT) é representado pela cor preta.



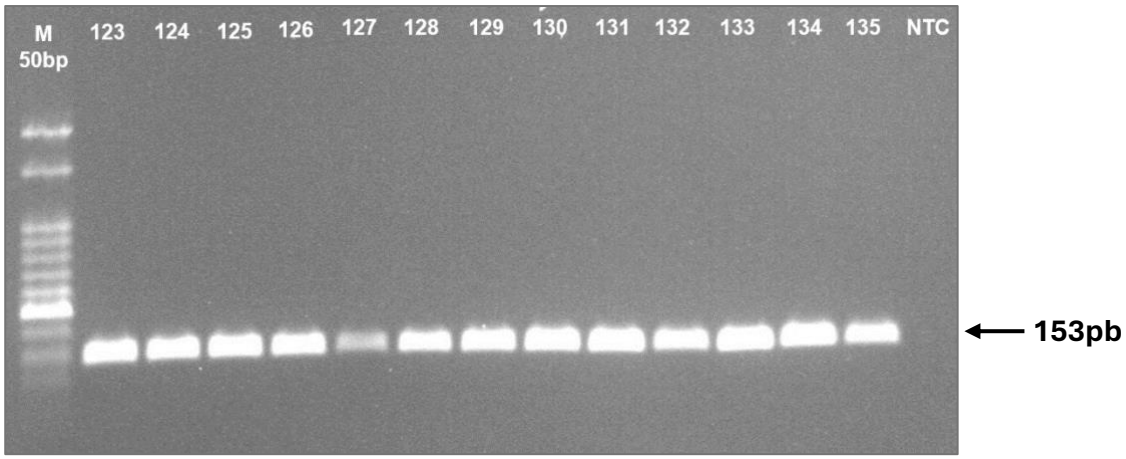
Fonte: Autor.

A distribuição dos sinais de fluorescência gerou três agrupamentos bem definidos no gráfico de dispersão: 88 amostras homozigotas para o alelo A (AA), 80 heterozigotas (AG) e 22 homozigotas para o alelo G (GG), totalizando 190 amostras. Com base nesses resultados, as frequências alélicas calculadas foram p = 0,674 para o alelo A e q = 0,326 para o alelo G. A partir dessas proporções, as frequências genotípicas observadas foram 0,463 para AA, 0,421 para AG e 0,116 para GG. Pelo princípio do equilíbrio de Hardy-Weinberg (HWE), as frequências genotípicas esperadas seriam p^2 = 0,454 para AA, $2pq$ = 0,440 para AG e q^2 = 0,106 para GG. A comparação entre os valores observados e esperados revelou uma concordância próxima (χ^2 = 0,340; p > 0,05), indicando que a população amostral se encontra em equilíbrio de HWE para este *locus*, sem desvios significativos que sugiram seleção, mutação, migração ou erro de genotipagem. Ademais, os resultados obtidos são consistentes com os dados de frequência alélica do SNP –3826A/G (rs1800592), reportados no banco de dados *dbSNP* (NCBI) no qual o alelo G apresenta frequência menor em diferentes populações e q = 0,421 na classificação *Latin America 2*, que inclui a população brasileira, reforçando a validade dos resultados experimentais e a representatividade da amostra analisada.

5.3 Espectros obtidos dos *amplicons*

Após a etapa de otimização das condições de amplificação, observou-se que a combinação final de reagentes e parâmetros termocíclicos resultou em *amplicons* altamente específicos e de elevado rendimento (Figura 7).

Figura 7. Eletroforese em gel de agarose a 1,6% dos produtos amplificados por PCR qualitativa da região contendo o SNP –3826A/G do gene *UCP1*. Visualizam-se bandas únicas, nítidas e bem definidas correspondentes aos *amplicons* de 153 pares de bases (pb), compatíveis com o fragmento esperado, indicando alta especificidade e eficiência da reação de amplificação após a etapa de otimização dos *primers*. O marcador de peso molecular (M) utilizado foi de 50 pb, permitindo confirmar o tamanho do produto amplificado. A ausência de bandas inespecíficas ou de amplificação no controle negativo (NTC) confirma a ausência de contaminações e a fidelidade da reação.



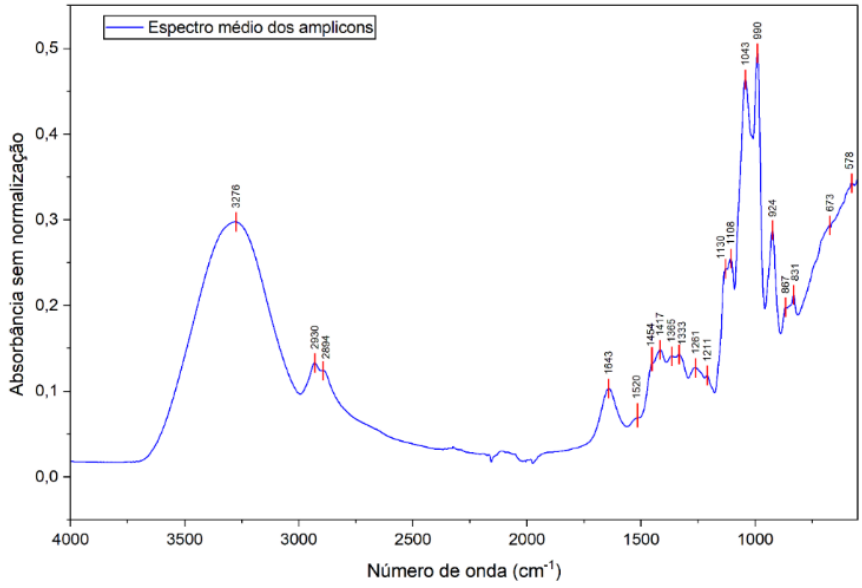
Fonte: Autor.

A visualização dos produtos amplificados em gel de agarose a 1,6% revelou bandas únicas, nítidas e bem definidas, com tamanho compatível ao fragmento esperado, sem a presença de bandas inespecíficas ou resíduos de *primers* ligantes entre si (*primers-dimer*). Esse resultado evidencia que o desenho dos oligonucleotídeos e as condições de reação empregadas foram eficientes para a amplificação seletiva do fragmento genômico contendo o SNP em estudo. Além disso, a uniformidade e intensidade das bandas observadas indicam boa eficiência na amplificação entre as amostras, refletindo a qualidade adequada do DNA e a reprodutibilidade da técnica. A ausência de artefatos visuais e de produtos inespecíficos no gel reforça a confiabilidade da metodologia estabelecida, permitindo a utilização segura dos *amplicons* obtidos nas análises espectroscópicas com base nos genótipos previamente identificados.

Os espectros obtidos dos *amplicons* amplificados por PCR, revelaram um perfil espectral característico de biomoléculas, com bandas bem distribuídas ao longo do infravermelho médio (4000 a 550 cm^{-1}). Na análise inicial, considerando o espectro médio de

todas as 190 amostras sem distinção de genótipos e sem normalizações, observou-se um conjunto de bandas atribuídas principalmente às vibrações típicas de biomoléculas (Figura 8). Para a região de alta frequência, destacam-se as bandas 3276, 2930 e 2894 cm^{-1}. Além destas, a maior informação espectral se concentra nas bandas 1643, 1520, 1417, 1365, 1333, 1261, 1211, 1130, 1108, 1043, 990, 924, 867 e 831 cm^{-1}, que representam a região de *fingerprint* da amostra (Tabela 3). Adicionalmente, observa-se a presença das bandas 673 e 578 cm^{-1}, mais atribuídas a constante sobreposições de vibrações. Embora o espectro observado seja característico de biomoléculas, não se pode descartar o fato de que ainda há presença residual de componentes do meio reacional, como o tampão Tris-KCl e a Taq DNA polimerase, que também contribuem parcialmente para o perfil obtido.

Figura 8. Espectro médio de absorbância no infravermelho na faixa espectral de 4000 a 550 cm^{-1} obtido por espectroscopia ATR-FTIR dos *amplicons* de PCR correspondentes à região do SNP −3826A/G do gene *UCP1*.



Fonte: Autor.

Tabela 3. Atribuição vibracional das bandas da região de *fingerprint* observadas nos espectros ATR-FTIR dos *amplicons* de PCR, com base em referências da literatura.

Bandas Observadas (cm^{-1})	Bandas de Referência (cm^{-1})	Referências na Literatura	Ligação e Modo Vibracional	Grupos Funcionais
1643	1637 / 1632 / 1655	Souza et al., 2024 / Banyay, Sarkar e Gräslund, 2003 / Brewer et al., 2002	ν(C = C) + ν(C = O) + ν(C = N)	Bases nitrogenadas
1520	1520	Banyay, Sarkar e Gräslund, 2003	ν(C = C)	Bases nitrogenadas

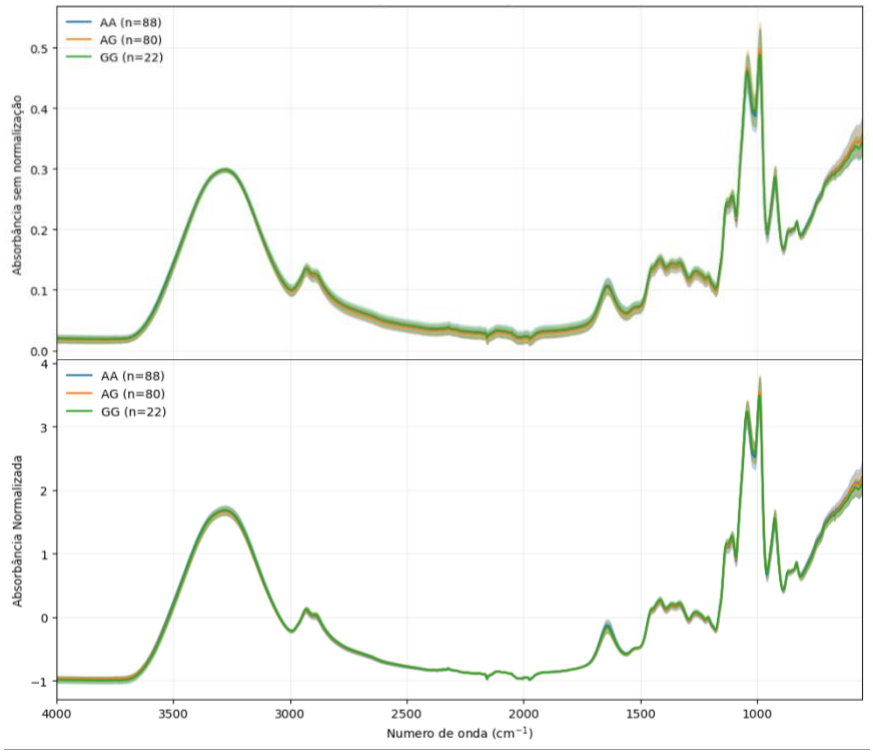
1454	1455 / 1453	Souza et al., 2024 / Banyay, Sarkar e Gräslund, 2003	δas(CH_3) + ν(C = N)	Bases nitrogenadas
1417	1412 / 1418	Souza et al., 2024 / Banyay, Sarkar e Gräslund, 2003	ν(C-N) + δ(N-H) + δ(C-H) + (C = N)	Bases nitrogenadas + desoxirribose
1365	1365	Banyay, Sarkar e Gräslund, 2003	δ(CH_2) + ν(C-N)	
1333	1326 / 1335	Souza et al., 2024 / Banyay, Sarkar e Gräslund, 2003	ν(C-N) + ν(C = N) + δ(C-H)	Bases nitrogenadas + desoxirribose
1261	1230 / 1244 / 1245 / 1238	Souza et al., 2024 / Mello e Vidal, 2012 / Banyay, Sarkar e Gräslund, 2003 / Brewer et al., 2002	νas(PO_2^-)	Grupos fosfato
1211	1207 / 1216	Souza et al., 2024 / Banyay, Sarkar e Gräslund, 2003	νas(PO_2^-)	Grupos fosfato
1130	1135	Banyay, Sarkar e Gräslund, 2003	νs(P-O-C) + ν(C-O) + ν(C-C)	Ligação fosfodiéster + desoxirribose
1108	1106 / 1116	Souza et al., 2024 / Banyay, Sarkar e Gräslund, 2003	νs(PO_2^-) + νs(P-O-C) + ν(C-O) + ν(C-C)	Ligação fosfodiéster + desoxirribose
1043	1031 / 1044	Souza et al., 2024 / Banyay, Sarkar e Gräslund, 2003	ν(C-C) + ν(C-O) + δ(C-O) + ν(CH_2OH)	Ligação fosfodiéster + desoxirribose
990	991 / 995	Souza et al., 2024 / Banyay, Sarkar e Gräslund, 2003	ν(C-O) + νs(P-O-C)	Ligação fosfodiéster + desoxirribose
924	923 / 924	Souza et al., 2024 / Banyay, Sarkar e Gräslund, 2003	νs(C-O-C)	Desoxirribose
867	865	Banyay, Sarkar e Gräslund, 2003	ν(C-O) + ν(C-C)	Desoxirribose
831	820	Banyay, Sarkar e Gräslund, 2003	ν(C-O) + ν(C-C)	Desoxirribose

ν = estiramento | **νs** = estiramento simétrico | **νas** = estiramento assimétrico | **δ** = deformação | **δas** = deformação assimétrica
Fonte: Autor.

A Figura 9 apresenta os espectros médios de ATR-FTIR das amostras agrupadas de acordo com os genótipos AA, AG e GG, acompanhados de seus respectivos desvios padrão, tanto na forma não normalizada quanto após a aplicação da normalização. De modo geral, observa-se elevada similaridade entre os espectros médios dos três grupos genotípicos ao longo de toda a região espectral analisada, independentemente do tratamento aplicado aos dados. As principais bandas espectrais estão presentes em posições semelhantes para todos os grupos, com

sobreposição expressiva dos perfis médios, indicando um comportamento espectral globalmente consistente entre os genótipos.

Figura 9. Espectros médios de ATR-FTIR das amostras de DNA amplificado por PCR, agrupadas de acordo com os genótipos AA (n = 88), AG (n = 80) e GG (n = 22), acompanhados do desvio padrão. O painel superior apresenta os espectros médios sem normalização, enquanto o painel inferior mostra os espectros após a aplicação da normalização por variância normal padrão (SNV).



Fonte: Autor.

Na representação sem normalização, os espectros médios apresentam intensidades absolutas comparáveis entre os grupos, com desvios padrão observável na maior parte do espectro. As variações observadas concentram-se principalmente nas regiões de maior intensidade espectral, refletindo flutuações experimentais associadas à aquisição dos dados e à variabilidade intrínseca das amostras. Ainda assim, o padrão geral de absorção mantém-se altamente semelhante entre os grupos AA, AG e GG.

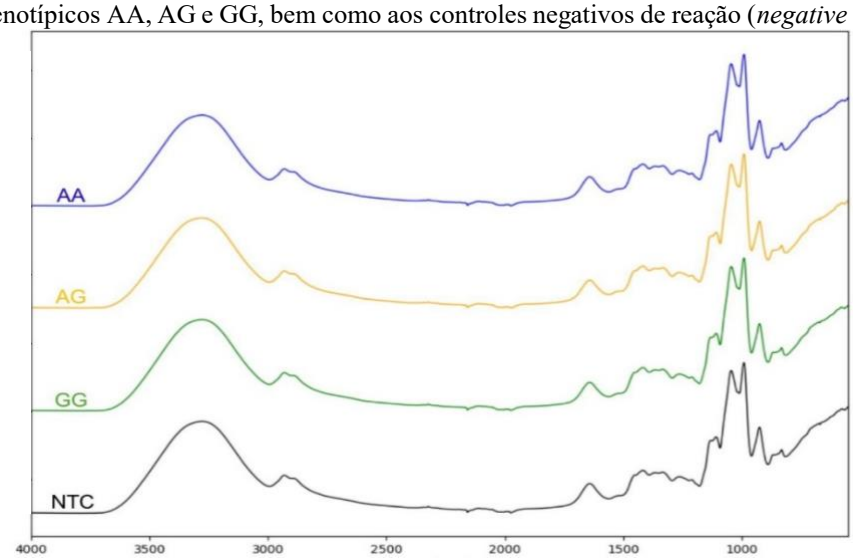
Após a normalização por SNV, os espectros médios preservam a forma global observada nos dados não normalizados, evidenciando que o procedimento de normalização não altera substancialmente o perfil espectral médio dos grupos. No entanto, após a aplicação do SNV, pode-se observar com maior clareza a presença de um maior desvio padrão em determinadas regiões do espectro, especialmente na porção de menores números de onda, mais próxima de 550 cm^{-1}. Nessa região, a normalização tende a evidenciar diferenças relativas entre as amostras, resultando em maior dispersão em torno do espectro médio.

De forma geral, a comparação entre os espectros normalizados e não normalizados indica que, embora a normalização realce a variabilidade relativa em regiões específicas do espectro, a similaridade global entre os perfis médios dos diferentes genótipos é mantida.

5.4 Espectros diferenciais em relação ao NTC

Os espectros médios de ATR-FTIR dos grupos genotípicos AA, AG e GG apresentaram elevada similaridade com o espectro médio dos NTCs ao longo de toda a região espectral analisada. As principais bandas encontram-se em posições coincidentes entre os *amplicons* e o NTCs, com padrões de absorção amplamente sobrepostos e diferenças visuais nulas entre os perfis espectrais. Esse comportamento é observado tanto nas regiões de maior intensidade espectral quanto na região de impressão digital, indicando que o perfil espectral global é fortemente semelhante entre todos os grupos avaliados (Figura 10).

Figura 10. Espectros médios normalizados de ATR-FTIR dos *amplicons* de PCR correspondentes aos diferentes grupos genotípicos AA, AG e GG, bem como aos controles negativos de reação (*negative template control*, NTC).



Fonte: Autor.

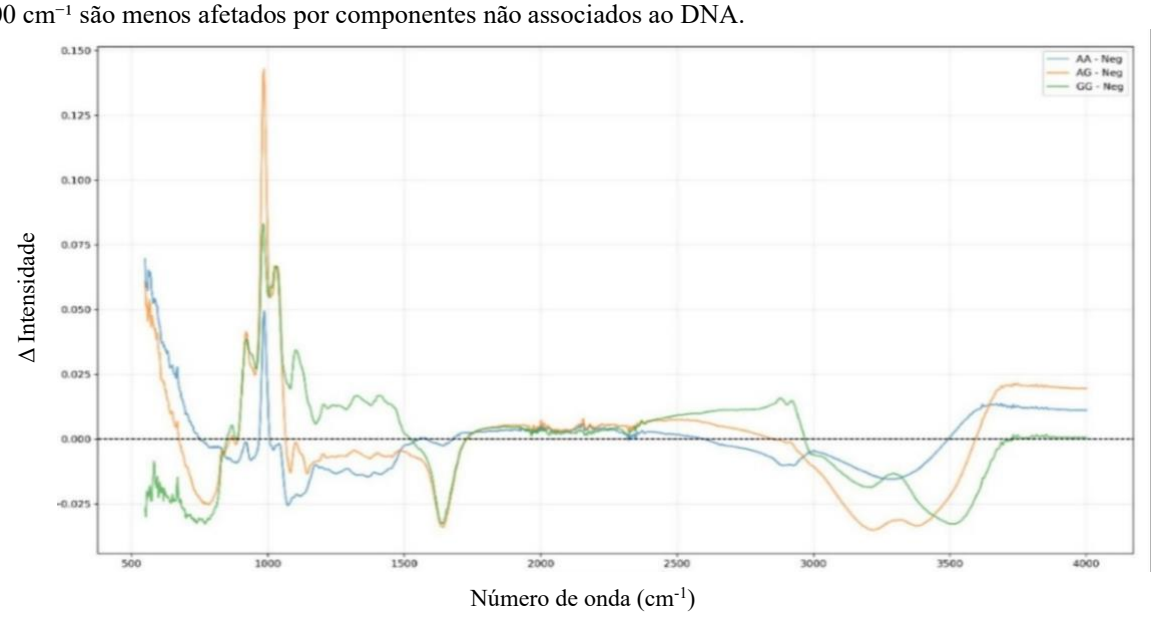
Essa similaridade é particularmente evidente nas regiões de maior intensidade espectral, incluindo as faixas associadas às vibrações do estiramento O–H/N–H na região de números de onda mais elevados, bem como na região de impressão digital, onde múltiplas bandas estreitas e intensas são observadas tanto nos espectros dos *amplicons* quanto no NTC. O comportamento espectral global sugere que uma parcela significativa das contribuições observadas nos

espectros dos *amplicons* está relacionada a componentes comuns ao sistema reacional, presentes também nos controles negativos.

