



UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE CIÊNCIAS BIOLÓGICAS
PROGRAMA DE PÓS-GRADUAÇÃO EM GENÉTICA E BIOLOGIA MOLECULAR

CARACTERIZAÇÃO DE VARIANTES DO GENE *CFTR* EM EXOMAS DE
POPULAÇÕES INDÍGENAS E AVALIAÇÃO DE DESEMPENHO DE PREDITORES
DE PATOGENICIDADE

ANA KATARINA CAMPOS NUNES

BELÉM
2025



UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE CIÊNCIAS BIOLÓGICAS
PROGRAMA DE PÓS-GRADUAÇÃO EM GENÉTICA E BIOLOGIA MOLECULAR

CARACTERIZAÇÃO DE VARIANTES DO GENE *CFTR* EM EXOMAS DE
POPULAÇÕES INDÍGENAS E AVALIAÇÃO DE DESEMPENHO DE PREDITORES
DE PATOGENICIDADE

Dissertação de Mestrado submetido ao Programa de Pós-Graduação em Genética e Biologia Molecular (PPGBM/UFPA), como requisito parcial para obtenção do grau de Mestre em Genética e Biologia Molecular.

Orientador: Prof. Dr. Gilderlanio Santana de Araújo

BELÉM

2025

Dados Internacionais de Catalogação na Publicação (CIP) de acordo com ISBD
Sistema de Bibliotecas da Universidade Federal do Pará
Gerada automaticamente pelo módulo Ficat, mediante os dados fornecidos pelo(a) autor(a)

C198c Campos Nunes, Ana Katarina.
CARACTERIZAÇÃO DE VARIANTES DO GENE CFTR
EM EXOMAS DE POPULAÇÕES INDÍGENAS E AVALIAÇÃO
DE DESEMPENHO DE PREDITORES DE PATOGENICIDADE
/ Ana Katarina Campos Nunes. — 2025.
89 f. : il. color.

Orientador(a): Prof. Dr. Gilderlanio Santana de Araújo
Dissertação (Mestrado) - Universidade Federal do Pará,
Instituto de Ciências Biológicas, Programa de Pós-Graduação em
Genética e Biologia Molecular, Belém, 2025.

1. População indígena. 2. Gene CFTR. 3. Variabilidade
Genética. 4. Preditores. I. Título.

CDD 599.935

*"Não pode chover o tempo todo.
O céu não pode cair para sempre.
E embora a noite pareça longa, suas lágrimas não
podem cair para sempre."*

(O Corvo – Edgar Allan Poe)

AGRADECIMENTOS

Agradeço, primeiramente, a Deus e a toda a espiritualidade que me rege e me guia, por me acolher e não me deixar perecer.

À minha mãe Zana, ao meu irmão Arthur, às minhas tias Conceição e Graciete, à minha avó Dona Graça e aos meus tios e tias de Manaus: meu sincero agradecimento pelo apoio constante, desde os meus primeiros passos na ciência.

Retomando de forma especial à minha mãe e ao meu irmão: vocês são a minha fortaleza e o meu refúgio. Tudo isso é por vocês! Mãe, um dia espero poder retribuir todo o seu esforço e ser o seu maior orgulho. Arthur, meu irmão, desejo estar sempre ao seu lado para te guiar e apoiar em cada etapa da sua caminhada. Amo vocês imensamente!

A toda a minha família, mesmo àqueles que estão fisicamente distantes, obrigada pela motivação, pelas orações e pelo suporte incondicional.

Aos meus filhotes felinos, Meleys e meu anjo Alberth: meus pedacinhos do céu!

Aos meus amigos Layna, Elson e Fábio: são poucos, mas são para sempre. Obrigada por suportarem minha ausência em longos períodos, por estarem ao meu lado nos momentos de ansiedade e frustração, e por me ensinarem, sobretudo, a enxergar o melhor em mim. Por acreditarem em mim, especialmente quando eu mesma não conseguia. Vocês sempre terão a minha eterna gratidão.

Agradeço também à Dra. Ândrea Ribeiro-dos-Santos, pela confiança em meu trabalho e na oportunidade de integrar o LGHM.

Ao meu orientador, Dr. Gilderlanio Santana de Araújo, minha profunda gratidão por me guiar nesta jornada do mestrado, pelos ensinamentos transmitidos tanto no ambiente acadêmico quanto na vida.

Aos colegas do laboratório, obrigada por tornarem o ambiente de trabalho mais leve, acolhedor e saudável. Compartilhar o dia a dia com vocês foi um privilégio.

Ao meu querido amigo Arthur Reis, que se tornou um grande professor e inspiração para mim, meu muito obrigada.

Aos colegas da equipe de Bioinformática, sou grata pelo apoio, pelos ensinamentos e pelo companheirismo ao longo dessa jornada.