Apesar da presença do DNA amplificado nos grupos genotípicos, as diferenças visuais entre os espectros médios dos *amplicons* e do NTC são sutis e não facilmente distinguíveis por inspeção direta. O padrão espectral dominante permanece amplamente conservado entre todos os grupos analisados, indicando que, antes de qualquer tratamento adicional dos dados, a contribuição espectral do meio reacional exerce influência relevante sobre o perfil observado.

A subtração do espectro médio de ATR-FTIR dos NTCs em relação aos espectros médios de cada grupo genotípico evidenciou regiões espectrais menos influenciadas por componentes não associados ao DNA amplificado. Os espectros diferenciais resultantes para os grupos AA, AG e GG são apresentados na Figura 11, nos quais desvios positivos e negativos em relação à linha de base indicam diferenças em relação ao fundo espectral representado pelos NTCs.

Figura 11. Espectros diferenciais obtidos pela subtração do espectro médio de ATR-FTIR dos controles negativos de reação (*negative template control*, NTC) em relação aos espectros médios dos *amplicons* de PCR para cada grupo genotípico (AA–Neg, AG–Neg e GG–Neg). A subtração evidencia regiões espectrais com menor contribuição de reagentes residuais da PCR, indicando que os intervalos de 900–1100 cm^{-1}, 950-1200 cm^{-1} e 2800–3800 cm^{-1} são menos afetados por componentes não associados ao DNA.



Fonte: Autor.

De modo geral, observa-se que grande parte do espectro apresenta variações de baixa intensidade, com flutuações próximas de zero ao longo da maioria da faixa de números de onda analisada. Esse comportamento reflete a elevada similaridade entre as assinaturas espectrais dos *amplicons* de PCR e dos NTCs, indicando sobreposição substancial das contribuições espectrais

associadas ao sistema reacional. No entanto, três regiões espectrais se destacam por apresentarem desvios mais consistentes e menor influência do sinal dos NTCs.

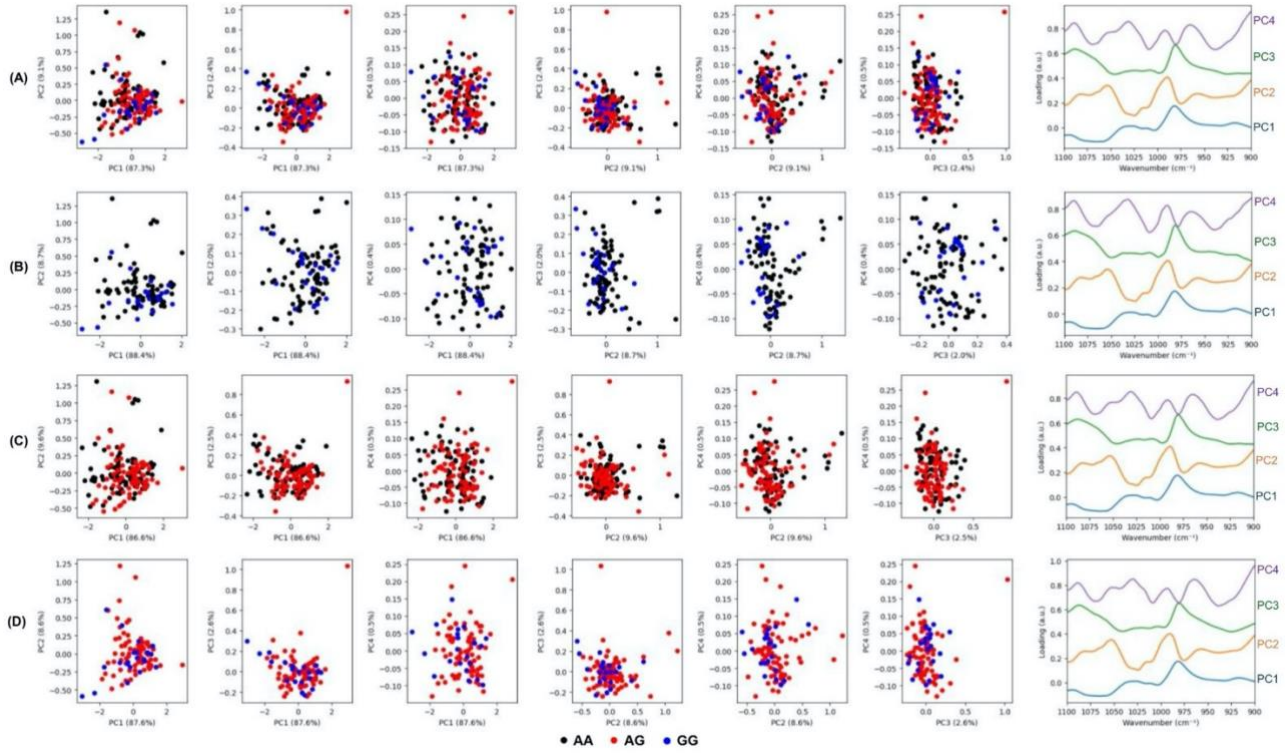
As duas primeiras correspondem ao intervalo entre 900 e 1100 cm^{-1} e 950 a 1200 cm^{-1}, respectivamente, no qual os espectros diferenciais exibem características mais pronunciadas e sistemáticas entre os grupos genotípicos. A terceira região localiza-se na região de alta frequência, abrangendo aproximadamente o intervalo entre 2800 e 3800 cm^{-1}. Em ambas as regiões, a menor contribuição do fundo espectral sugere redução da interferência de reagentes residuais da PCR.

As regiões espectrais de 900–1100 e 950–1200 cm^{-1} são comumente associadas a modos vibracionais relacionados ao arcabouço fosfato e às porções sacarídicas dos ácidos nucleicos, enquanto a região de 2800–3800 cm^{-1} abrange vibrações de estiramento coletivas, incluindo modos O–H e C–H, que refletem características estruturais mais amplas e aspectos relacionados à hidratação do DNA. A menor influência do sinal dos NTCs nessas regiões indica que elas capturam predominantemente informações vibracionais intrínsecas associadas ao DNA amplificado, com contribuição reduzida de componentes do meio reacional. Com base nessas observações, os três intervalos espectrais foram selecionados para as análises multivariadas subsequentes voltadas à avaliação de diferenças entre genótipos.

5.5 Machine learning

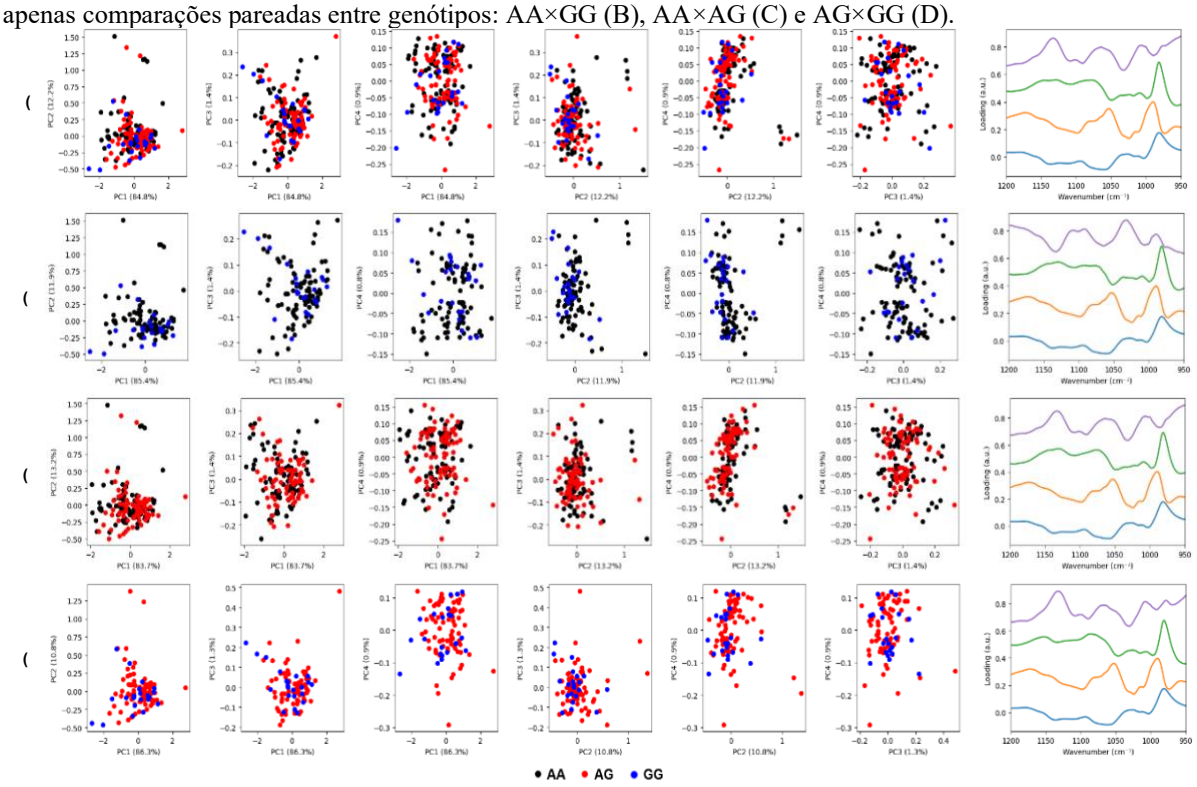
A análise não supervisionada de PCA aplicada aos espectros de ATR-FTIR normalizados por SNV, nas regiões de 900–1100 cm^{-1} (Figura 12), 950–1200 cm^{-1} (Figura 13) e 2800–3800 cm^{-1} (Figura 14), revelou ampla sobreposição entre as amostras dos grupos genotípicos AA, AG e GG. Em todas as regiões analisadas, os gráficos de escores (*score plot*) não evidenciaram separação visual consistente entre os genótipos, tanto nas análises conjuntas quanto nas comparações pareadas (AA×GG, AA×AG e AG×GG).

Figura 12. Análise de componentes principais (*principal component analysis*, PCA) aplicada aos espectros de ATR-FTIR normalizados por SNV na região de 900–1100 cm^{-1}, selecionada com base na análise dos espectros diferenciais como um intervalo espectral com menor interferência de reagentes. (A) Gráficos de escores da PCA considerando conjuntamente todos os grupos genotípicos, com combinações pareadas entre os quatro primeiros componentes principais (PC1×PC2, PC1×PC3, PC1×PC4, PC2×PC3, PC2×PC4 e PC3×PC4), nos quais as amostras AA, AG e GG são representadas, respectivamente, por pontos pretos, vermelhos e azuis; à direita são apresentados os perfis de *loadings* correspondentes PC1–PC4 ao longo da faixa de números de onda analisada. (B–D) Gráficos de escores da PCA construídos utilizando as mesmas combinações de componentes, considerando apenas comparações pareadas entre genótipos: AA×GG (B), AA×AG (C) e AG×GG (D).



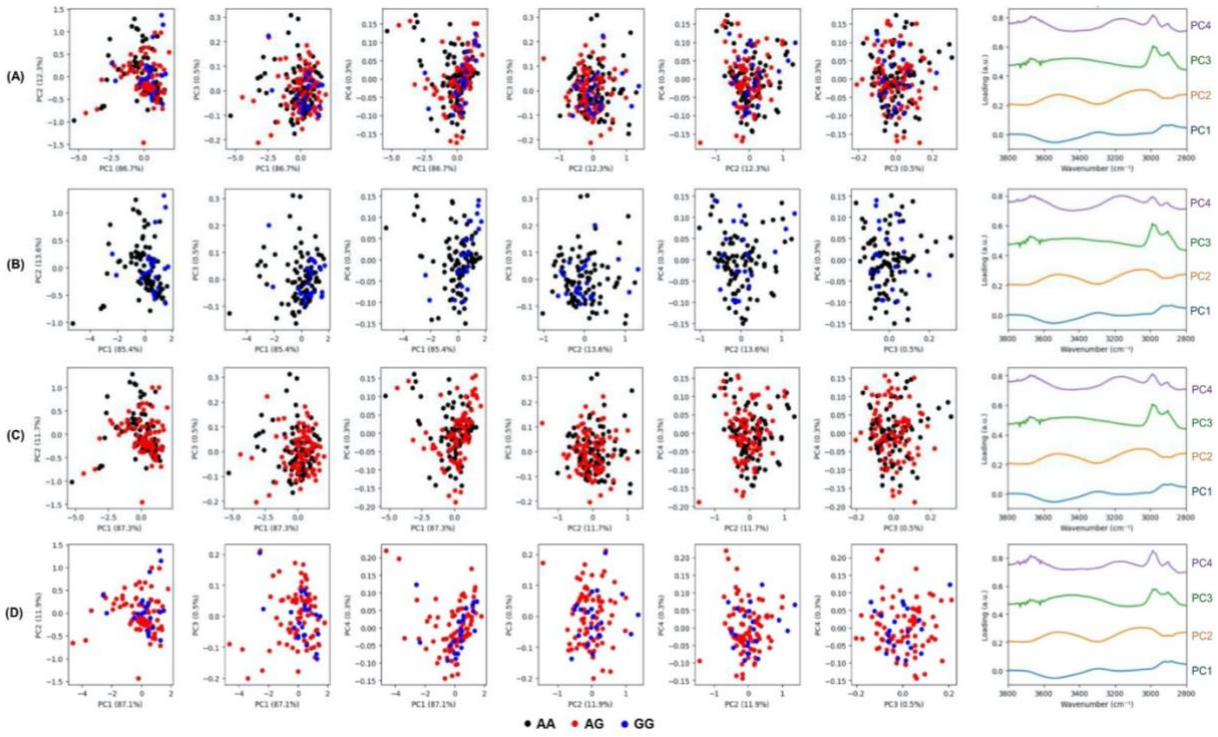
Fonte: Autor.

Figura 13. Análise de componentes principais (*principal component analysis*, PCA) aplicada aos espectros de ATR-FTIR normalizados por SNV na região de 950–1200 cm^{-1}, selecionada com base na análise dos espectros diferenciais como um intervalo espectral com menor interferência de reagentes. (A) Gráficos de escores da PCA considerando conjuntamente todos os grupos genotípicos, com combinações pareadas entre os quatro primeiros componentes principais (PC1×PC2, PC1×PC3, PC1×PC4, PC2×PC3, PC2×PC4 e PC3×PC4), nos quais as amostras AA, AG e GG são representadas, respectivamente, por pontos pretos, vermelhos e azuis; à direita são apresentados os perfis de *loadings* correspondentes PC1–PC4 ao longo da faixa de números de onda analisada. (B–D) Gráficos de escores da PCA construídos utilizando as mesmas combinações de componentes, considerando apenas comparações pareadas entre genótipos: AA×GG (B), AA×AG (C) e AG×GG (D).



Fonte: Autor.

Figura 14. Análise de componentes principais (*principal component analysis*, PCA) aplicada aos espectros de ATR-FTIR normalizados por SNV na região de 2800–3800 cm^{-1}, selecionada com base na análise dos espectros diferenciais como um intervalo espectral com menor interferência de reagentes. (A) Gráficos de escores da PCA considerando conjuntamente todos os grupos genotípicos, com combinações pareadas entre os quatro primeiros componentes principais (PC1×PC2, PC1×PC3, PC1×PC4, PC2×PC3, PC2×PC4 e PC3×PC4), nos quais as amostras AA, AG e GG são representadas, respectivamente, por pontos pretos, vermelhos e azuis; à direita são apresentados os perfis de *loadings* correspondentes PC1–PC4 ao longo da faixa de números de onda analisada. (B–D) Gráficos de escores da PCA construídos utilizando as mesmas combinações de componentes, considerando apenas comparações pareadas entre genótipos: AA×GG (B), AA×AG (C) e AG×GG (D).



Fonte: Autor.

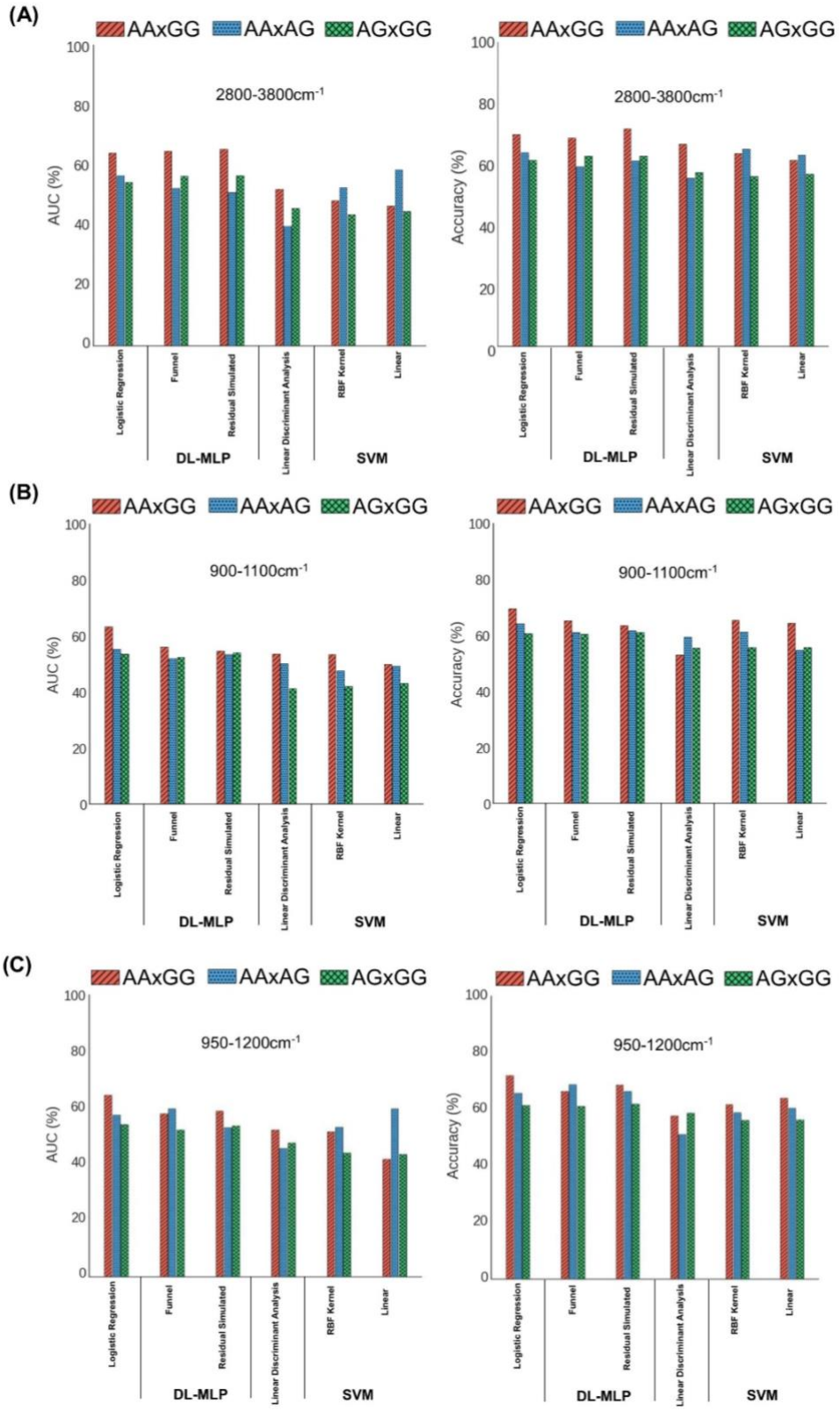
Nas três regiões espectrais avaliadas, os dois primeiros componentes principais (PC1 e PC2) explicaram aproximadamente 98–99% da variância total, tanto nas análises envolvendo todos os genótipos quanto nas comparações pareadas. De acordo com o critério de Kaiser, apenas PC1 e PC2 apresentaram autovalores superiores a 1, sendo, portanto, considerados os componentes mais relevantes para a descrição da variabilidade dos dados. Apesar da elevada variância capturada por esses componentes, as projeções PC1×PC2, assim como aquelas envolvendo combinações com PC3 e PC4, não resultaram em agrupamentos distintos entre os grupos genotípicos. Os *loading plots* indicam que a variabilidade espectral está distribuída ao longo dos respectivos intervalos de números de onda, sem predomínio de contribuições localizadas associadas à discriminação visual entre genótipos.

Em conjunto, esses resultados indicam que a análise não supervisionada de PCA não foi suficiente para discriminar os grupos genotípicos a partir dos espectros avaliados. Esse comportamento sustenta a aplicação de métodos de ML supervisionados nas etapas

subsequentes, com o objetivo de explorar padrões multivariados mais sutis presentes nos dados espectrais do DNA amplificado que contêm o SNP -3826 A/G do gene *UCP1*.

Subsequentemente, modelos de ML e DL foram aplicados aos espectros de ATR-FTIR normalizados por SNV dos *amplicons* do SNP -3826 A/G do gene *UCP1*com o objetivo de avaliar a capacidade de discriminação entre os grupos genotípicos (AA×GG, AA×AG e AG×GG). O desempenho dos modelos foi avaliado nos intervalos espectrais previamente selecionados (2800–3800 cm^{-1}, 900–1100 cm^{-1} e 950–1200 cm^{-1}) e utilizando a AUC e a acurácia como métricas principais. Os resultados comparativos estão resumidos na Figura 15, enquanto o conjunto completo de medidas de desempenho está representado no Apêndice F.

Figura 15. Desempenho dos modelos de aprendizado de máquina na discriminação entre pares de genótipos com base nos espectros de ATR-FTIR dos *amplicons* de PCR, considerando as comparações AA×GG, AA×AG e AG×GG. Os valores de AUC são apresentados nos painéis à esquerda e os de acurácia nos painéis à direita, para todos os intervalos espectrais analisados. (A) Resultados obtidos para a região de 2800–3800 cm^{-1}. (B) Resultados obtidos para a região de 900–1100 cm^{-1}. (C) Resultados obtidos para a região de 950–1200 cm^{-1}.



Fonte: Autor.

De forma geral, os modelos supervisionados apresentaram desempenho de classificação moderado, com variação dependente tanto do intervalo espectral analisado quanto da arquitetura do modelo empregado. Em contraste com as análises não supervisionadas de PCA, que não evidenciaram separação clara entre os genótipos, os modelos supervisionados foram capazes de explorar padrões espectrais multivariados sutis, resultando em desempenho discriminatório consistente entre pares genotípicos específicos.

Na região de 2800–3800 cm^{-1} (Figura 15-A), foram observados os melhores desempenhos globais, especialmente para comparações envolvendo o genótipo AA. Nesse intervalo espectral, a arquitetura de DL-MLP Residual Simulated apresentou o melhor desempenho entre os modelos avaliados, alcançando uma AUC média de aproximadamente 65,4% e acurácia de 71,6% na comparação AA×GG. Para a discriminação AA×AG, no entanto, o mesmo modelo obteve AUC média de 56,5% e acurácia de 62,7%, indicando maior adequação dessa arquitetura para capturar padrões espectrais não lineares distribuídos ao longo desse intervalo de números de onda mais amplo.

Em contraste, os resultados obtidos na região de 900–1100 cm^{-1} (Figura 15-B), associada predominantemente a modos vibracionais do arcabouço fosfodiéster do DNA, indicaram uma preferência por modelos mais simples. Nesse intervalo, a regressão logística apresentou o melhor desempenho entre os modelos avaliados, alcançando uma AUC média de 63,4% e acurácia de 69,6% para a melhor comparação genotípica, que foi AA×GG. Embora os valores de desempenho nessa região tenham sido ligeiramente inferiores aos observados no intervalo de maior número de onda, eles representam os melhores resultados médios obtidos dentro de uma região espectral dominada por contribuições estruturais do DNA.

A Figura 15-C apresenta os resultados de classificação para a região de 950–1200 cm^{-1}, selecionada com base em análises preliminares e em evidências da literatura quanto à relevância de números de onda específicos para a discriminação genotípica. Nesse intervalo, a regressão logística novamente apresentou o melhor desempenho médio, com AUC de 64,4% e acurácia de 72% para a comparação AA×GG. As métricas obtidas nessa região foram semelhantes às observadas no intervalo de 900–1100 cm^{-1}, mantendo um padrão de desempenho moderado e consistente.

Os números de onda identificados como mais relevantes para cada intervalo espectral, bem como as respectivas comparações genotípicas, modelos com melhor desempenho e valores médios de AUC, estão sumarizados no Apêndice G. Em conjunto, esses resultados demonstram que para a amostras estudada, o desempenho dos modelos de ML e DL depende fortemente tanto da seleção do intervalo espectral quanto da arquitetura do modelo. Embora as análises não

supervisionadas não tenham promovido separação genotípica evidente, as abordagens supervisionadas permitiram a extração de informações discriminatórias a partir de variações espectrais sutis, possibilitando a diferenciação moderada entre pares de genótipos em alguns modelos aplicados.

5.6 Viabilidade da espectroscopia ATR-FTIR associado ao ML

A avaliação comparativa dos parâmetros de tempo e custo evidenciou diferenças substanciais entre as metodologias analisadas, permitindo evidenciar a viabilidade operacional da abordagem baseada em ATR-FTIR associada ao ML discriminação das variantes do SNP −3826A/G do gene *UCP1*. A Tabela 4 reúne os valores obtidos para o tempo de execução manual (*TEM*), tempo de operação instrumental (*TOI*), tempo total de processamento (*TTP*) e custo estimado por amostra, considerando condições experimentais equivalentes entre as metodologias avaliadas. De forma geral, os resultados indicam que a estratégia proposta apresenta desempenho operacional relevante quando comparada às técnicas convencionalmente empregadas para genotipagem de SNPs.

Tabela 4. Comparação do tempo de execução manual (TEM), tempo de operação instrumentação (TOI), tempo total de processamento (TTP) e custos de reagentes para as metodologias de genotipagem de SNPs analisadas

Método	***TEM* (min./am.) ***	***TOI* (min./am.) ***	***TTP* (min./am.) ***	**Custo (USD, por amostra) ****
Sequenciamento NGS	2,19	12,81	15,00	60,00-100,00
qPCR SNP genotyping	1,88	0,62	2,50	3,00-4,00
ATR-FTIR associada ao ML	2,28	3,52	5,80	0,50-1,00

***** Os resultados das estimativas de tempo foram definidos em minutos por amostra (min./am.)
****** Os custos estimados de reagentes são baseados em preços médios de mercado expressos em dólar americano (USD). Os valores indicam faixas aproximadas que podem variar de acordo com fornecedor, região, infraestrutura laboratorial e condições operacionais. Custos de mão de obra, consumo de energia e depreciação de equipamentos não foram incluídos. TEM, tempo de execução manual; TOI, tempo de operação instrumental; TTP, tempo total de processamento; NGS, sequenciamento de nova geração; *qPCR SNP genotyping,* genotipagem de SNP por reação de cadeia em polimerase quantitativa em tempo real.

Fonte: Autor.

A abordagem baseada em ATR-FTIR associada ao aprendizado de máquina apresentou TTP de 5,80 min por amostra. Esse valor corresponde a aproximadamente 39% do tempo requerido pelo sequenciamento baseado em NGS (15,00 min/amostra) e a cerca de 2,3 vezes o

tempo observado para qPCR (2,50 min/amostra). Embora a qPCR apresente o menor TTP absoluto em função do elevado grau de automação e aquisição paralela de dados, o tempo obtido para o ATR-FTIR indica a possibilidade de realização de análises em intervalos reduzidos, considerando a simplicidade instrumental e o menor número de etapas bioquímicas envolvidas no fluxo experimental.

A análise dos componentes temporais evidencia que a espectroscopia ATR-FTIR associada ao ML apresentou TEM de 2,28 min por amostra, valor próximo ao observado para qPCR (1,88 min/amostra) e ao NGS (2,19 min/amostra), indicando que o esforço operacional requerido permanece semelhante entre as abordagens. O TOI foi de 3,52 min por amostra, superior ao da qPCR (0,62 min/amostra), porém substancialmente inferior ao do NGS (12,81 min/amostra). Essa diferença reflete a natureza sequencial da aquisição espectral em ATR-FTIR, em contraste com a leitura simultânea característica da qPCR e com as longas corridas instrumentais necessárias para o sequenciamento de nova geração.

Em relação aos custos de reagentes, a abordagem baseada em ATR-FTIR e ML apresentou valores estimados entre USD 0,50 e 1,00 por amostra, correspondendo à menor faixa de custo entre as metodologias avaliadas. Esse intervalo representa uma redução aproximada de 70–85% em relação à qPCR (USD 3,00–4,00 por amostra) e superior a 98% quando comparado ao NGS (USD 60,00–100,00 por amostra). A diferença observada está associada principalmente à utilização exclusiva de reagentes convencionais de PCR previamente empregados na amplificação do DNA, sem necessidade de sondas fluorescentes específicas ou reagentes de preparo de biblioteca e sequenciamento, o que ressalta a vantagem da espectroscopia ATR-FTIR como uma ferramenta que dispensa o uso de reagentes especializados.

6 DISCUSSÃO

Este estudo propõe o uso da espectroscopia ATR-FTIR combinada com ML como uma estratégia analítica alternativa e de custo reduzido para a discriminação genotípica de *amplicons* de PCR quanto a presença de variantes de um SNP, motivada pela crescente relevância da análise de SNPs na investigação de distúrbios multifatoriais, os quais representam importantes desafios para a saúde pública global. Dentro dessa proposta, torna-se fundamental compreender inicialmente quais informações moleculares estão efetivamente representadas nos espectros obtidos.

Nesse sentido, a análise de caracterização espectral dos *amplicons* de PCR por ATR-FTIR permitiu a identificação de um conjunto consistente de bandas distribuídas ao longo da região de *fingerprint* do DNA, refletindo contribuições vibracionais associadas tanto às bases nitrogenadas quanto ao esqueleto açúcar-fosfato. De modo geral, as atribuições observadas mostram elevada concordância com descrições previamente reportadas para ácidos nucleicos, reforçando a natureza molecular dos sinais detectados e a predominância de modos vibracionais intrínsecos à estrutura do DNA.

Na região de maior número de onda dentro da *fingerprint*, a banda em 1643 cm^{-1} foi atribuída predominantemente aos estiramentos ν(C=O), ν(C=N) e contribuições de ν(C=C) das bases nitrogenadas, em concordância com atribuições clássicas para modos vibracionais de adenina, timina, citosina e guanina descritas em estudos de espectroscopia de infravermelho com DNA (Souza *et al*., 2024; Banyay; Sarkar; Gräslund, 2003; Brewer *et al*., 2002). Ainda associadas às bases, as bandas em 1520 e 1454 cm^{-1} refletem vibrações do anel heterocíclico, incluindo estiramentos ν(C=C) e deformações assimétricas de grupos CH, combinadas a contribuições de ν(C=N), também amplamente descritas como marcadores vibracionais das estruturas aromáticas das bases (Souza *et al*., 2024; Banyay; Sarkar; Gräslund, 2003).

A banda em 1417 cm^{-1} apresenta caráter vibracional misto, envolvendo deformações δ(C–H), contribuições ν(C–N) e vibrações associadas a grupos N–H, refletindo simultaneamente a participação das bases nitrogenadas e da desoxirribose (Souza *et al*., 2024; Banyay; Sarkar; Gräslund, 2003). De forma semelhante, as bandas em 1365 e 1333 cm^{-1} também evidenciam contribuições combinadas entre bases e açúcar, incluindo deformações δ(C–H) e estiramentos ν(C–N) e ν(C=N), indicando acoplamento vibracional entre anéis nitrogenados e a estrutura do esqueleto molecular (Souza *et al*., 2024; Banyay; Sarkar; Gräslund, 2003).

Na faixa associada aos grupos fosfato, destacam-se as bandas em 1261 e 1211 cm^{-1}, atribuídas aos estiramentos assimétricos $\nu as(PO_2^-)$, característicos das ligações fosfodiéster que conectam os nucleotídeos ao longo da cadeia polimérica (Souza *et al*., 2024; Mello; Vidal, 2012; Banyay; Sarkar; Gräslund, 2003; Brewer *et al*., 2002). Essas vibrações representam marcadores estruturais importantes da integridade do *backbone* do DNA. Complementarmente, as bandas em 1130 e 1108 cm^{-1} envolvem contribuições combinadas de $\nu s(PO_2^-)$, ν(P–O–C), ν(C–O) e ν(C–C), refletindo vibrações coletivas da ligação fosfodiéster acopladas à desoxirribose (Souza *et al*., 2024; Banyay; Sarkar; Gräslund, 2003).

Na região inferior da *fingerprint*, as bandas em 1043 e 990 cm^{-1} foram atribuídas a vibrações do esqueleto açúcar-fosfato, incluindo estiramentos ν(C–O), ν(P–O–C), ν(C–C) e deformações associadas a grupos CH_2OH da desoxirribose. Essas bandas refletem modos vibracionais da cadeia principal do DNA e são frequentemente relacionadas à organização estrutural do *backbone* (Souza *et al*., 2024; Banyay; Sarkar; Gräslund, 2003). A banda em 924 cm^{-1}, por sua vez, foi atribuída ao estiramento ν(C–O–C) do anel da desoxirribose, enquanto as bandas em 867 e 831 cm^{-1} correspondem a vibrações combinadas ν(C–O) e ν(C–C) do açúcar, caracterizando deformações estruturais do anel furanose (Banyay; Sarkar; Gräslund, 2003).

De forma integrada, o conjunto de bandas identificado demonstra que os espectros obtidos para os *amplicons* de PCR podem ser, em partes, relacionados a contribuições vibracionais do DNA, abrangendo modos associados às bases nitrogenadas, aos grupos fosfato e à desoxirribose. A concordância entre as bandas observadas e aquelas descritas na literatura reforça a consistência espectral dos dados obtidos e indica que as regiões selecionadas para as análises multivariadas concentram informações estruturais relevantes da molécula amplificada. No entanto, não se pode descartar completamente a presença de contribuições residuais provenientes de reagentes da reação de PC, que podem apresentar sobreposições espectrais pontuais ao longo da região de *fingerprint*.

Embora o presente trabalho represente, até onde é de nosso conhecimento, o primeiro estudo a investigar *amplicons* de PCR humanos por espectroscopia ATR-FTIR combinada a abordagens de aprendizado de máquina para discriminação orientada por genótipo, estudos anteriores (Souza *et al*., 2024; Rios *et al*., 2021; Han *et al*., 2018; Nurdalila *et al*., 2015; Qiu *et al*., 2015; Song *et al*., 2014; Emura *et al*., 2006) já demonstraram o potencial de técnicas baseadas em FTIR para a análise de DNA em diferentes contextos biológicos.

Emura *et al*. (2006) aplicaram com sucesso a espectroscopia FTIR para diferenciar DNA genômico de plantas, incluindo a discriminação de variedades de milho e arroz. Song *et al*. (2014) investigaram linhagens híbridas e parentais de *Brassica campestris* por meio de análise

multivariada, revelando variações sutis em grupos funcionais do DNA. Qiu *et al*. (2015) também relataram elevada acurácia de classificação baseada em impressões digitais espectrais de DNA para identificação de variedades vegetais, como demonstrado para *Camellia reticulata*. Nurdalila *et al*. (2015) aplicaram ATR-FTIR em amostras animais e classificaram populações de peixes utilizando marcadores gênicos mitocondriais. Han *et al*. (2018) demonstraram alta sensibilidade e especificidade para discriminação de DNA animal, sugerindo aplicações em áreas como controle de alimentos e ciência forense. Até o momento, a única aplicação reportada da análise de *amplicons* de PCR por FTIR para discriminação genotípica foi descrita por Rios *et al*. (2021), que investigaram *amplicons* de DNA bovino direcionados a um SNP associado a características de crescimento em bovinos, combinando espectroscopia FTIR com abordagens quimiométricas e de ML. Mais recentemente, Souza *et al*. (2024) demonstraram que a análise ATR-FTIR de produtos de DNA amplificado provenientes de amostras humanas pode ser explorada para discriminar indivíduos saudáveis daqueles com síndrome metabólica utilizando DNA livre circulante, reforçando ainda mais a aplicabilidade da espectroscopia vibracional na investigação de distúrbios genéticos e metainflamatórios.

Uma comparação direta com dois estudos que investigaram *amplicons* de PCR por FTIR reforça ainda mais os padrões espectrais observados no presente trabalho. Rios *et al*. (2021) relataram contribuições vibracionais dominantes em *amplicons* de DNA bovino na região de 800–1250 cm^{-1}, associadas a vibrações do esqueleto fosfato e da desoxirribose, bem como bandas entre 1250–1500 cm^{-1} e 1500–1800 cm^{-1} relacionadas a modos das bases nitrogenadas e interações de empilhamento de bases. Essas regiões correspondem de forma estreita às identificadas neste estudo, particularmente as bandas relacionadas ao fosfato próximas de ~1230–1245 cm^{-1} ($\nu as\ PO_2^-$) e ~1100–1050 cm^{-1} ($\nu s\ PO_2^-$ e estiramento P–O–C). De maneira semelhante, Souza *et al*. (2024) descreveram bandas FTIR bem definidas em *amplicons* de DNA em aproximadamente 1230, 1106, 1031 e 923–991 cm^{-1}, todas atribuídas a componentes nucleotídicos, em forte concordância com as principais feições de absorção observadas neste estudo. Apesar dessas similaridades, diferenças nas intensidades relativas das bandas e no comportamento da linha de base foram evidentes, refletindo variações na química de PCR entre kits comerciais, incluindo principalmente a composição de tampões e estabilizadores enzimáticos (Green; Sambrook, 2019; Lorenz, 2012). Esses componentes introduzem contribuições vibracionais sobrepostas que podem modular o perfil espectral global, indicando que, embora a assinatura vibracional central do DNA seja preservada, a expressão espectral específica de *amplicons* de PCR é influenciada pelo sistema de amplificação, reforçando a

importância de pré-processamentos espectrais adequados e do uso de controles internos (Baker *et al*., 2014; Lasch, 2012).

Um aspecto metodológico importante do presente estudo é a subtração explícita do espectro médio dos NTCs a partir dos espectros médios dos *amplicons* de PCR, etapa que não havia sido sistematicamente abordada em análises anteriores de FTIR em *amplicons* de DNA. Enquanto estudos prévios (Souza *et al*., 2024; Rios *et al*., 2021) demonstraram que o DNA amplificado mantém assinaturas vibracionais características detectáveis por FTIR, a contribuição dos reagentes de PCR para o perfil espectral global não havia sido isolada. Em contraste, a inclusão da subtração média dos NTCs neste trabalho permitiu uma avaliação mais direta de regiões espectrais menos afetadas por contribuições de fundo oriundas da química de amplificação. Diversos componentes comumente presentes em misturas de PCR são conhecidos por apresentar fortes bandas de absorção no infravermelho, incluindo o glicerol utilizado como estabilizador enzimático e tampões à base de Tris (Lorenz, 2012). O glicerol apresenta bandas intensas de estiramento O–H na região de 3200–3600 cm^{-1} e vibrações de estiramento C–O entre aproximadamente 1000 e 1150 cm^{-1}, que podem se sobrepor a modos associados ao DNA (Khatib; Ramburrun; Choonara, 2025; Sirbu *et al*., 2024; Stuart, 2004) enquanto compostos de Tris contribuem com bandas relacionadas a vibrações O–H, N–H e C–O, particularmente na região de 1000–1200 cm^{-1} (Kondratenko *et al*., 2022; Sugiura; Makita, 2019).

A partir da subtração do espectro médio dos NTCs, a presente análise revelou que os intervalos de 900–1100 cm^{-1} e 2800–3800 cm^{-1} apresentaram influência comparativamente reduzida de sinais derivados de reagentes. A região de menor número de onda está predominantemente associada a vibrações do esqueleto fosfato e do açúcar do DNA, enquanto a região de maior número de onda engloba modos coletivos de estiramento que podem refletir características de hidratação e conformação do ácido nucleico (Kahn *et al*., 2009). Essas observações sustentam o uso da subtração média dos NTCs como uma estratégia eficaz para mitigar interferências espectrais relacionadas a reagentes e aumentar a sensibilidade à informação vibracional intrínseca do DNA em análises FTIR de *amplicons* de PCR.