E à equipe do Genoma SUS, minhas queridas companheiras, obrigada pela paciência ao ensinar, pela generosidade e, acima de tudo, por me acolherem com tanta empatia e compreensão.

SUMÁRIO

1. INTRODUÇÃO	9
1.1 Gene <i>CFTR</i>	9
<i>CFTR</i> (Cystic fibrosis transmembrane regulador)	9
Mutações do gene <i>CFTR</i> e a Fisiopatologia da Fibrose Cística	10
Epidemiologia da Fibrose Cística	12
Integração do diagnóstico genético no contexto hospitalar universitário da Região Norte.....	15
A importância do estudo do <i>CFTR</i> em populações indígenas	16
Possibilidades da influência da ancestralidade na expressão do gene <i>CFTR</i>	22
1.2 Preditores de Patogenicidade	23
Utilização e Eficácia.....	23
Utilização dos preditores na detecção de variantes raras em populações sub-representadas	26
2. Justificativa.....	29
3. OBJETIVOS.....	31
3.1 Objetivo Geral	31
3.2 Objetivos Específicos	31
4. Capítulo I.....	32
5. Capítulo II	59
6. DISCUSSÃO GERAL	75
6.1 Previsão de variantes de <i>CFTR</i> por AlphaMissense, SIFT e PolyPhen-2	78
6.2 Previsão de patogenicidade do conjunto	80
6.3 Análise dos preditores com o FoldX	81
7. Conclusão geral	84
8. REFERÊNCIAS BIBLIOGRÁFICAS	85

RESUMO

A fibrose cística (FC) é uma doença genética monogênica com grande variabilidade clínica e genotípica, o que torna o diagnóstico e a interpretação de variantes um desafio constante, especialmente em populações sub-representadas. Esta dissertação foi estruturada em dois capítulos interligados que exploram, de forma complementar, a caracterização genética do gene *CFTR* e a avaliação de ferramentas preditoras de patogenicidade. No Capítulo 1, foi realizada uma análise do gene *CFTR* em uma amostra composta por 110 indivíduos indígenas da Amazônia brasileira. Através de abordagens bioinformáticas, foram identificadas quatro variantes de potencial valor clínico, ainda que se apresentem como *singletons*. Estas variantes foram avaliadas com o preditor AlphaMissense, uma ferramenta de aprendizado profundo capaz de integrar informações evolutivas, estruturais e de sequência, evidenciando seu potencial mesmo em contextos de baixa recorrência, como é o caso de populações geneticamente isoladas. No Capítulo 2, foi conduzida uma análise comparativa entre três preditores amplamente utilizados — AlphaMissense, SIFT e PolyPhen-2 — utilizando 165 variantes do banco clínico CFTR2. Além disso, foi empregada uma abordagem bioenergética com o software FoldX, que estima o impacto de mutações na estabilidade proteica ($\Delta\Delta G$). Os resultados mostraram que o AlphaMissense apresentou desempenho superior na classificação de variantes tanto benignas quanto patogênicas. Mais ainda, observou-se uma correlação significativa entre os escores de patogenicidade gerados por AlphaMissense e os valores de $\Delta\Delta G$ estimados pelo FoldX, indicando que as duas ferramentas, quando usadas em conjunto, oferecem uma estratégia complementar poderosa para a interpretação de variantes. A correlação entre os capítulos evidencia a relevância de incorporar preditores robustos como AlphaMissense em estudos populacionais e clínicos, principalmente em grupos geneticamente diversos ou negligenciados. Além disso, a integração de análises estruturais, como as realizadas com FoldX, pode servir para validar e aprofundar a interpretação funcional dessas variantes. Assim, a dissertação contribui para o avanço da genômica médica de precisão, destacando a importância de abordagens integrativas para a avaliação de variantes do *CFTR* em diferentes contextos populacionais e clínicos.

Palavras-chave: População indígena; Gene *CFTR*; Variabilidade Genética, Preditores.

ABSTRACT

Cystic fibrosis (CF) is a monogenic genetic disease with high clinical and genotypic variability, which makes diagnosis and interpretation of variants a constant challenge, especially in underrepresented populations. This dissertation was structured in two interconnected chapters that explore, in a complementary manner, the genetic characterization of the CFTR gene and the evaluation of pathogenicity prediction tools. In Chapter 1, an analysis of the CFTR gene was performed on a sample of 110 indigenous individuals from the Brazilian Amazon. Through bioinformatic approaches, four variants of potential clinical value were identified, even though they are singletons. These variants were evaluated with the AlphaMissense predictor, a deep learning tool capable of integrating evolutionary, structural, and sequence information, demonstrating its potential even in contexts of low recurrence, as is the case with genetically isolated populations. In Chapter 2, a comparative analysis was conducted between three widely used predictors—AlphaMissense, SIFT, and PolyPhen-2—using 165 variants from the CFTR2 clinical database. In addition, a bioenergetic approach was employed with the FoldX software, which estimates the impact of mutations on protein stability ($\Delta\Delta G$). The results showed that AlphaMissense performed better in classifying both benign and pathogenic variants. Furthermore, a significant correlation was observed between the pathogenicity scores generated by AlphaMissense and the $\Delta\Delta G$ values estimated by FoldX, indicating that the two tools, when used together, offer a powerful complementary strategy for variant interpretation. The correlation between the chapters highlights the relevance of incorporating robust predictors such as AlphaMissense in population and clinical studies, especially in genetically diverse or neglected groups. In addition, the integration of structural analyses, such as those performed with FoldX, can serve to validate and deepen the functional interpretation of these variants. Thus, the dissertation contributes to the advancement of precision medical genomics, highlighting the importance of integrative approaches for the evaluation of CFTR variants in different population and clinical contexts.