A análise exploratória por PCA realizada sobre os espectros ATR-FTIR normalizados por SNV ilustra de forma adicional os desafios associados à discriminação não supervisionada de genótipos com base em espectros de *amplicons* de PCR. No presente estudo, a PCA aplicada às regiões de 2800–3800 cm^{-1} e 900–1100 cm^{-1} capturou uma proporção substancial da variância espectral total nos dois primeiros componentes principais, alcançando valores próximos de 99%. No entanto, os gráficos de escores construídos a partir de combinações pareadas dos quatro primeiros PCs demonstraram, de forma consistente, ampla sobreposição

entre os grupos genotípicos, tanto quando todas as amostras foram analisadas conjuntamente quanto nas comparações pareadas entre genótipos. Esse resultado indica que as principais fontes de variabilidade presentes nos espectros não estão diretamente alinhadas às diferenças relacionadas ao genótipo, mas refletem, sobretudo, características espectrais globais compartilhadas entre as amostras.

Resultado semelhante foi reportado no estudo com DNA bovino Rios *et al*. (2021), no qual a PCA aplicada a espectros FTIR normalizados por SNV de *amplicons* de DNA, em diferentes intervalos espectrais, demonstrou que, apesar de os três primeiros componentes principais explicarem mais de 98% da variância total, não foi observada formação de agrupamentos claros de acordo com o genótipo. Em ambos os estudos, o elevado grau de similaridade espectral entre os *amplicons* limitou a capacidade de métodos lineares não supervisionados em resolver padrões específicos de genótipo. Em conjunto, esses achados reforçam a interpretação de que, embora a PCA seja eficiente para resumir a variância espectral global e identificar os principais contribuintes da variabilidade, ela se mostra insuficiente para a discriminação genotípica nesse contexto, justificando assim a aplicação de abordagens supervisionadas de aprendizado de máquina capazes de capturar relações sutis e não lineares incorporadas aos espectros FTIR.

A aplicação de modelos supervisionados de aprendizado de máquina aos espectros ATR-FTIR de *amplicons* de PCR neste estudo demonstrou que informações relacionadas ao genótipo podem ser extraídas de padrões espectrais multivariados sutis, ainda que com desempenho de classificação moderado. Os melhores resultados foram obtidos nas comparações pareadas entre genótipos, particularmente aquelas envolvendo o genótipo AA, com o modelo DL-MLP Residual Simulated alcançando AUC média de aproximadamente 65,4% e acurácia de 71,6% para a discriminação AA×GG na região de 2800–3800 cm^{-1}. Por outro lado, nas regiões de *fingerprint* dominadas por DNA (900–1100 cm^{-1} e 950–1200 cm^{-1}), modelos lineares mais simples, como a regressão logística, apresentaram melhor desempenho, atingindo valores de AUC de até ~64% e acurácia em torno de 70%.

Em comparação, Rios *et al*. (2021) reportaram acurácias de classificação superiores para *amplicons* de DNA bovino após redução de dimensionalidade baseada em PCA, alcançando aproximadamente 75% de acurácia em classificação de três classes e até 95% em comparações pareadas específicas utilizando modelos SVM e k-nearest neighbors (KNN) na faixa de 1800–800 cm^{-1}, com mais de 98–99% da variância dos dados explicada pelos componentes principais selecionados. Embora o presente trabalho tenha adotado uma estratégia distinta, aplicando modelos supervisionados diretamente sobre espectros normalizados por

SNV dentro de intervalos espectrais previamente selecionados por subtração média dos NTCs, ambos os estudos indicam de forma consistente que métodos não supervisionados baseados apenas em variância são insuficientes para discriminação genotípica, sendo necessário o uso de aprendizado supervisionado para resolver as diferenças espectrais sutis associadas à variação alélica em *amplicons* de PCR.

Para além do desempenho analítico, a estratégia proposta combinando ATR-FTIR ML responde diretamente à crescente demanda por abordagens mais eficientes em termos de tempo e custo para triagem de SNPs. No presente estudo, o fluxo de trabalho baseado em ATR-FTIR demonstrou consumo de reagentes significativamente reduzido, com custos estimados inferiores a USD 1 por amostra, mantendo um tempo total de análise de aproximadamente 6 minutos por amostra em lotes de 96 amostras.

Esses achados posicionam a abordagem proposta como uma estratégia de pré-triagem de baixa carga operacional, particularmente adequada para contextos exploratórios e de genotipagem em larga escala. Quando comparada a metodologias consolidadas, essa vantagem torna-se ainda mais evidente. Abordagens de sequenciamento direcionado, incluindo fluxos baseados em sequenciamento de nova geração, estão consistentemente associadas a maior complexidade operacional, maiores tempos de processamento (aproximadamente 15 minutos por amostra) e custos de reagentes substancialmente mais elevados, podendo ultrapassar USD 60 por amostra, dependendo da configuração da plataforma e da profundidade de sequenciamento (Rios *et al*., 2021; Green; Sambrook, 2019; Lorenz, 2012). A genotipagem de SNPs por PCR quantitativa em tempo real utilizando ensaios TaqMan, embora mais rápida e acessível, ainda apresenta custos de reagentes na faixa de USD 3–4 por amostra e requer aproximadamente 2,5 minutos por amostra em formatos padrão de processamento em lote.

Essas observações estão em concordância com relatos prévios de Rios *et al*. (2021), que destacaram que o rastreamento de polimorfismos baseado em sequenciamento tradicional demanda múltiplas reações de sequenciamento para assegurar confiança em nível de nucleotídeo, resultando em aumentos significativos tanto no tempo de trabalho quanto nos custos operacionais. Em contraste, abordagens baseadas em FTIR acopladas ao aprendizado de máquina foram propostas como uma etapa eficaz de pré-triagem, permitindo reduções substanciais de tempo e custo ao restringir o sequenciamento confirmatório a um número limitado de regiões candidatas.

Os resultados aqui apresentados expandem esse conceito ao fornecer uma avaliação experimental e em nível de fluxo de trabalho da combinação entre ATR-FTIR e ML, demonstrando que tal abordagem pode, de forma realista, reduzir o consumo de reagentes e o

esforço operacional, ao mesmo tempo em que preserva informações discriminatórias relacionadas ao genótipo. Embora o desempenho de classificação obtido neste estudo permaneça moderado, o conjunto das evidências sugere que, com otimizações adicionais, incluindo aumento do número de amostras, aprimoramento de estratégias de pré-processamento espectral e avaliação de novas arquiteturas de ML, o fluxo baseado em ATR-FTIR possui potencial para evoluir como uma alternativa viável e escalável para triagem de SNPs e aplicações exploratórias de genotipagem.

7 CONCLUSÃO

Este estudo demonstra o potencial da aplicação da espectroscopia ATR-FTIR associada ao *machine learning* para a discriminação de *amplicons* de PCR de acordo com diferenças genotípicas relacionadas a SNPs, constituindo-se como uma ferramenta complementar para a triagem de genotipagem em larga escala e não como uma técnica substituta dos métodos moleculares já estabelecidos. A partir da análise do SNP −3826A/G no gene *UCP1*, os resultados evidenciaram que as características vibracionais do DNA contêm informações associadas ao tipo de SNP (GG, GA ou AA) que podem ser visualizadas por meio de modelos multivariados, mesmo diante da elevada similaridade espectral e da presença de contribuições de fundo relacionadas aos reagentes da reação.

O desempenho dos modelos aplicados para a classificação das classes genotípicas foi moderado, com AUC e acurácia máximas de 0,654 e 0,716 respectivamente, na discriminação entre os genótipos AA e GG com o modelo de DL-MLP com arquitetura residual simulada na região de 2800–3800 cm^{-1}. Adicionalmente, constatou-se que a técnica apresenta viabilidade de aplicação em relação ao tempo e custo empregados na sua aplicação quando comparado com as técnicas padrão ouro de genotipagem por qPCR e sequenciamento NGS.

Embora o desempenho de classificação obtido ainda foi sucinto, vale ressaltar que este é um estudo exploratório e pioneiro na aplicação de ML em amostras de DNA humano amplificado. O desenvolvimento deste estudo possibilitou a consolidação de competências técnico-científicas avançadas, abrangendo desde a padronização experimental em biologia molecular e espectroscopia até a implementação e interpretação de modelos multivariados e algoritmos de ML, contribuindo de forma significativa para a formação acadêmica e profissional e para o estabelecimento de novas perspectivas de investigação em genotipagem. Estudos futuros com um maior conjunto amostral, com SNPs mais heterogêneos, inclusão de novos SNPs como alvos e a expansão das estratégias de ML poderão aumentar o potencial dessa análise na discriminação de polimorfismos do genoma humano.

REFERÊNCIAS

AGUIAR-PULIDO, Vanessa *et al.* Machine Learning Techniques for Single Nucleotide Polymorphism—Disease Classification Models in Schizophrenia. **Molecules**, v. 15, n. 7, p. 4875–4889, 12 jul. 2010.

ALBEGALI, Abdullah Abdo *et al.* Genetic association of insulin receptor substrate-1 (IRS-1, rs1801278) gene with insulin resistant of type 2 diabetes mellitus in a Pakistani population. **Molecular Biology Reports**, v. 46, n. 6, p. 6065–6070, 24 dez. 2019.

ALI, Yasir *et al.* MIR149 rs2292832 and MIR499 rs3746444 Genetic Variants Associated with the Risk of Rheumatoid Arthritis. **Genes**, v. 14, n. 2, p. 431, 8 fev. 2023.

AZEVEDO, Pedro Guimarães de *et al.* Genetic association of the PERIOD3 (PER3) Clock gene with extreme obesity. **Obesity Research & Clinical Practice**, v. 15, n. 4, p. 334–338, jul. 2021.

BAKAY, Kadir *et al.* Effects of HRG and TP73 gene variations on ovarian response. **Gynecological Endocrinology**, v. 38, n. 3, p. 243–247, 4 mar. 2022.

BAKER, Matthew J. *et al.* Using Fourier transform IR spectroscopy to analyze biological materials. **Nature Protocols**, v. 9, n. 8, p. 1771–1791, 3 ago. 2014.

BANYAY, Martina; SARKAR, Munna; GRÄSLUND, Astrid. A library of IR bands of nucleic acids in solution. **Biophysical Chemistry**, v. 104, n. 2, p. 477–488, jun. 2003.

BONAKDARI, Hossein *et al.* Single nucleotide polymorphism genes and mitochondrial DNA haplogroups as biomarkers for early prediction of knee osteoarthritis structural progressors: use of supervised machine learning classifiers. **BMC Medicine**, v. 20, n. 1, p. 316, 12 set. 2022.

BOUILLAUD, Frédéric; ALVES-GUERRA, Marie-Clotilde; RICQUIER, Daniel. UCPs, at the interface between bioenergetics and metabolism. **Biochimica et Biophysica Acta (BBA) - Molecular Cell Research**, v. 1863, n. 10, p. 2443–2456, out. 2016.

BRASIL. **Vigitel Brasil 2023:** vigilância de fatores de risco e proteção para doenças crônicas por inquérito telefônico: estimativas sobre frequência e distribuição sociodemográfica de fatores de risco e proteção para doenças crônicas nas capitais dos 26 estados brasileiros e no Distrito Federal em 2023. Brasília: Ministério da Saúde, 2023.

BREWER, Scott H. *et al.* Detection of DNA Hybridization on Gold Surfaces by Polarization Modulation Infrared Reflection Absorption Spectroscopy. **Langmuir**, v. 18, n. 11, p. 4460–4464, 1 maio 2002.

BRONDANI, Letícia de Almeida *et al.* The role of the uncoupling protein 1 (UCP1) on the development of obesity and type 2 diabetes mellitus. **Arquivos Brasileiros de Endocrinologia & Metabologia**, v. 56, n. 4, p. 215–225, jun. 2012.

BRONDANI, Letícia de Almeida *et al.* Association of the UCP polymorphisms with susceptibility to obesity: case–control study and meta-analysis. **Molecular Biology Reports**, v. 41, n. 8, p. 5053–5067, 22 ago. 2014.

CAIXETA, Douglas Carvalho *et al.* Salivary ATR-FTIR Spectroscopy Coupled with Support Vector Machine Classification for Screening of Type 2 Diabetes Mellitus. **Diagnostics**, v. 13, n. 8, p. 1396, 12 abr. 2023.

CHAN, Eugene Y. Next-Generation Sequencing Methods: Impact of Sequencing Accuracy on SNP Discovery. **Method. Mol. Biol.**, v. 578, p. 95–111, 2009.

CHATHOTH, Shahanas *et al.* Association of Uncoupling Protein 1 (UCP1) gene polymorphism with obesity: a case-control study. **BMC Medical Genetics**, v. 19, n. 1, p. 203, 20 dez. 2018.

CHAUDHARY, Rishabh; GUPTA, Sumeet; CHAUHAN, Samrat. Protein Uncoupling as an Innovative Practice in Diabetes Mellitus Treatment: A Metabolic Disorder. **Endocrine, Metabolic & Immune Disorders - Drug Targets**, v. 23, n. 4, p. 494–502, abr. 2023.

CHAUDHURI, Plaban *et al.* Role of Metabolic Risk Factors, Family History, and Genetic Polymorphisms (PPARγ and TCF7L2) on Type 2 Diabetes Mellitus Risk in an Asian Indian Population. **Public Health Genomics**, v. 24, n. 3–4, p. 131–138, 2021.

CHENG, Chu; FEI, Zhongjie; XIAO, Pengfeng. Methods to improve the accuracy of next-generation sequencing. **Frontiers in Bioengineering and Biotechnology**, v. 11, 20 jan. 2023.

CHIARELLA, Pieranna; CAPONE, Pasquale; SISTO, Renata. Contribution of Genetic Polymorphisms in Human Health. **International Journal of Environmental Research and Public Health**, v. 20, n. 2, p. 912, 4 jan. 2023.

CHO, Eunjin *et al.* Single nucleotide polymorphism marker combinations for classifying Yeonsan Ogye chicken using a machine learning approach. **Journal of Animal Science and Technology**, v. 64, n. 5, p. 830–841, set. 2022.

CHOI, R. Y. *et al.* Introduction to Machine Learning, Neural Networks, and Deep Learning. **Translational vision science & technology**, v. 9, n. 2, p. 14, 2020.

DEBERNEH, Henock M.; KIM, Intaek. Prediction of Type 2 Diabetes Based on Machine Learning Algorithm. **International Journal of Environmental Research and Public Health**, v. 18, n. 6, p. 3317, 23 mar. 2021.

DÍAZ-GARCÍA, Juan Daniel *et al.* Association Study of CACNA1D, KCNJ11, KCNQ1, and CACNA1E Single-Nucleotide Polymorphisms with Type 2 Diabetes Mellitus. **International Journal of Molecular Sciences**, v. 25, n. 17, p. 9196, 24 ago. 2024.

DONG, Liu *et al.* Evaluation of Fourier transform infrared (FTIR) spectroscopy with multivariate analysis as a novel diagnostic tool for lymph node metastasis in gastric cancer. **Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy**, v. 289, p. 122209, mar. 2023.

ELMI, Fatemeh *et al.* Application of FT-IR spectroscopy on breast cancer serum analysis. **Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy**, v. 187, p. 87–91, dez. 2017.

EMURA, Koji *et al.* Estimation for Different Genotypes of Plants based on DNA Analysis using Near-infrared (NIR) and Fourier-transform Infrared (FT-IR) Spectroscopy. **Breeding Science**, v. 56, n. 4, p. 399–403, 2006.

FORGA, L. I. *et al.* [Influence of the polymorphism 03826 A --> G in the UCP1 gene on the components of metabolic syndrome. **Anales del Sistema Sanitario de Navarra**, v. 26, n. 2, jan. 2003.

GAIOLLA, Rafael Dezen; MORAES, Marcelo Padovani de Toledo; OLIVEIRA, Deilson Elgui. SNPs in genes encoding for IL-10, TNF-α, and NFκB p105/p50 are associated with clinical prognostic factors for patients with Hodgkin lymphoma. **PLOS ONE**, v. 16, n. 3, p. e0248259, 8 mar. 2021.

GAUDILLO, Joverlyn *et al.* Machine learning approach to single nucleotide polymorphism-based asthma prediction. **PLOS ONE**, v. 14, n. 12, p. e0225574, 4 dez. 2019.

GENITSARIDI, Irini *et al.* 11th edition of the IDF Diabetes Atlas: global, regional, and national diabetes prevalence estimates for 2024 and projections for 2050. **The Lancet Diabetes & Endocrinology**, v. 14, n. 2, p. 149–156, fev. 2026.

GONZÁLEZ-HERRERA, Lizbeth *et al.* Genetic variation of FTO: rs1421085 T>C, rs8057044 G>A, rs9939609 T>A, and copy number (CNV) in Mexican Mayan school-aged children with obesity/overweight and with normal weight. **American Journal of Human Biology**, v. 31, n. 1, 10 jan. 2019.

GONZÁLEZ-RECIO, Oscar *et al.* Detecting single-nucleotide polymorphism by single-nucleotide polymorphism interactions in rheumatoid arthritis using a two-step approach with machine learning and a Bayesian threshold least absolute shrinkage and selection operator (LASSO) model. **BMC Proceedings**, v. 3, n. S7, p. S63, 15 dez. 2009.

GOODWIN, Sara; MCPHERSON, John D.; MCCOMBIE, W. Richard. Coming of age: ten years of next-generation sequencing technologies. **Nature Reviews Genetics**, v. 17, n. 6, p. 333–351, 17 jun. 2016.

GREEN, Michael R.; SAMBROOK, Joseph. Polymerase Chain Reaction. **Cold Spring Harbor Protocols**, v. 2019, n. 6, p. pdb.top095109, 3 jun. 2019.

GREENER, Joe G. *et al.* A guide to machine learning for biologists. **Nature Reviews Molecular Cell Biology**, v. 23, n. 1, p. 40–55, 13 jan. 2022.

GUL, Ali *et al.* Role of the Polymorphisms of Uncoupling Protein Genes in Childhood Obesity and Their Association with Obesity-Related Disturbances. **Genetic Testing and Molecular Biomarkers**, v. 21, n. 9, p. 531–538, set. 2017.

GUTIÉRREZ-GALLEGO, Alberto *et al.* Combination of Machine Learning Techniques to Predict Overweight/Obesity in Adults. **Journal of Personalized Medicine**, v. 14, n. 8, p. 816, 31 jul. 2024.

HAN, Yahong *et al.* Insight into Rapid DNA-Specific Identification of Animal Origin Based on FTIR Analysis: A Case Study. **Molecules**, v. 23, n. 11, p. 2842, 1 nov. 2018.

HASHEMIAN, Leila *et al.* The role of the PPARG (Pro12Ala) common genetic variant on type 2 diabetes mellitus risk. **Journal of Diabetes & Metabolic Disorders**, v. 20, n. 2, p. 1385–1390, 20 dez. 2021.

HUI, Lester; DELMONTE, Terrye; RANADE, Koustubh. Genotyping Using the TaqMan Assay. **Current Protocols in Human Genetics**, v. 56, n. 1, jan. 2008.

HWA, Hsiao-Lin *et al.* A single nucleotide polymorphism panel for individual identification and ancestry assignment in Caucasians and four East and Southeast Asian populations using a machine learning classifier. **Forensic Science, Medicine and Pathology**, v. 15, n. 1, p. 67–74, 16 mar. 2019.

JANIĆ, Miodrag *et al.* Obesity: Recent Advances and Future Perspectives. **Biomedicines**, v. 13, n. 2, p. 368, 5 fev. 2025.

JIANG, Tammy; GRADUS, Jaimie L.; ROSELLINI, Anthony J. Supervised Machine Learning: A Brief Primer. **Behavior Therapy**, v. 51, n. 5, p. 675–687, set. 2020.

JIANG, Xiaohui *et al.* A development strategy to fast establish the Taqman qPCR based method to detect SNP mutations. **Human Cell**, v. 33, n. 4, p. 1331–1333, 1 jul. 2020.

KAABI, Yahia. MTNR 1B (rs10830963) Gene Polymorphism, but not MTNR 1A (rs2119882), Associated with Type 2 Diabetes Mellitus Risk in Saudi Arabia. **Clinical Laboratory**, v. 70, n. 01/2024, 2024.

KAHN, Talia R. *et al.* An FTIR Investigation of Flanking Sequence Effects on the Structure and Flexibility of DNA Binding Sites. **Biochemistry**, v. 48, n. 6, p. 1315–1321, 17 fev. 2009.

KAPOOR, Nitin *et al.* Prevalence of normal weight obesity and its associated cardio-metabolic risk factors – Results from the baseline data of the Kerala Diabetes Prevention Program (KDPP). **PLOS ONE**, v. 15, n. 8, p. e0237974, 25 ago. 2020.

KAZARIAN, Sergei G.; CHAN, K. L. Andrew. ATR-FTIR spectroscopic imaging: recent advances and applications to biological systems. **The Analyst**, v. 138, n. 7, p. 1940, 2013.

KHATIB, Sameera; RAMBURRUN, Poornima; CHOONARA, Yahya E. A Thermo-Photo-Ionic Crosslinked Gellan Gum Hydrogel with Gradient Biomechanic Modulation as a Neuromaterial for Peripheral Nerve Injury. **Gels**, v. 11, n. 9, p. 720, 10 set. 2025.

KIEĆ-WILK, B. *et al.* Correlation of the -3826A >G polymorphism in the promoter of the uncoupling protein 1 gene with obesity and metabolic disorders in obese families from southern Poland. **Journal of physiology and pharmacology: an official journal of the Polish Physiological Society**, v. 53, n. 3, p. 477–490, 2002.

KIM, Hyun Jun; LEE, Sang Yeoup; KIM, Cheol Min. Association between gene polymorphisms and obesity and physical fitness in Korean children. **Biology of Sport**, v. 35, n. 1, p. 21-27, 2018.

KINO, Saiko *et al.* Distinguishing IDH mutation status in gliomas using FTIR-ATR spectra of peripheral blood plasma indicating clear traces of protein amyloid aggregation. **BMC Cancer**, v. 24, n. 1, p. 222, 16 fev. 2024.

KLUSEK, Justyna *et al.* NOS2 Polymorphism in Aspect of Left and Right-Sided Colorectal Cancer. **Journal of Clinical Medicine**, v. 13, n. 4, p. 937, 6 fev. 2024.

KOCKUM, Ingrid; HUANG, Jesse; STRIDH, Pernilla. Overview of Genotyping Technologies and Methods. **Current Protocols**, v. 3, n. 4, 7 abr. 2023.

KONDRATENKO, Y. A. *et al.* Synthesis, structure and properties of tris(hydroxymethyl)aminomethane complexes with biogenic metal salts. **Inorganica Chimica Acta**, v. 530, p. 120705, jan. 2022.

KUMAR, Santosh *et al.* Genetic Association of Transcription Factor 7-Like-2 rs7903146 Polymorphism With Type 2 Diabetes Mellitus. **Cureus**, 22 jan. 2024a.

KUMAR, Harshit *et al.* Machine Learning-Aided Ultra-Low-Density Single Nucleotide Polymorphism Panel Helps to Identify the Tharparkar Cattle Breed: Lessons for Digital Transformation in Livestock Genomics. **OMICS: A Journal of Integrative Biology**, v. 28, n. 10, p. 514–525, 1 out. 2024b.

LASCH, Peter. Spectral pre-processing for biomedical vibrational spectroscopy and microspectroscopic imaging. **Chemometrics and Intelligent Laboratory Systems**, v. 117, p. 100–114, ago. 2012.

LEE, K. H. *et al.* Association of UCP1 -3826A/G and UCP3 -55C/T gene polymorphisms with obesity and its related traits among multi-ethnic Malaysians. **Ethnicity & disease**, v. 25, n. 1, p. 65–71, 2015.

LESLIE, L. Suzanne *et al.* High Definition Infrared Spectroscopic Imaging for Lymph Node Histopathology. **PLOS ONE**, v. 10, n. 6, p. e0127238, 3 jun. 2015.

LEWIS, Lavinia M. *et al.* Characterizing the Freeze–Drying Behavior of Model Protein Formulations. **AAPS PharmSciTech**, v. 11, n. 4, p. 1580–1590, 6 dez. 2010.

LI, Qing-Bo *et al*. *In vivo* and *in situ* detection of colorectal cancer using Fourier transform infrared spectroscopy. **World Journal of Gastroenterology**, v. 11, n. 3, p. 327, 2005.

LI, Yan-yan; WANG, Hui; ZHANG, Yang-yang. Melatonin receptor 1B gene rs10830963 C/G polymorphism associated with type 2 diabetes mellitus: An updated meta-analysis of 13,752 participants. **Heliyon**, v. 8, n. 11, p. e11786, nov. 2022.

LOBSTEIN, T. *et al*. **World Obesity Atlas 2022**. Londres: World Obesity Federation, 2022.

LONGO, Michele *et al.* Adipose Tissue Dysfunction as Determinant of Obesity-Associated Metabolic Complications. **International Journal of Molecular Sciences**, v. 20, n. 9, p. 2358, 13 maio 2019.

LORENZ, Todd C. Polymerase Chain Reaction: Basic Protocol Plus Troubleshooting and Optimization Strategies. **Journal of Visualized Experiments**, n. 63, 22 maio 2012.

LU, Ake Tzu-Hui *et al.* Prediction of serotonin transporter promoter polymorphism genotypes from single nucleotide polymorphism arrays using machine learning methods. **Psychiatric Genetics**, v. 22, n. 4, p. 182–188, ago. 2012.

LU, Wen-Hsin; CHANG, Yao-Ming; HUANG, Yi-Shuian. Alternative Polyadenylation and Differential Regulation of Ucp1: Implications for Brown Adipose Tissue Thermogenesis Across Species. **Frontiers in Pediatrics**, v. 8, 9 fev. 2021.

MARDIS, Elaine R. DNA sequencing technologies: 2006–2016. **Nature Protocols**, v. 12, n. 2, p. 213–218, 5 fev. 2017.

MARTÍNEZ-LÓPEZ, Erika *et al.* FTO rs9939609: T>A Variant and Physical Inactivity as Important Risk Factors for Class III Obesity: A Cross-Sectional Study. **Healthcare**, v. 12, n. 7, p. 787, 4 abr. 2024.

MARTINEZ-MARIN, David *et al.* Accounting for tissue heterogeneity in infrared spectroscopic imaging for accurate diagnosis of thyroid carcinoma subtypes. **Vibrational Spectroscopy**, v. 91, p. 77–82, jul. 2017.

MELLO, Maria Luiza S.; VIDAL, B. C. Changes in the Infrared Microspectroscopic Characteristics of DNA Caused by Cationic Elements, Different Base Richness and Single-Stranded Form. **PLoS ONE**, v. 7, n. 8, p. e43169, 24 ago. 2012.

MIZANI, Mehrdad A. *et al.* Identifying subtypes of type 2 diabetes mellitus with machine learning: development, internal validation, prognostic validation and medication burden in linked electronic health records in 420 448 individuals. **BMJ Open Diabetes Research & Care**, v. 12, n. 3, p. e004191, 4 jun. 2024.

MOHAMED, Amal Ahmed *et al.* Leptin Rs7799039 polymorphism is associated with type 2 diabetes mellitus Egyptian patients. **Archives of Physiology and Biochemistry**, p. 1–13, 15 out. 2023.

MONDAL, Partha Pratim *et al.* Review on machine learning-based bioprocess optimization, monitoring, and control systems. **Bioresource Technology**, v. 370, p. 128523, fev. 2023.

MOVASAGHI, Zanyar; REHMAN, Shazza; UR REHMAN, Dr. Ihtesham. Fourier Transform Infrared (FTIR) Spectroscopy of Biological Tissues. **Applied Spectroscopy Reviews**, v. 43, n. 2, p. 134–179, fev. 2008.

NELSON, D. L.; COX, M. M.; HOSKINS, A. A. **Princípios de bioquímica de Lehninger**. 8. ed. PortoAlegre: Artmed, 2022.

NICOLETTI, Carolina Ferreira *et al.* UCP1 -3826 A>G polymorphism affects weight, fat mass, and risk of type 2 diabetes mellitus in grade III obese patients. **Nutrition**, v. 32, n. 1, p. 83–87, jan. 2016.

NISHITA, Denise M. *et al.* Clinical trial participant characteristics and saliva and DNA metrics. **BMC Medical Research Methodology**, v. 9, n. 1, p. 71, 29 dez. 2009.

NOGUEIRA, Marcelo Saito *et al.* FTIR spectroscopy as a point of care diagnostic tool for diabetes and periodontitis: A saliva analysis approach. **Photodiagnosis and Photodynamic Therapy**, v. 40, p. 103036, dez. 2022.

NURDALILA, A'wani *et al.* Homogeneous Nature of Malaysian Marine Fish Epinephelus fuscoguttatus (Perciformes; Serranidae): Evidence Based on Molecular Markers, Morphology and Fourier Transform Infrared Analysis. **International Journal of Molecular Sciences**, v. 16, n. 7, p. 14884–14900, 2 jul. 2015.

NUSSBAUM, R. L.; MCINNES, R. R.; WILLARD, H. F. **Thompson & Thompson:** Genética Médica. 8. ed. Rio de Janeiro: Elsevier, 2016.

OKTAVIANTHI, Sukma *et al.* Uncoupling protein 2 gene polymorphisms are associated with obesity. **Cardiovascular Diabetology**, v. 11, n. 1, p. 41, 25 dez. 2012.

ONG, Kanyin Liane *et al.* Global, regional, and national burden of diabetes from 1990 to 2021, with projections of prevalence to 2050: a systematic analysis for the Global Burden of Disease Study 2021. **The Lancet**, v. 402, n. 10397, p. 203–234, jul. 2023.

PARK, Hyeon-Gyo; CHOI, Jeong-Hwa. Genetic variant rs9939609 in FTO is associated with body composition and obesity risk in Korean females. **BMJ Open Diabetes Research & Care**, v. 11, n. 6, p. e003649, 22 nov. 2023.

PEI, Xiaoting *et al.* Haplotype-based interaction of the PPARGC1A and UCP1 genes is associated with impaired fasting glucose or type 2 diabetes mellitus. **Medicine (United States)**, v. 96, n. 23, 1 jun. 2017.

PHU, Sagawah *et al.* Single Nucleotide Polymorphism at rs7903146 of Transcription Factor 7-like 2 gene Among Subjects with Type 2 Diabetes Mellitus in Myanmar. **Journal of the ASEAN Federation of Endocrine Societies**, v. 38, n. S1, p. 41–47, 10 maio 2023.

QADDOUMI, Mohammad *et al.* GALNT2 rs4846914 SNP Is Associated with Obesity, Atherogenic Lipid Traits, and ANGPTL3 Plasma Level. **Genes**, v. 13, n. 7, p. 1201, 4 jul. 2022.

QIU, Lu *et al.* Fourier Transform Infrared Spectroscopy of the DNA of the Chuxiong Population of Camellia reticulata Lindl. of China. **Spectroscopy Letters**, v. 48, n. 2, p. 120–127, 7 fev. 2015.

RASOOL, Shayaq Ul Abeer *et al.* Insulin Receptor Substrate 1 Gly972Arg (rs1801278) Polymorphism Is Associated with Obesity and Insulin Resistance in Kashmiri Women with Polycystic Ovary Syndrome. **Genes**, v. 13, n. 8, p. 1463, 17 ago. 2022.

RESENDE, Cristina Maria Mendes *et al.* Polymorphisms on rs9939609 FTO and rs17782313 MC4R genes in children and adolescent obesity: A systematic review. **Nutrition**, v. 91–92, p. 111474, nov. 2021.

RICCI, Claudia *et al.* The impact of CPT1B rs470117, LEPR rs1137101 and BDNF rs6265 polymorphisms on the risk of developing obesity in an Italian population. **Obesity Research & Clinical Practice**, v. 15, n. 4, p. 327–333, jul. 2021.

RIOS, Thaynádia Gomes *et al.* FTIR spectroscopy with machine learning: A new approach to animal DNA polymorphism screening. **Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy**, v. 261, p. 120036, nov. 2021.

RYMSZA, Taciana *et al.* Human papillomavirus detection using PCR and ATR-FTIR for cervical cancer screening. **Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy**, v. 196, p. 238–246, maio 2018.

SAFAEI, Mahmood *et al.* A systematic literature review on obesity: Understanding the causes & consequences of obesity and reviewing various machine learning approaches used to predict obesity. **Computers in Biology and Medicine**, v. 136, p. 104754, set. 2021.

SCHLEINITZ, Dorit; DISTEFANO, Johanna K.; KOVACS, Peter. Targeted SNP Genotyping Using the TaqMan® Assay. **Methods Mol. Biol.**, v. 700, p. 77–87, 2011.

SIQUEIRA, Laurinda F. S. *et al.* SVM for FT-MIR prostate cancer classification: An alternative to the traditional methods. **Journal of Chemometrics**, v. 32, n. 12, 23 dez. 2018.

SIRBU, Elena-Emilia *et al.* Influence of Plasticizers Concentration on Thermal, Mechanical, and Physicochemical Properties on Starch Films. **Processes**, v. 12, n. 9, p. 2021, 19 set. 2024.

SNP9(rs)Report. **Rs1800592.** 2024. Disponível em: https://www.ncbi.nlm.nih.gov/snp/rs1800592#seq_hash. Acesso em: 10 jan. 2026.

SONG, Seung Yeub *et al.* Fourier transform infrared (FT-IR) spectroscopy of genomic DNA to discriminate F1 progenies from their paternal lineage of Chinese cabbage (Brassica rapa subsp. pekinensis). **Molecular Breeding**, v. 33, n. 2, p. 453–464, 28 fev. 2014.

SOUZA, Bianca M. *et al.* Associations between UCP1 -3826A/G, UCP2 -866G/A, Ala55Val and Ins/Del, and UCP3 -55C/T Polymorphisms and Susceptibility to Type 2 Diabetes Mellitus: Case-Control Study and Meta-Analysis. **PLoS ONE**, v. 8, n. 1, p. e54259, 24 jan. 2013.

SOUZA, Nikolas M. P. *et al.* Structural characterization of DNA amplicons by ATR-FTIR spectroscopy as a guide for screening metainflammatory disorders in blood plasma. **Spectrochimica Acta - Part A: Molecular and Biomolecular Spectroscopy**, v. 310, 5 abr. 2024.

SRAMKOVA, D. *et al.* The UCP1 Gene Polymorphism A-3826G in Relation to DM2 and Body Composition in Czech Population. **Experimental and Clinical Endocrinology & Diabetes**, v. 115, n. 05, p. 303–307, 21 maio 2007.

STUART, Barbara H. **Infrared Spectroscopy**: Fundamentals and Applications. *[S.l.]*: John Wiley & Sons, 2004.

SUGIURA, Yuki; MAKITA, Yoji. Tris(hydroxymethyl)aminomethane Substitution into Octacalcium Phosphate. **Chemistry Letters**, v. 48, n. 11, p. 1304–1307, 5 nov. 2019.

TAM, Vivian *et al.* Benefits and limitations of genome-wide association studies. **Nature Reviews Genetics**, v. 20, n. 8, p. 467–484, 8 ago. 2019.

TEKLEMARIAM, Thomas A. *et al.* ATR-FTIR spectroscopy and machine/deep learning models for detecting adulteration in coconut water with sugars, sugar alcohols, and artificial sweeteners. **Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy**, v. 322, p. 124771, dez. 2024.

UMPIERREZ, G. E. *et al.* Hyperglycemic Crises in Adults With Diabetes: A Consensus Report. **Diabetes Care**, v. 47, n. 8, p. 1257–1275, 2024.

VIMALESWARAN, Karani S. *et al.* A Haplotype at the *UCP1* Gene Locus Contributes to Genetic Risk for Type 2 Diabetes in Asian Indians (CURES-72). **Metabolic Syndrome and Related Disorders**, v. 8, n. 1, p. 63–68, fev. 2010.

VISSCHER, Peter M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. **The American Journal of Human Genetics**, v. 101, n. 1, p. 5–22, jul. 2017.

VOURDOUMPA, Aikaterini; PALTOGLOU, George; CHARMANDARI, Evangelia. The Genetic Basis of Childhood Obesity: A Systematic Review. **Nutrients**, v. 15, n. 6, p. 1416, 15 mar. 2023.

WALD, N. *et al.* Identification of melanoma cells and lymphocyte subpopulations in lymph node metastases by FTIR imaging histopathology. **Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease**, v. 1862, n. 2, p. 202–212, fev. 2016.

WANG, Yijun *et al.* Associations between Aquaglyceroporin Gene Polymorphisms and Risk of Type 2 Diabetes Mellitus. **BioMed Research International**, v. 2018, p. 1–7, 27 nov. 2018.

WORLD OBESITY FEDERATION. **World Obesity Atlas 2024**. London, UK: World Obesity Federation, 2024.

WU, Lijun *et al.* A new risk locus in CHCHD5 for hypertension and obesity in a Chinese child population: a cohort study. **BMJ Open**, v. 7, n. 9, p. e016241, 11 set. 2017.

WU, Yanling *et al.* Risk Factors Contributing to Type 2 Diabetes and Recent Advances in the Treatment and Prevention. **International Journal of Medical Sciences**, v. 11, n. 11, p. 1185–1200, 2014.

XIE, Chenyao *et al.* The ADRB3 rs4994 polymorphism increases risk of childhood and adolescent overweight/obesity for East Asia's population: an evidence-based meta-analysis. **Adipocyte**, v. 9, n. 1, p. 77–86, 1 jan. 2020.

YANASEGARAN, Kevina *et al.* Single nucleotide polymorphisms (SNPs) that are associated with obesity and type 2 diabetes among Asians: a systematic review and meta-analysis. **Scientific Reports**, v. 14, n. 1, p. 20062, 29 ago. 2024.

YIN, Dan *et al.* FTO: a critical role in obesity and obesity-related diseases. **British Journal of Nutrition**, v. 130, n. 10, p. 1657–1664, 28 nov. 2023.

YU, Keping *et al.* Association between MC4R rs17782313 genotype and obesity: A meta-analysis. **Gene**, v. 733, p. 144372, abr. 2020.

YU, Songcheng *et al.* Characteristic and influencing factors of Taqman genotyping calling error. **Journal of Clinical Laboratory Analysis**, v. 32, n. 9, 26 nov. 2018.

ZHANG, Weitao *et al.* Noninvasive surface detection of papillary thyroid carcinoma by Fourier transform infrared spectroscopy. **Chemical Research in Chinese Universities**, v. 31, n. 2, p. 198–202, 13 abr. 2015.

ZHANG, Xiangyan *et al.* Analysis and comparison of machine learning methods for species identification utilizing ATR-FTIR spectroscopy. **Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy**, v. 308, p. 123713, mar. 2024.

GLOSSÁRIO

Amplicon · Produto de DNA amplificado obtido a partir de uma PCR

Background · Sinal de fundo de um espectro de FTIR

DL-MLP Funnel · Aprendizado Profundo com Perceptron Multicamadas em Arquitetura Funil (Deep Learning with Multilayer Perceptron Funnel Architecture)

DL-MLP Residual Simulated · Aprendizado Profundo com Perceptron Multicamadas com Conexões Residuais Simuladas (Deep Learning with Multilayer Perceptron with Simulated Residual Connections)

Fingerprint · Impressão digital molecular de uma amostra no FTIR na região de 1800-800 cm^{-1}

Medidas de Desempenho · conjunto de métricas que compõe os resultados de um modelo de ML e incluem AUC, acurácia, sensibilidade, especificidade e F1-score

Primers · Par de iniciadores que flanqueiam um *amplicon*

qPCR SNP Genotyping · Genotipagem de SNP por Reação em Cadeia da Polimerase Quantitativa em Tempo Real (qPCR-based SNP Genotyping)

Range · Faixa espectral em números de onda (cm^{-1})

Scans · Varreduras realizadas em um ATR-FTIR

SVM Linear · Máquina de Vetores de Suporte Linear (Linear Support Vector Machine)

SVM RBF Kernel · Máquina de Vetores de Suporte com Função de Base Radial (Support Vector Machine with Radial Basis Function Kernel)

ν · Estiramento

νs · Estiramento Simétrico

νas · Estiramento Assimétrico

δ · Deformação

δas · Deformação Assimétrica

APÊNDICE A: TERMO DE CONSENTIMENTO LIVRE E ESCLARECIDO – TCLE

Dados de Identificação da Pesquisa

Título do projeto: Relação do polimorfismo -866G/A do gene *UCP2* com o desenvolvimento de distúrbios multifatoriais na população brasileira.

Pesquisador Responsável: Prof.ª Dra. Renata de Azevedo Canevari.

Graduandos: Bianca Espindola Martins, Letícia Rodrigues Sobrinho, Aleph Vaz Arruda, João Victor Castro de Souza e Rayssa Gomes dos Santos.

Mestrando: Ramon Varella Diniz.

Doutorando: Igor Martins Alves Melo.

Instituições: UNIVERSIDADE DO VALE DO PARAÍBA – Univap e Instituto de Pesquisa e Desenvolvimento – IP&D. Avenida Shishima Hifumi, 2911. CEP: 12244-000. São José dos Campos - SP

Telefones para contato: (12) 3947-1165 e (12) 98159-1607

Nome do participante:

_____

Data de nascimento: _____ CPF: _____ RG:

_____

Idade: _____ Gênero: _____ Raça/cor: _____ Descendência (s):

_____

Telefone: _____ CEP: _____

Endereço:

_____

E-mail (para receber o resultado laboratorial bioquímico):

_____

Código de identificação:

O (A) Sr.(a) está sendo convidado(a) a participar do projeto de pesquisa que tem como tema/título "Relação do polimorfismo -866g/a do gene *UCP2* com o desenvolvimento de distúrbios multifatoriais na população brasileira em desenvolvimento" de responsabilidade da pesquisadora Prof.ª Dra. Renata de Azevedo Canevari e financiamento da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP Processo 2023/16314-6).

O objetivo do estudo é coletar uma amostra do seu sangue (3 a 4 ml), identificá-la por código, armazená-la em geladeira ou freezer e neste material analisar o DNA como também realizar exames de sangue. No DNA será pesquisado a presença de uma alteração específica que pode ter relação com uma maior chance de se desenvolver a Diabetes Mellitus tipo 2, ou ter facilidade para ganho de peso ou talvez uma maior possibilidade de desenvolver futuramente um câncer em relação a população geral. Os exames de sangue realizados serão para avaliar a glicemia, insulina, hemoglobina glicada, colesterol e triglicérides. Como auxílio à esta pesquisa, com a sua permissão e após a assinatura deste termo, será aplicado à você um questionário para obtermos suas informações de histórico pessoal e familiar em relação as doenças que serão avaliadas neste estudo.

Para a coleta de sangue, que será realizada por um biomédico no laboratório pertencente a universidade, será necessário que você esteja em jejum de 8 horas. Durante a sua participação, podem ocorrer hematoma, inflamação e alguma dor no local do braço em que foi coletado o sangue. Para minimizar isto, medidas de segurança serão tomadas, como a coleta de sangue em ambiente adequado e no local correto do corpo, higienização e descontaminação correta, ângulo e lado adequado da agulha corretos, uso de materiais estéreis e descartáveis, além do uso correto e completo dos equipamentos de proteção individual (EPI) pelo profissional que faz parte da Equipe de Pesquisa.

Esperamos, que esta pesquisa possa auxiliar a adoção de hábitos de vida mais saudáveis, no sentido de prevenir o surgimento de algumas dessas doenças. Esclarecemos que os usos das informações por você oferecidas estão submetidas às normas éticas destinadas à pesquisa envolvendo seres humanos, da Comissão Nacional de Ética em Pesquisa (CONEP) do Conselho Nacional de Saúde, do Ministério da Saúde.

A sua participação é de livre escolha e de forma anônima, em que permitirá a coleta de amostras do seu sangue para que sejam realizadas as análises descritas a partir da assinatura desta autorização. Informamos que há risco relativo a quebra de sigilo e confidencialidade na manipulação e publicação das informações obtidas. Contudo, para minimizar este risco iremos utilizar um código numérico correspondente ao seu nome e que será dado a você no momento da coleta de sangue no box do laboratório. Este código estará disponível em uma tabela que será consultada apenas pelos pesquisadores do projeto para garantirmos a confidencialidade das informações e a sua privacidade. Contudo, vale ressaltar que durante o atendimento na recepção da CPS da FCS todos as informações serão coletadas como qualquer outro cliente seguindo as normativas da Vigilância Sanitária citadas na RDC nº 302/2005 que são: nome completo, endereço, sexo, idade, telefone para contato e e-mail para envio do laudo, que consta os resultados laboratoriais, se assim você desejar.

As informações e materiais obtidos nesta pesquisa não poderão ser utilizados para outras finalidades que não sejam a desta pesquisa científica. Garantimos a você de se retirar desta pesquisa a qualquer momento, sem prejuízos ou sofrer quaisquer sanções ou constrangimentos.

No caso de gastos decorrentes da participação nesta pesquisa como, por exemplo, transporte e alimentação, você e o seu acompanhante serão imediatamente e integralmente ressarcidos de todos os gastos. No caso de algum dano, imediato ou tardio, decorrente desta pesquisa, você também tem direito de ser indenizado pelo pesquisador desta pesquisa, bem como a ter assistência gratuita, integral e imediata.

Sempre que desejar, você poderá entrar em contato para obter informações sobre este projeto de pesquisa, sobre sua participação ou resultados da pesquisa, com a pesquisadora responsável pelo telefone (12) 3947-1165 ou e-mail (rcanevari@univap.br). Você também pode entrar em contato com o CEP – Comitê de Ética em Pesquisa da Universidade do Vale do Paraíba (UNIVAP), corresponsável por garantir e zelar pelos direitos do participante da pesquisa, pelo telefone (12) 3947-1111, pelo e-mail cep@univap.br ou pessoalmente na Av. Shishima Hifumi, 2911, Urbanova – Bloco 11 – Instituto de Pesquisa e Desenvolvimento II, sala 13, de segunda a sexta-feira, das 08:00h às 12:00h.

Este termo está elaborado em duas vias, rubricadas em todas as suas páginas e assinadas, ao seu término, pelo participante da pesquisa e pelo pesquisador, sendo uma das vias entregue ao participante.

Eu, _____, fui informado e concordo em participar do projeto de pesquisa acima descrito.

São José dos Campos – SP, _____ de _____ de _____

_____	_____
Participante	Pesquisador Responsável Prof.ª Dra. Renata de Azevedo Canevari
_____	_____
Testemunha nº 1 Igor Martins Alves Melo	Testemunha nº 2 Ramon Varella Diniz

APÊNDICE B: CARACTERIZAÇÃO DEMOGRÁFICA E CLÍNICA DA POPULAÇÃO DO ESTUDO

Número da amostra	Gênero	Idade	Quadro de diabetes tipo 2	Quadro de IMC	Presença de câncer
1	Feminino	20	Não	19,27	Não
2	Feminino	20	Não	17,85	Não
3	Feminino	20	Não	25,59	Não
4	Feminino	22	Não	22,96	Não
5	Feminino	20	Não	32,77	Não
6	Masculino	21	Não	22,15	Não
7	Feminino	50	Tipo 2	19,53	Não
8	Feminino	21	Não	24,34	Não
9	Feminino	21	Não	18,66	Não
10	Feminino	21	Não	17,71	Não
11	Feminino	69	Pré-diabetes	26,83	Não
12	Feminino	73	Não	25,97	Não
13	Feminino	50	Não	32,78	Não
14	Feminino	65	Tipo 2	33,02	Sim
15	Masculino	23	Não	24	Não
16	Feminino	25	Não	21,99	Não
17	Feminino	25	Não	22,6	Não
18	Feminino	40	Sim	29,5	Não
19	Feminino	50	Não	19,23	Não
20	Feminino	33	Não	39,56	Não
21	Feminino	30	Tipo 2	23,11	Não
22	Masculino	49	Não	24,69	Não
23	Masculino	48	Não	27,7	Não
24	Feminino	72	Tipo 2	29,03	Não
25	Feminino	77	Tipo 2	29,29	Não
26	Feminino	78	Tipo 2	26,02	Não
27	Masculino	76	Tipo 2	39,45	Sim
28	Masculino	79	Tipo 2	25,39	Não
29	Feminino	74	Tipo 2	21,91	Sim
30	Feminino	56	Tipo 2	21,48	Não
31	Feminino	68	Tipo 2	22,88	Sim
32	feminino	56	Tipo 2	30,11	não
33	Masculino	67	Pré-diabetes	31,27	Sim
34	Feminino	52	Não	22,95	Sim
35	Masculino	69	Tipo 2	20,19	Sim
36	Feminino	53	Não	30,26	Não

37	Feminino	52	Pré-diabetes	36,57	Não
38	Feminino	36	Tipo 2	28	Não
39	Feminino	63	Não	22,74	Não
40	Feminino	67	Não	31,3	Não
41	Feminino	58	Pré-diabetes	31,9	Não
42	Masculino	74	Não	25,5	Não
43	Feminino	49	Não	34,55	Não
44	Feminino	56	Não	29,8	Não
45	Feminino	75	Não	26,8	Não
46	Feminino	79	Não	28,9	Não
47	Feminino	37	Não	35	Não
48	Feminino	20	Não	31	Não
49	Masculino	51	Tipo 2	34	Não
50	Feminino	22	Não	37	Não
51	Feminino	72	Tipo 2	27,2	Não
52	Feminino	47	Não	36	Não
53	Masculino	22	Não	37,9	Não
54	Feminino	19	Não	37,7	Não
55	Masculino	58	Não	27,88	Sim
56	Feminino	21	Não	38,16	Não
57	Feminino	20	Não	33,4	Não
58	Masculino	63	Tipo 2	29,6	Sim
59	Masculino	53	Tipo 2	34,07	Não
60	Feminino	61	Tipo 2	38,08	Sim
61	Feminino	42	Não	36,13	Não
62	Feminino	37	Não	33,5	Não
63	Masculino	41	Não	32,69	Não
64	Masculino	48	Tipo 2	30,11	Não
65	Masculino	32	Pré-diabetes	33,9	Não
66	Masculino	35	Não	44,9	Não
67	Feminino	66	Tipo 2	23,59	Sim
68	Feminino	82	Não	24,88	Sim
69	Masculino	51	Não	34,79	Não
70	Masculino	64	Não	29	Sim
71	Masculino	53	Não	31,02	Não
72	Masculino	56	Não	28,2	Não
73	Feminino	52	Pré-diabetes	30,74	Não
74	Feminino	51	Não	33,5	Não
75	Masculino	56	Não	20,16	Não

76	Feminino	52	Não	25,19	Sim
77	Masculino	74	Não	28,09	Não
78	Masculino	18	Não	33,03	Não
79	Masculino	58	Não	29,69	Sim
80	Feminino	54	Não	36,44	Não
81	Masculino	52	Não	31,56	Não
82	Feminino	75	Pré-diabetes	25,24	Não
83	Masculino	25	Não	40,74	Não
84	Masculino	62	Não	36,69	Não
85	Masculino	19	Não	39,6	Não
86	Feminino	63	Não	22,6	Não
87	Masculino	74	Tipo 2	21,45	Não
88	Feminino	50	Pré-diabetes	23,9	Não
89	Feminino	25	Tipo 2	29,1	Não
90	Feminino	56	Pré-diabetes	27,1	Não
91	Feminino	22	Não	42	Não
92	Feminino	65	Não	25,59	Não
93	Masculino	52	Não	30,9	Não
94	Feminino	57	Pré-diabetes	21,3	Não
95	Feminino	34	Não	56,64	Não
96	Feminino	19	Não	41,62	Não
97	Feminino	51	Não	27,55	Sim
98	Masculino	56	Não	25,36	Não
99	Masculino	67	Não	26,06	Não
100	Feminino	59	Não	29,01	Não
101	Feminino	64	Tipo 1	26,05	Não
102	Masculino	52	Não	24,87	Não
103	Masculino	47	Não	26,16	-
104	Feminino	34	Não	38.59	Não
105	Feminino	69	Não	39,18	Não
106	Feminino	28	Não	32.82	Não
107	feminino	53	Não	21.45	Não
108	Feminino	59	Tipo 2	25,07	Não

109	Feminino	60	Não	26,48	Não
110	Masculino	45	Não	28	-
111	Masculino	67	Tipo 2	32,35	Não
112	Feminino	60	Não	33,47	Não
113	Feminino	52	Não	37,98	Não
114	Feminino	53	Não	23,34	Sim
115	Feminino	46	Não	28,22	Não
116	Masculino	52	Não	29,87	Não
117	Feminino	25	Não	38,65	Não
118	Masculino	52	Tipo 2	24,37	Não
119	Feminino	77	Não	18,75	Não
120	Masculino	26	Não	25,6	Não
121	Feminino	20	Não	24,2	Não
122	Masculino	21	Não	29,05	Não
123	Masculino	29	Não	22,95	Não
124	Masculino	51	Não	23,49	Não
125	Feminino	53	Não	22,14	Não
126	Masculino	43	Tipo 2	31,66	Não
127	Masculino	50	Não	24,9	Não
128	Feminino	66	Não	25,9	Não
129	Masculino	50	Não	24.44	Não
130	Feminino	46	Não	24.57	Não
131	Feminino	69	Não	20.83	Não
132	Feminino	55	Não	23.44	Não
133	Feminino	52	Não	23.73	Não
134	Masculino	60	Não	22.22	Não
135	Feminino	55	Não	21,31	Não
136	Masculino	53	Não	22,4	Não
137	Masculino	55	Pré-diabético	22,94	Não
138	Feminino	42	Não	22,1	Não
139	Feminino	42	Não	21,83	Não
140	Masculino	59	Não	23,62	Não

141	Feminino	51	Não	22,49	Não
142	Feminino	52	Não	17,92	Não
143	Masculino	62	Não	27,3	Não
144	Feminino	58	Pré-diabético	25,24	Não
145	Feminino	54	Não	25,24	Sim
146	Feminino	54	Não	23,38	Não
147	Feminino	76	Tipo 2	29,69	Não
148	Feminino	36	Pré-diabético	29,02	Não
149	Feminino	71	Tipo 2	30,82	Não
150	Masculino	21	Não	39,42	Não
151	Feminino	46	Não	35,93	Não
152	Feminino	51	Não	31,44	Não
153	Masculino	48	Não	41,52	Não
154	Feminino	49	Não	30,86	Não
155	Masculino	54	Tipo 2	30,1	Não
156	Masculino	52	Não	23,84	Não
157	Feminino	63	Não	30,1	Não
158	Feminino	52	Tipo 2	26,48	Não
159	Masculino	54	Pré-diabético	33.02	Não
160	Feminino	68	Não	20,55	Não
161	Feminino	49	Não	17,8	Sim
162	Feminino	54	Não	26,44	Não
163	Feminino	54	Não	24,98	Não
164	Feminino	69	Não	30,3	Não
165	Feminino	52	Não	26,64	Não
166	Feminino	61	Não	27,34	Não
167	Feminino	64	Não	26,4	Não
168	Masculino	62	Pré-diabético	28,38	Não
169	Feminino	54	Não	23,73	Não
170	Feminino	57	Não	34,08	Não
171	Feminino	67	Tipo 2	47,86	Sim
172	Feminino	82	Não	23,62	Não

173	Feminino	67	Não	27,18	Não
174	Feminino	61	Não	26,78	Não
175	Feminino	74	Não	22,48	Sim
176	Feminino	51	Não	27,29	Não
177	Feminino	65	Pré-diabético	28,04	Sim
178	Feminino	52	Não	28	Não
179	Feminino	53	Não	23,42	Não
180	Feminino	45	Tipo 2	37,25	Não
181	Feminino	51	Não	26,56	Não
182	Feminino	32	Não	40,39	Não
183	Feminino	59	Não	28,26	Não
184	Feminino	50	Não	22,94	Sim
185	Feminino	39	Não	23,34	Não
186	Feminino	54	Não	21,64	Não
187	Feminino	35	Não	31,8	Não
188	Feminino	59	Tipo 2	22,41	Não
189	Feminino	65	Não	27,41	Não
190	Feminino	60	Não	21,34	Sim

APÊNDICE C: DADOS DE CONCENTRAÇÃO E RAZÕES DE CONTAMINAÇÃO DE PROTEÍNAS (260/280) E REAGENTES (260/230) DAS AMOSTRAS DE DNA EXTRAÍDAS DE SANGUE DOS 190 PARTICIPANTES

Número da Amostra	Método de extração	Concentração de DNA	Razão 260/280	Razão 260/230
1	Fenol-Clorofórmio	202,98	1,62	1,42
2	Fenol-Clorofórmio	145,27	1,42	1,14
3	Fenol-Clorofórmio	162,66	1,83	1,82
4	Fenol-Clorofórmio	210,24	1,51	1,27
5	Fenol-Clorofórmio	83,50	1,48	0,99
6	Fenol-Clorofórmio	57,50	1,41	0,88
7	Fenol-Clorofórmio	100,20	1,43	1,19
8	Fenol-Clorofórmio	1.242,91	1,90	2,00
9	Fenol-Clorofórmio	2.356,92	1,90	2,13
10	Fenol-Clorofórmio	41,20	1,69	1,46
11	Fenol-Clorofórmio	754,50	1,54	1,66
12	Fenol-Clorofórmio	1.101,70	1,73	1,95
13	Fenol-Clorofórmio	1.422,70	1,72	1,93
14	Fenol-Clorofórmio	1.540,50	1,67	1,86
15	Fenol-Clorofórmio	896,45	1,90	1,70
16	Fenol-Clorofórmio	668,50	1,80	1,89
17	Fenol-Clorofórmio	1.113,50	1,83	1,84
18	Fenol-Clorofórmio	2.723,62	1,87	2,21
19	Fenol-Clorofórmio	321,30	1,38	1,69
20	Fenol-Clorofórmio	262,80	2,20	1,80
21	Fenol-Clorofórmio	1.098,60	1,92	2,32
22	Fenol-Clorofórmio	139,00	1,91	1,88
23	Fenol-Clorofórmio	178,40	1,85	2,00
24	Fenol-Clorofórmio	401,10	1,88	1,43
25	Fenol-Clorofórmio	535,80	1,74	1,70
26	Fenol-Clorofórmio	465,60	1,77	1,53
27	Fenol-Clorofórmio	170,80	1,79	1,38
28	Fenol-Clorofórmio	236,80	1,77	1,72
29	Fenol-Clorofórmio	291,40	1,73	1,74
30	Fenol-Clorofórmio	237,50	1,82	1,49
31	Fenol-Clorofórmio	483,70	1,83	1,17
32	Fenol-Clorofórmio	514,50	1,72	0,91
33	Fenol-Clorofórmio	1.560,20	1,99	2,24
34	Fenol-Clorofórmio	1.438,08	1,84	2,13
35	Fenol-Clorofórmio	19,90	1,75	0,30
36	Fenol-Clorofórmio	698,70	1,99	1,93
37	Fenol-Clorofórmio	1.665,60	1,90	2,04
38	Fenol-Clorofórmio	1.003,20	1,98	1,99
39	Fenol-Clorofórmio	147,80	1,85	1,07
40	Fenol-Clorofórmio	823,80	1,95	1,96
41	Kit De Extração	272,33	1,81	1,51
42	Fenol-Clorofórmio	670,24	1,92	2,21
43	Fenol-Clorofórmio	118,99	1,99	1,87
44	Fenol-Clorofórmio	114,37	1,85	1,77
45	Fenol-Clorofórmio	470,86	1,89	2,18
46	Fenol-Clorofórmio	1.171,63	1,96	2,16
47	Fenol-Clorofórmio	945,18	1,77	2,25
48	Fenol-Clorofórmio	1.575,39	1,90	2,27
49	Fenol-Clorofórmio	984,01	1,99	1,89
50	Fenol-Clorofórmio	1.479,04	1,96	2,11
51	Fenol-Clorofórmio	177,50	1,83	1,31
52	Fenol-Clorofórmio	2.550,40	1,88	2,14
53	Fenol-Clorofórmio	244,35	1,80	1,42
54	Fenol-Clorofórmio	168,48	1,86	1,25
55	Fenol-Clorofórmio	926,30	1,98	2,07
56	Fenol-Clorofórmio	142,46	1,85	1,21
57	Fenol-Clorofórmio	531,04	1,92	1,90
58	Fenol-Clorofórmio	1.671,26	2,00	2,12
59	Fenol-Clorofórmio	1.884,09	2,02	2,12
60	Fenol-Clorofórmio	706,86	1,92	1,84

61	Fenol-Clorofórmio	294,16	1,85	1,58
62	Fenol-Clorofórmio	311,39	1,84	1,34
63	Fenol-Clorofórmio	309,85	1,78	1,25
64	Fenol-Clorofórmio	1.226,90	1,89	2,21
65	Fenol-Clorofórmio	1.441,60	1,89	2,23
66	Fenol-Clorofórmio	1.005,60	1,91	2,09
67	Fenol-Clorofórmio	2.260,58	2,04	2,27
68	Fenol-Clorofórmio	2.330,58	1,90	2,14
69	Fenol-Clorofórmio	3.588,45	1,93	2,20
70	Fenol-Clorofórmio	585,53	2,24	2,06
71	Fenol-Clorofórmio	169,08	2,02	1,43
72	Fenol-Clorofórmio	96,18	1,94	0,96
73	Fenol-Clorofórmio	273,72	1,95	1,56
74	Fenol-Clorofórmio	2.232,39	1,88	2,26
75	Fenol-Clorofórmio	1.158,04	1,90	2,29
76	Fenol-Clorofórmio	537,33	1,92	2,21
77	Fenol-Clorofórmio	650,32	1,95	2,26
78	Kit De Extração	167,08	1,88	1,92
79	Kit De Extração	123,26	1,96	2,28
80	Kit De Extração	606,40	1,94	2,14
82	Kit De Extração	934,17	1,97	2,26
83	Kit De Extração	964,99	1,96	2,23
84	Kit De Extração	478,21	1,87	2,00
85	Kit De Extração	1.010,22	1,91	2,00
86	Kit De Extração	504,28	1,84	1,76
87	Kit De Extração	200,60	1,81	1,16
88	Kit De Extração	414,87	1,86	1,96
89	Kit De Extração	710,00	1,89	1,67
90	Kit De Extração	452,40	1,86	1,98
91	Kit De Extração	849,77	1,83	1,40
92	Kit De Extração	579,03	1,80	1,60
93	Kit De Extração	308,56	1,86	1,02
94	Kit De Extração	313,82	1,84	1,63
95	Kit De Extração	426,73	1,80	1,24
96	Kit De Extração	521,38	1,85	1,91
97	Kit De Extração	362,40	1,85	1,79
98	Kit De Extração	323,24	1,82	1,54
99	Kit De Extração	74,73	1,84	0,89
100	Kit De Extração	475,10	1,88	1,89
101	Kit De Extração	881,18	1,89	2,09
102	Kit De Extração	597,67	1,92	1,95
103	Kit De Extração	936,86	1,89	2,17
104	Kit De Extração	130,10	1,94	1,56
105	Kit De Extração	35,58	1,95	0,78
106	Kit De Extração	62,51	2,09	1,17
107	Kit De Extração	327,87	1,82	1,35
108	Kit De Extração	62,51	1,95	1,33
109	Kit De Extração	972,63	1,86	2,11
110	Kit De Extração	42,77	1,08	1,08
111	Kit De Extração	1.128,69	1,86	1,83
112	Kit De Extração	410,98	1,84	1,89
113	Kit De Extração	677,40	1,85	1,82
114	Kit De Extração	472,40	1,84	1,71
115	Kit De Extração	865,21	1,88	2,11
116	Kit De Extração	62,95	1,81	1,18
117	Kit De Extração	239,41	1,86	1,63
118	Kit De Extração	625,92	1,86	1,73
119	Kit De Extração	186,15	2,09	2,36
120	Fenol-Clorofórmio	1.128,77	1,94	2,14
121	Fenol-Clorofórmio	757,87	1,91	1,84
122	Fenol-Clorofórmio	1.061,80	1,89	2,03
123	Fenol-Clorofórmio	213,90	1,77	1,79
124	Kit De Extração	344,13	1,96	1,94
125	Kit De Extração	911,10	1,87	2,00
126	Kit De Extração	948,90	1,93	2,04
127	Kit De Extração	93,70	1,55	1,11
128	Kit De Extração	165,5	1,85	1,16

129	Kit De Extração	58,35	1,95	1,31
130	Kit De Extração	566,3	1,82	1,76
131	Kit De Extração	260,98	1,89	1,9
132	Kit De Extração	47,12	1,48	0,86
133	Kit De Extração	445,75	1,84	1,93
134	Kit De Extração	43,99	1,79	1,17
135	Kit De Extração	170,81	1,78	1,05
136	Kit De Extração	108,9	1,72	0,5
137	Kit De Extração	28,48	1,82	0,68
138	Kit De Extração	297,79	1,84	1,92
139	Kit De Extração	147,77	1,81	1,14
140	Kit De Extração	913,98	1,84	1,79
141	Kit De Extração	778,07	1,86	1,96
142	Kit De Extração	171,59	1,82	1,18
143	Kit De Extração	187,47	1,81	1,24
144	Kit De Extração	413,7	1,83	1,21
145	Kit De Extração	73,99	1,57	1,2
146	Fenol-Clorofórmio	322,16	1,94	2
147	Fenol-Clorofórmio	542,6	1,65	1,46
148	Fenol-Clorofórmio	605,07	1,84	1,87
149	Kit De Extração	347,19	1,87	2,03
150	Kit De Extração	242,31	1,88	1,92
151	Kit De Extração	161,48	1,59	1,53
152	Kit De Extração	275,5	1,89	1,78
153	Kit De Extração	92,91	1,9	1,94
154	Kit De Extração	215,75	1,88	0,98
155	Kit De Extração	178,33	1,73	1,57
156	Kit De Extração	298,21	1,75	1,7
157	Kit De Extração	1142,32	1,88	1,93
158	Kit De Extração	918,28	1,92	2,14
159	Kit De Extração	910,65	1,91	2,14
160	Kit De Extração	374,34	1,8	1,92
161	Kit De Extração	381,26	1,78	1,91
162	Kit De Extração	168,02	1,89	0,88
163	Kit De Extração	58,08	1,71	1,12
164	Kit De Extração	229,67	1,85	1,81
165	Kit De Extração	52,48	1,87	1,21
166	Kit De Extração	21,76	1,76	0,7
167	Kit De Extração	145,19	1,85	1,47
168	Kit De Extração	118,49	1,9	1,72
169	Fenol-Clorofórmio	1220,1	1,92	2,08
170	Kit De Extração	304,46	1,78	1,44
171	Kit De Extração	88,11	1,79	0,89
172	Kit De Extração	186,27	1,75	1,14
173	Kit De Extração	113,56	1,81	1,06
174	Kit De Extração	646	1,85	1,63
175	Kit De Extração	101,27	1,81	0,9
176	Kit De Extração	498,65	1,84	1,92
177	Kit De Extração	209,46	1,85	1,8
178	Kit De Extração	270,17	1,82	1,24
179	Kit De Extração	235,59	1,86	1,84
180	Kit De Extração	520,64	1,81	1,92
181	Kit De Extração	360,89	1,82	1,65
182	Kit De Extração	122,3	1,88	1,52
183	Kit De Extração	14,98	1,84	0,44
184	Kit De Extração	16,29	1,7	0,78
185	Kit De Extração	12,02	1,54	0,35
186	Kit De Extração	17,95	1,54	0,35
187	Kit De Extração	13,78	1,59	0,33
188	Kit De Extração	17,04	1,67	0,31
189	Kit De Extração	15,43	1,59	0,74
190	Kit De Extração	66,62	1,36	0,25

APÊNDICE D: VALORES DE FLUORESCÊNCIA (ΔRN) OBTIDOS PELA *QPCR SNP GENOTYPING* PARA O SNP RS1800592 DO GENE *UCP1*, DISCRIMINANDO OS SINAIS DOS ALELOS G (ALELO 1) E A (ALELO 2) E OS PARÂMETROS DE QUALIDADE DA REAÇÃO EM 190 AMOSTRAS ANALISADAS.

Amostra	Alelo 1 (G) ΔRn	Alelo 2 (A) ΔRn	Referência Passiva	Qualidade (%)
01	0,699971	3,270536	320342,4	99,548523
02	0,657517	3,570789	321656,1	99,533401
03	0,581859	3,003681	316919,3	98,737244
04	0,617611	3,705499	337400,6	99,266151
05	0,71843	4,919535	372943,8	99,445442
06	0,56243	2,050976	324569,3	97,711243
07	0,876616	0,845062	334157,6	61,466969
08	1,891876	2,063389	326520,8	98,76203
09	0,88209	4,664988	331679,1	99,9726
10	0,796809	4,896676	331237,3	99,753212
11	1,800012	2,425587	344648	98,58551
12	2,607021	0,591592	356308,1	99,292801
13	1,783557	2,511924	342474,7	98,454086
14	0,799383	4,102865	318095,2	99,77193
15	0,872367	4,776809	344604,1	99,91123
16	1,997333	2,899069	296529,7	99,392502
17	2,104046	2,772575	304571,9	99,764343
18	2,543995	3,462559	365498,1	99,84951
19	0,342905	0,428867	339646,3	13,699243
20	0,714229	3,836185	346542,7	99,1834
21	0,674697	3,811929	354457,3	98,85799
22	1,980071	2,648873	368857,7	99,6461
23	1,956466	2,630501	377305,7	99,5797
24	0,61113	3,331463	307564,4	99,144783
25	2,200724	3,080256	358200,9	99,73535
26	2,150347	3,128866	323483	99,38068
27	0,904708	4,377745	310704	99,96124
28	3,01923	0,37437	336848,1	99,99075
29	0,78852	3,693334	314835,4	99,32111
30	0,769358	3,916805	311501,4	99,942451
31	0,599359	3,434626	316690,8	99,068344
32	0,578392	3,208707	336303,5	98,791618
33	2,122194	0,253321	296480,7	98,83268
34	2,147812	2,727636	311561,3	99,91624
35	0,643797	3,977739	338850,9	99,456612
36	1,666277	2,175846	322639,1	98,87885
37	2,932166	0,385402	349396,7	99,93981
38	1,296047	1,280761	304945,5	97,01542
39	2,244427	3,183224	326635,2	99,918884
40	2,02672	2,94058	348606,9	99,19262
41	0,877781	4,454775	331048,5	99,99712
42	0,931807	4,605995	293096,2	99,95702
43	0,917378	4,43582	308036,8	99,96116
44	0,89528	4,776659	343172,5	99,92779
45	3,352479	0,447142	405280,4	99,70455
46	0,726971	3,195904	347138,9	98,22608
47	0,811009	3,808259	351253,3	99,50939
48	1,001011	5,118737	338590,2	99,54563
49	0,895798	4,633635	334804,4	99,98431
50	0,807172	4,561603	353879,3	99,86729
51	0,905596	5,050917	306558,2	99,70495
52	2,090313	2,430909	337379,6	99,4829
53	3,383978	0,408648	335506,3	99,64005
54	1,152892	5,627896	317259,5	98,27625

55	2,51144	3,290018	336552,8	99,92876
56	1,179467	5,391505	327216,7	98,44855
57	1,06328	4,68257	345940,9	99,4488
58	2,492818	3,185476	231145,9	99,91634
59	0,953335	4,768848	341363,1	99,88563
60	0,892661	4,583407	332179,2	99,99414
61	0,895272	4,481205	233286,3	99,99501
62	2,393912	3,248401	353549,2	99,93784
63	0,817062	4,284518	325030,5	99,89693
64	0,7499	3,681735	358385,2	99,19523
65	1,519211	1,519791	375731,5	97,39332
66	0,936499	4,6108	369771,2	99,94877
67	0,965931	4,941366	352582,7	99,77041
68	1,010016	4,971087	410755,1	99,64253
69	1,162864	4,969167	267130,1	98,79079
70	2,170136	3,010121	322669,7	99,914619
71	3,041225	0,319941	344907,1	99,96374
72	2,521069	3,342871	403762,2	99,91687
73	2,378473	3,1344	404185,3	99,99228
74	0,789093	4,134178	348995,6	99,76057
75	0,786733	4,220594	370190,9	99,79053
76	2,253825	2,938442	384030,4	99,98742
77	2,463476	3,3398	365025,1	99,91604
78	2,387798	3,07908	382811,3	99,98497
79	3,132765	0,43409	332448,4	99,978
80	2,547376	3,430982	345171,1	99,87223
81	0,904023	4,614763	306545	99,871201
82	3,062792	0,417308	307151,5	99,99784
83	1,023607	4,577835	393330,8	99,6371
84	0,971574	4,480555	326778	99,8305
85	2,326856	3,008242	366055,2	99,99669
86	0,841298	4,356392	367914,8	99,95743
87	0,749225	4,143229	341352,2	99,5993
88	2,232276	2,902091	374144,9	99,97957
89	0,792943	4,14385	327325,4	99,77745
90	2,567703	3,213904	226087,1	99,78886
91	2,619856	3,3491	406527,4	99,7901
92	2,837209	3,763023	448417,9	99,45155
93	1,065096	5,785441	319591,8	98,28873
94	2,841432	3,797148	368200,7	99,43468
95	2,841018	3,747242	361669,7	99,4463
96	2,45599	3,23331	367791,3	99,96463
97	2,557153	3,460426	364112,2	99,85284
98	0,870252	4,646889	362584,2	99,97045
99	0,776757	4,278976	359671,7	99,76898
100	0,824941	4,319358	283504,1	99,92252
101	2,307615	3,164423	391788,4	99,89866
102	2,511998	3,273604	311603,2	99,92664
103	2,227783	3,156858	271903,8	99,916054
104	2,470902	3,154559	341696,7	99,92982
105	2,287745	3,102049	401470,1	99,93311
106	2,303474	2,938825	382320,5	99,98067
107	0,953444	4,790747	344519,5	99,87701
108	0,92703	4,415844	332826,2	99,936
109	0,98802	5,498841	327171,6	99,274269
110	0,938107	5,254843	368982,5	99,552635
111	3,24082	0,439632	365264,3	99,88594
112	0,961153	4,516368	235677,1	99,8782
113	2,343133	3,168123	349968,9	99,95157
114	2,286748	3,063449	332504,9	99,9652
115	0,90556	4,44055	329277,5	99,85434

116	0,793862	4,413066	315319,4	99,84626
117	0,79683	4,627241	424449,5	99,80621
118	2,534437	3,233523	373296,5	99,87846
119	2,765845	3,750104	375368,5	99,55784
120	2,727391	3,708539	340297,6	99,61808
121	2,155003	2,653035	335759	99,84071
122	0,767942	4,423642	364699	99,7385
123	0,978282	4,990755	349470,5	99,593521
124	3,232387	0,386826	337502,8	99,797401
125	2,321729	3,115388	333070,6	99,989105
126	2,335322	3,204738	360755,2	99,973503
127	2,925638	0,263856	353566,1	99,931404
128	2,348924	3,339829	301357,4	99,864655
129	0,858829	4,496839	345238,7	99,961105
130	2,302843	3,183485	331122,8	99,979767
131	0,93428	4,699312	315152,5	99,782669
132	3,042811	0,50082	319368,6	99,881653
133	0,965919	4,680892	302280,2	99,663071
134	0,977848	4,483356	335439,6	99,539421
135	0,955923	4,375009	226892,1	99,605408
136	2,492409	3,568425	335831,5	99,542236
137	0,740777	3,942637	321215,9	99,911682
138	2,178292	2,826238	230936,5	99,847931
139	2,801663	0,280982	352721,3	99,942116
140	2,224487	2,961909	325294,4	99,964233
141	2,201213	2,892234	360733,3	99,915146
142	2,504997	3,192868	386849,5	99,643814
143	3,42292	0,470086	341226,4	99,483238
144	1,014946	5,324167	323117,1	99,331192
145	2,670188	3,735921	351808,6	98,98407
146	2,743935	3,609176	246022,3	98,823624
147	0,972536	4,713463	318883,9	99,63797
148	2,523909	3,439241	392965,4	99,646011
149	2,319153	3,164592	380613,3	99,988678
150	0,76728	4,209531	356683,6	99,967041
151	2,129356	2,919595	381558,5	99,861443
152	2,369454	3,186241	382667,4	99,956345
153	2,288747	3,138312	348728,6	99,993866
154	0,920765	4,479656	364667,4	99,809639
155	2,331757	3,114099	313171,3	99,981133
156	0,94581	4,68726	336950,5	99,743721
157	2,318714	3,163029	289089,2	99,989113
158	0,935849	4,269722	368588,1	99,651909
159	2,325626	3,275467	316416,3	99,921387
160	0,949292	4,409415	350545,6	99,659363
161	0,859063	4,219355	370801	99,932602
162	2,699287	0,273789	340689	99,882843
163	2,00723	2,7263	363764,5	99,546822
164	2,37922	3,021777	396895,4	99,817162
165	2,742339	2,778124	244394,4	90,696106
166	3,365115	0,342315	422081,7	99,577576
167	2,60994	3,44933	440223,4	99,400444
168	0,980169	5,450295	339134,3	99,331161
169	0,91377	4,83288	387948	99,811035
170	3,349075	0,477551	380623,3	99,608582
171	2,449062	3,240223	380627,2	99,836639
172	0,895439	4,620662	370028,1	99,890068
173	2,832196	0,252131	355197,7	99,914459
174	2,326229	3,142147	359925	99,988899
175	0,812017	4,245151	288582,9	99,99913
176	2,979187	0,346377	379010,1	99,985435

177	0,856604	4,210773	300529,8	99,935966
178	2,350994	2,981758	317575,6	99,828789
179	2,402105	3,298867	368660,3	99,898758
180	0,965327	4,675313	360316,4	99,664894
181	2,289571	3,066127	321046,9	99,994881
182	2,408243	3,202969	303235,2	99,905792
183	2,803946	0,466195	342279,2	99,894569
184	2,061935	2,72969	221594,5	99,673286
185	2,494266	0,221681	322694,3	99,591957
186	0,675517	4,391352	380527,3	99,546295
187	1,942792	2,436172	322541,6	99,040955
188	0,762603	4,496655	379156,2	99,887657
189	0,801643	4,439702	418120,4	99,977203
190	0,59578	2,529002	383288,2	98,505226
***NTC – 1ª Placa**	0,479623	0,147943	231604,8	100
***NTC – 2ª Placa**	0,410711	0,067106	199213,5	100

*NTC = controle negativo (*negative template control*, NTC)

APÊNDICE E: DETALHAMENTO DOS CÁLCULOS DE TEMPO E CUSTO E APLICAÇÃO DAS EQUAÇÕES EM EXEMPLOS TEÓRICOS

➢ Equações para os cálculos de tempo:

Para uma **batelada** de 96 amostras:

$$TEM_{lote(N)} = \sum_{k=1}^{m} t_k^{manual}$$

$$TOI_{lote(N)} = \sum_{k=1}^{m} t_k^{instrumental}$$

$$TTP_{lote(N)} = TEM_{lote(N)} + TOI_{lote(N)}$$

Por **amostra individual** (considerando uma batelada de 96 amostras):

$$TEM_{por\ amostra\,(N)} = \frac{1}{N} \sum_{k=1}^{m} t_k^{manual}$$

$$TOI_{por\ amostra\,(N)} = \frac{1}{N} \sum_{k=1}^{m} t_k^{instrumental}$$

$$TTP_{por\ amostra(N)} = TEM_{por\ amostra(N)} + TOI_{por\ amostra(N)}$$

➢ Elementos das equações:

k

- Índice que representa cada etapa do fluxo de trabalho
- Vai de 1 até m
- Exemplo no seu estudo:
  - k_1 = preparo da PCR
  - k_2 = reação no termociclador
  - k_3 = medição no FTIR
  - k_4 = processamento em ML

m

- Número total de etapas experimentais do fluxo
- Exemplo:
  - Sequenciamento → m = 4 etapas
  - qPCR → m = 3 etapas
  - ATR-FTIR associada ao ML → m = 4 etapas

t_k^{manual}

- Tempo manual da etapa k
- Representa o tempo em que o operador está efetivamente trabalhando

Exemplo real:

- Preparação da qPCR: 150 min
- Aplicar amostra no ATR: 1,3 min por amostra

t_k^{total}

- Tempo total decorrido da etapa k
- Inclui tempo de equipamento funcionando sem operador

Exemplo real:

- Reação no termociclador: 120 min
- Reação do sequenciador: 960 min

Σ (somatório)

- Indica a soma de todas as etapas do fluxo
- Exemplo:
  - $TEM_{lote(N)}$ = preparo + interpretação + etc...

N

- Número total de amostras do lote
- Neste estudo: N = 96

➢ **Exemplo (Sequenciamento NGS):**

$$TEM_{lote(N)} = \sum_{k=1}^{m} t_k^{manual}$$

$$TEM_{por\ amostra\,(N)} = \frac{1}{N} \sum_{k=1}^{m} t_k^{manual}$$

- ***N*** = 96 amostras por lote

- **Considere *m* = 3 etapas com trabalho manual relevante:**
  k = 1: Preparação de biblioteca (library prep)
  k = 2: Limpeza/pipetagem
  k = 3: Setup/carregamento do sequenciador

- **Atribuindo os tempos manuais:**
  t_1^{manual} = 120 min
  (pipetagem, preparo de mixes, indexação/adaptadores, organização das bibliotecas)
  t_2^{manual} = 60 min
  (clean-up com beads/colunas, normalização e pooling manual)
  t_3^{manual} = 30 min
  (preparo final e carregamento no equipamento)

$$TEM_{lote(N)} = \sum_{k=1}^{3} t_k^{manual} = t_1^{manual} + t_2^{manual} + t_3^{manual}$$

$$TEM_{lote(96)} = 120 + 60 + 30 = 210\text{min.}$$

$$TEM_{por\ amostra(N)} = \frac{1}{\mathrm{N}} \sum_{k=1}^{m} t_k^{manual}$$

$$TEM_{por\ amostra(1)} = \frac{1}{96} . 210 = 2{,}19\text{min./amostra}$$

- **Resumo de TEM (para sequenciamento NGS):**

- $N = 96$
- $m = 3$ etapas manuais
- $(t_1,t_2,t_3) = (120,60,30)$ min.
- $TEM_{lote(96)} = 210$ min
- $TEM_{por\ amostra(1)} = \frac{210min.}{96\ \text{amostras}} = 2{,}19$ min./amostra

Equações para os cálculos de custo:

Para uma **batelada** de 96 amostras:

$$Custo_{lote(N)} = \sum_{k=1}^{m} \sum_{j=1}^{p_k} q_{k,j} \cdot P_{k,j}$$

Por **amostra individual** (considerando uma batelada de 96 amostras):

$$Custo_{por\ amostra(N)} = \frac{1}{\mathrm{N}} \sum_{k=1}^{m} \sum_{j=1}^{p_k} q_{k,j} \cdot P_{k,j}$$

- **Elementos das equações:**

k

- Índice que representa cada etapa do fluxo de trabalho
- Vai de 1 até m
- Exemplo no seu estudo:
  - k_1 = preparo da PCR
  - k_2 = reação no termociclador
  - k_3 = medição no FTIR
  - k_4 = processamento em ML

m

- Número total de etapas experimentais do fluxo
- Exemplo:
  - Sequenciamento → m = 4 etapas
  - qPCR → m = 3 etapas
  - ATR-FTIR associada ao ML → m = 4 etapas

P_k

- Número de reagentes ou consumíveis utilizados na etapa k.
- Exemplo:
  Se na etapa "library prep" foi usado:
  - kit library prep.
  - adaptadores indexados
- Então:
- $P_1 = 2$

j

- Índice que representa cada reagente dentro de uma etapa.
- Exemplo:
  - Na etapa PCR (k = 1):
    - j = 1 → Taq DNA polimerase
    - j = 2 → dNTPs
    - j = 3 → primers
    - j = 4 → água

$q_{k,j}$

- Quantidade total do reagente j utilizada na etapa k para o lote inteiro.
- Pode ser:
  - número de reações usadas
  - número de microlitros consumidos
  - número de kits utilizados
- Exemplos:
  - 1 kit de library prep para 96 amostras
    → $q_{1,1} = 1$
  - 96 reações usando mastermix
    → $q_{1,1}$ = volume total consumido (ex: 960 μL)

$P_{k,j}$

- Preço unitário do reagente j na etapa k (em USD).
- Exemplos:
  - Library prep kit: 2000USD
  - Sequencing kit: 3000USD

Σ (somatório)

- Indica que estamos somando todas as etapas do fluxo
- Exemplo:
    - TEM_{lote} = preparo + interpretação + etc...

N

- Número total de amostras do lote
- Neste estudo: N = 96

➢ **Exemplo (Sequenciamento NGS):**

$$Custo_{lote(N)} = \sum_{k=1}^{m} \sum_{j=1}^{p_k} q_{k,j} \cdot P_{k,j}$$

$$Custo_{por\ amostra(N)} = \frac{1}{\mathrm{N}} \sum_{k=1}^{m} \sum_{j=1}^{p_k} q_{k,j} \cdot P_{k,j}$$

- **N** = 96

- ***m* (número de etapas)**

  Para NGS:
  - ***k* = 1**: preparo de biblioteca (library preparation)
  - ***k* = 2**: sequenciamento (reagentes do run)

  Logo:
  m = 2

- **Etapa k = 1 (montagem da biblioteca)**
  Itens (reagentes/consumíveis) usados nessa etapa:
  - ***j* = 1:** Library prep kit (enzimas + buffers)
  - ***j* = 2:** Index/adapters kit
  - ***j* = 3:** Cleanup beads / purification
  - ***j* = 4:** Consumíveis plásticos (pontas, tubos, placas)

  Logo:
  $\mathbf{P_1}$ = 4

- **Etapa k = 2 (reação de sequenciamento)**
  Itens usados no run:
  - ***j* = 1:** Kit de sequenciamento / flow cell (1 run)
  - ***j* = 2:** Consumíveis auxiliares do run

  Logo:
  $\mathbf{P_2}$ = 2

- **Atribuindo valores de $q_{k,j}$ e $P_{k,j}$**

Etapa *k* = 1 (montagem das bibliotecas)

***j* = 1 (library prep kit)**

- Preço do kit: 3600 USD para 96 reações

$$q_{1,1} = 1 \qquad P_{1,1} = 3600$$

***j* = 2 (index/adapters kit)**

- Preço do kit: 700 USD para 96 reações

$$q_{1,2} = 1 \qquad P_{1,2} = 700$$

***j* = 3 (cleanup beads)**

- Preço: 350 USD por frasco (rende 200 reações)
- Para 96 reações, usamos uma fração do frasco:

$$q_{1,3} = \frac{96}{200} = 0{,}48 \qquad P_{1,3} = 350$$

***j* = 4 (consumíveis)**

- Estimativa: 200 USD por lote (ponteiras, tubos, placas)

$$q_{1,4} = 1 \qquad P_{1,2} = 200$$

Etapa *k* = 2 (reação de sequenciamento)

***j* = 1 (sequencing kit/run)**

- 1 run: 2200 USD

$$q_{2,1} = 1 \qquad P_{2,1} = 2200$$

***j* = 2 (consumíveis da reação)**

- Estimativa: 140 USD por lote

$$q_{2,2} = 1 \qquad P_{2,2} = 140$$

- **Aplicando os valores nas equações:**

Etapa *k* = 1 (montagem das bibliotecas)

$$Custo_{k=1} = (q_{1,1}P_{1,1}) + (q_{1,2}P_{1,2}) + (q_{1,3}P_{1,3}) + (q_{1,4}P_{1,4})$$

$$Custo_{k=1} = (1 \,.\, 3600) + (1 \,.\, 700) + (0{,}48 \,.\, 350) + (1 \,.\, 200)$$

$$Custo_{k=1} = 3600 + 700 + 168 + 200$$

$$Custo_{k=1} = 4668 \text{ USD}$$

Etapa *k* = 2 (reação de sequenciamento)

$$Custo_{k=2} = (q_{2,1}P_{2,1}) + (q_{2,2}P_{2,2})$$

$$Custo_{k=2} = (1 \,.\, 2200) + (1 \,.\, 140)$$

$$Custo_{k=2} = 2340 \text{ USD}$$

Custo total do lote (96 amostras)

$$Custo_{lote(N)} = Custo_{k=1} + Custo_{k=2}$$

$$Custo_{lote(96)} = 4668 + 2340$$

$$Custo_{lote(96)} = 7008 \text{ USD}$$

Custo total por amostra:

$$Custo_{por\ amostra(N)} = \frac{1}{N} \,.\, Custo_{lote(N)}$$

$$Custo_{por\ amostra(1)} = \frac{1}{96} \,.\, 7008$$

$$Custo_{por\ amostra(1)} = 73 \text{ USD}$$

APÊNDICE F: MEDIDAS DE DESEMPENHO DOS MODELOS DE *MACHINE LEARNING* ORGANIZADAS EM ORDEM DECRESCENTE DE ACORDO COM A FAIXA ESPECTRAL (*RANGE*)

AUC*	Acurácia	Sensibilidade	Especificidade	F1-score	Pares de genótipos	*Range* (cm^{-1})	Modelos de ML
0.654722	0.716364	0.76	0.706667	0.508704	AA_vs_GG	2800-3800	**DL-MLP Res. Sim.**
0.6475	0.686364	0.785	0.664444	0.494817	AA_vs_GG	2800-3800	**DL-MLP Funnel**
0.6425	0.697273	0.765	0.682222	0.49526	AA_vs_GG	2800-3800	**Log. Reg.**
0.58575	0.630476	0.76	0.59	0.487177	AG_vs_GG	2800-3800	**SVM Linear**
0.5665	0.639048	0.704	0.61875	0.477682	AG_vs_GG	2800-3800	**Log. Reg.**
0.565903	0.627059	0.62875	0.625556	0.581842	AA_vs_AG	2800-3800	**DL-MLP Res. Sim.**
0.565069	0.627647	0.61625	0.637778	0.577681	AA_vs_AG	2800-3800	**DL-MLP Funnel**
0.544444	0.613529	0.6125	0.614444	0.565471	AA_vs_AG	2800-3800	**Log. Reg.**
0.52725	0.650476	0.64	0.65375	0.444313	AG_vs_GG	2800-3800	**SVM RBF Kernel**
0.52525	0.592381	0.732	0.54875	0.457474	AG_vs_GG	2800-3800	**DL-MLP Funnel**
0.521111	0.665455	0.635	0.672222	0.412279	AA_vs_GG	2800-3800	**LDA**
0.483333	0.635455	0.645	0.633333	0.379914	AA_vs_GG	2800-3800	**SVM RBF Kernel**
0.465	0.612727	0.66	0.602222	0.3911	AA_vs_GG	2800-3800	**SVM Linear**
0.457292	0.573529	0.58125	0.566667	0.510769	AA_vs_AG	2800-3800	**LDA**
0.447604	0.568824	0.48375	0.644444	0.440432	AA_vs_AG	2800-3800	**SVM Linear**
0.436389	0.56	0.52875	0.587778	0.463785	AA_vs_AG	2800-3800	**SVM RBF Kernel**
0.399	0.554286	0.648	0.525	0.364619	AG_vs_GG	2800-3800	**LDA**
0.644444	0.72	0.725	0.718889	0.507751	AA_vs_GG	950-1500	**Log. Reg.**
0.59675	0.687619	0.688	0.6875	0.502781	AG_vs_GG	950-1500	**DL-MLP Funnel**
0.5965	0.604762	0.832	0.53375	0.51334	AG_vs_GG	950-1500	**SVM Linear**
0.588889	0.687273	0.675	0.69	0.45624	AA_vs_GG	950-1500	**DL-MLP Res. Sim.**
0.579444	0.663636	0.705	0.654444	0.454914	AA_vs_GG	950-1500	**DL-MLP Funnel**
0.57375	0.658095	0.7	0.645	0.496485	AG_vs_GG	950-1500	**Log. Reg.**
0.541181	0.614118	0.565	0.657778	0.54529	AA_vs_AG	950-1500	**Log. Reg.**
0.535486	0.618824	0.5625	0.668889	0.541154	AA_vs_AG	950-1500	**DL-MLP Res. Sim.**
0.53075	0.589524	0.764	0.535	0.469291	AG_vs_GG	950-1500	**SVM RBF Kernel**
0.53025	0.66381	0.608	0.68125	0.447124	AG_vs_GG	950-1500	**DL-MLP Res. Sim.**
0.522431	0.610588	0.51	0.7	0.517594	AA_vs_AG	950-1500	**DL-MLP Funnel**
0.521667	0.578182	0.765	0.536667	0.397879	AA_vs_GG	950-1500	**LDA**

0.514722	0.618182	0.735	0.592222	0.41084	AA_vs_GG	950-1500	**SVM RBF Kernel**
0.475139	0.587647	0.59	0.585556	0.523712	AA_vs_AG	950-1500	**LDA**
0.439479	0.561176	0.48375	0.63	0.435869	AA_vs_AG	950-1500	**SVM RBF Kernel**
0.435139	0.563529	0.4525	0.662222	0.42069	AA_vs_AG	950-1500	**SVM Linear**
0.418056	0.640909	0.565	0.657778	0.343427	AA_vs_GG	950-1500	**SVM Linear**
0.634722	0.696364	0.755	0.683333	NA	AA_vs_GG	900-1100	**Log. Reg.**
0.563056	0.653636	0.705	0.642222	NA	AA_vs_GG	900-1100	**DL-MLP Funnel**
0.55375	0.642857	0.7	0.625	NA	AG_vs_GG	900-1100	**Log. Reg.**
0.547778	0.637273	0.71	0.621111	NA	AA_vs_GG	900-1100	**DL-MLP Res. Sim.**
0.542222	0.611765	0.61	0.613333	NA	AA_vs_AG	900-1100	**DL-MLP Res. Sim.**
0.538681	0.608824	0.61	0.607778	NA	AA_vs_AG	900-1100	**Log. Reg.**
0.5375	0.531818	0.87	0.456667	NA	AA_vs_GG	900-1100	**LDA**
0.536111	0.656364	0.695	0.647778	NA	AA_vs_GG	900-1100	**SVM RBF Kernel**
0.535	0.619048	0.716	0.58875	NA	AG_vs_GG	900-1100	**DL-MLP Res. Sim.**
0.526042	0.605882	0.55125	0.654444	NA	AA_vs_AG	900-1100	**DL-MLP Funnel**
0.5225	0.613333	0.696	0.5875	NA	AG_vs_GG	900-1100	**DL-MLP Funnel**
0.501875	0.59619	0.692	0.56625	NA	AG_vs_GG	900-1100	**LDA**
0.501389	0.646364	0.67	0.641111	NA	AA_vs_GG	900-1100	**SVM Linear**
0.49375	0.548571	0.788	0.47375	NA	AG_vs_GG	900-1100	**SVM Linear**
0.478	0.614286	0.632	0.60875	NA	AG_vs_GG	900-1100	**SVM RBF Kernel**
0.433299	0.557647	0.575	0.542222	NA	AA_vs_AG	900-1100	**SVM Linear**

* AUC = área sob a curva (*area under the curve*, AUC)

APÊNDICE G: PRINCIPAIS NÚMEROS DE ONDA COM MAIOR PODER DISCRIMINATIVO AVALIADOS PELOS VALORES DE ÁREA SOB A CURVA (AUC) EM ORDEM DECRESCENTE DE NÚMERO DE ONDA (CM^{-1})

Pares de genótipos	Modelos	AUC	Número de onda (cm^{-1})
AA vs GG	DL-MLP Res. Sim.	0.6508	**3583.56**
AA vs GG	DL-MLP Res. Sim.	0.6508	**3564.0**
AA vs AG	DL-MLP Res. Sim.	0.5697	**3521.44**
AA vs GG	DL-MLP Res. Sim.	0.6508	**3520.29**
AA vs AG	DL-MLP Res. Sim.	0.5697	**3516.84**
AA vs AG	DL-MLP Res. Sim.	0.5697	**3511.09**
AA vs GG	DL-MLP Res. Sim.	0.6508	**3498.43**
AA vs GG	DL-MLP Res. Sim.	0.6508	**3367.29**
AA vs AG	DL-MLP Res. Sim.	0.5697	**2909.44**
AA vs AG	DL-MLP Res. Sim.	0.5697	**2904.83**
AG vs GG	DL-MLP Funnel	0.5752	**1466.86**
AG vs GG	DL-MLP Funnel	0.5752	**1211.47**
AG vs GG	DL-MLP Funnel	0.5752	**1209.17**
AG vs GG	DL-MLP Funnel	0.5752	**1208.02**
AG vs GG	DL-MLP Funnel	0.5752	**1202.27**
AA vs AG	DL-MLP Res. Sim.	0.5289	**1017.06**
AA vs GG	Log. Reg.	0.6347	**1094.13**
AA vs GG	Log. Reg.	0.6347	**1092.98**
AA vs GG	Log. Reg.	0.6347	**1069.97**
AA vs GG	Log. Reg.	0.6347	**1068.82**
AG vs GG	Log. Reg.	0.5538	**1029.71**
AG vs GG	Log. Reg.	0.5538	**1012.45**
AG vs GG	Log. Reg.	0.5538	**1011.3**

AG vs GG	Log. Reg.	0.5538	**1010.15**
AA vs AG	DL-MLP Res. Sim.	0.5289	**996.35**
AA vs GG	Log. Reg.	0.6347	**994.05**
AG vs GG	Log. Reg.	0.5538	**989.45**
AA vs AG	DL-MLP Res. Sim.	0.5289	**983.69**
AA vs AG	DL-MLP Res. Sim.	0.5289	**973.34**
AA vs AG	DL-MLP Res. Sim.	0.5289	**971.04**

ANEXO A: PARECER DO COMITÊ DE ÉTICA EM PESQUISA COM SERES HUMANOS – CEP

UNIVERSIDADE DO VALE DO PARAÍBA - UNIVAP

PARECER CONSUBSTANCIADO DO CEP

DADOS DA EMENDA

Título da Pesquisa: Relação do polimorfismo -866G/A do gene UCP2 com o desenvolvimento de distúrbios multifatoriais na população brasileira

Pesquisador: Renata de Azevedo Canevari

Área Temática: Genética Humana:
(Trata-se de pesquisa envolvendo Genética Humana que não necessita de análise ética por parte da CONEP;);

Versão: 5

CAAE: 62978022.0.0000.5503

Instituição Proponente: Universidade do Vale do Paraíba - UNIVAP

Patrocinador Principal: FUNDACAO DE AMPARO A PESQUISA DO ESTADO DE SAO PAULO

DADOS DO PARECER

Número do Parecer: 8.345.826

Apresentação do Projeto:

Trata-se de emenda a projeto aprovado por este CEP em 13/12/2022, parecer n.: 5.810.129.

Em decorrência da aprovação do projeto pela Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), sob o processo nº 2023/16314-6, e da consequente ampliação da infraestrutura, dos recursos financeiros e da capacidade técnico-operacional para a condução da pesquisa, submeto o presente pedido de emenda ao protocolo originalmente aprovado. A referida emenda fundamenta-se na possibilidade de expansão do número de participantes, com aumento do total de amostras inicialmente previstas (n=100) para 300. Essa ampliação permitirá maior robustez estatística, aumento do poder analítico e maior representatividade dos resultados, fortalecendo a validade científica do estudo. Adicionalmente, solicita-se a inclusão de dois novos membros na equipe executora, os quais atuarão nas etapas de análises moleculares, processamento e análise dos dados, contribuindo para a adequada execução metodológica e para o cumprimento do cronograma estabelecido. Encaminho, em anexo a esta plataforma, carta formal direcionada ao Comitê de Ética em Pesquisa, contendo a justificativa detalhada das alterações propostas nesta emenda.

Situação do Parecer:
Aprovado

Necessita Apreciação da CONEP:
Não

SAO JOSE DOS CAMPOS, 09 de Abril de 2026

Assinado por:
Mauricio Martins Alves
(Coordenador(a))