Keywords: Amerindian population; *CFTR* gene; Genetic Variability; Predictors.

1. INTRODUÇÃO

1.1 GENE *CFTR*

CFTR (CYSTIC FIBROSIS TRANSMEMBRANE REGULADOR)

O gene *CFTR* é responsável pela expressão da proteína CFTR, por meio da tradução do mRNA, constituída de 1.480 aminoácidos, presente na membrana plasmática das células humanas, sendo encontrado no braço longo do cromossomo 7, na região 7.q31, com 27 éxon codificantes e abrange, aproximadamente, 190 kB de DNA genômico humano (Farinha; Callebaut, 2022). Essa proteína é expressa no pulmão, pâncreas, glândulas sudoríparas, intestino, fígado, mucosa nasal, glândulas salivares e no trato reprodutivo.

Dessa forma, a estrutura da proteína CFTR possui dois domínios trans membranares (TMD1 e TMD2), visto que cada um contém seis segmentos que atravessam completamente a bicamada fosfolipídica, formando assim o poro do canal onde os ânions podem fluir. Estes mesmos segmentos são unidos por três alças extracelulares e duas intracelulares, possuindo função imprescindível na biogênese do *CFTR* e no dobramento pós-traducional para estabelecer interações entre os domínios necessárias para a estabilidade conformacional (Lopes-Pacheco, 2020).

Outras estruturas presentes são os domínios de ligação de nucleotídeos (NBD1 e NBD2) e um domínio regulatório (RD). Estes estão expostos no citosol, e os NBD's possuem uma sequência com alta conservação para ligação e hidrólise de ATP e o RD é desordenado, possuindo inúmeras sequências de consenso que contém serinas e treoninas para a fosforilação pela proteína quinase A (PKA) e proteína quinase C (PKC) (Borkenhagen; Prehm, 2022).

A proteína CFTR é responsável pela formação de um canal de passagem iônica, em que há o transporte de íons cloreto, portanto a interação entre essa com o canal epitelial de sódio (ENaC) permite principalmente a regulação da secreção e absorção de fluidos, mesmo que seu mecanismo ainda não tenha sido completamente elucidado. Fisiologicamente, o funcionamento da proteína CFTR depende da fosforilação do domínio R pela proteína quinase, assim o ATP impulsiona a abertura dos poros possibilitando o funcionamento dos canais iônicos. Após isso, a hidrólise do ATP

no sítio “consenso” (cataliticamente ativo) leva ao fechamento destes poros (Levring et al., 2023).

MUTAÇÕES DO GENE *CFTR* E A FISIOPATOLOGIA DA FIBROSE CÍSTICA

A mutação no gene *CFTR* é monogênica, autossômica e recessiva, visto que para herdar a mutação, deve-se herdar tanto do pai quanto da mãe os alelos recessivos que possuem o *CFTR* irregular, ocorrendo uma homozigose. Caso seja herdado um alelo normal e outro com o gene defeituoso, o indivíduo afetado não costuma manifestar sintomas da fibrose cística (Clancy et al., 2014).

Mais de um tipo de mutação ocorre no gene *CFTR*, logo são divididas em seis classes conforme a disfunção da proteína, caracterizado por: (I) nenhuma produção da proteína de comprimento total, (II) dobramento e tráfego de proteínas prejudicados, (III) controle de canal defeituoso, (IV) reduziu a condutância do canal, (V) reduziu a produção de proteínas e (VI) reduziu a estabilidade da proteína na membrana plasmática (Bacalhau et al., 2023) (Figura 1). A mutação mais comum encontrada no gene *CFTR*, em pacientes de FC é a deleção da fenilalanina na posição 508 (f508del), sendo que apesar das manifestações genotípicas e fenotípicas da FC, segundo Santos et al. (2024) , fatores ambientais também devem ser investigados, ao aparentarem ter influência na gravidade da doença, além disso, os indivíduos heterozigotos para a f508del apresentaram associação para a insuficiência pancreática (Sharma; Cutting, 2020) .

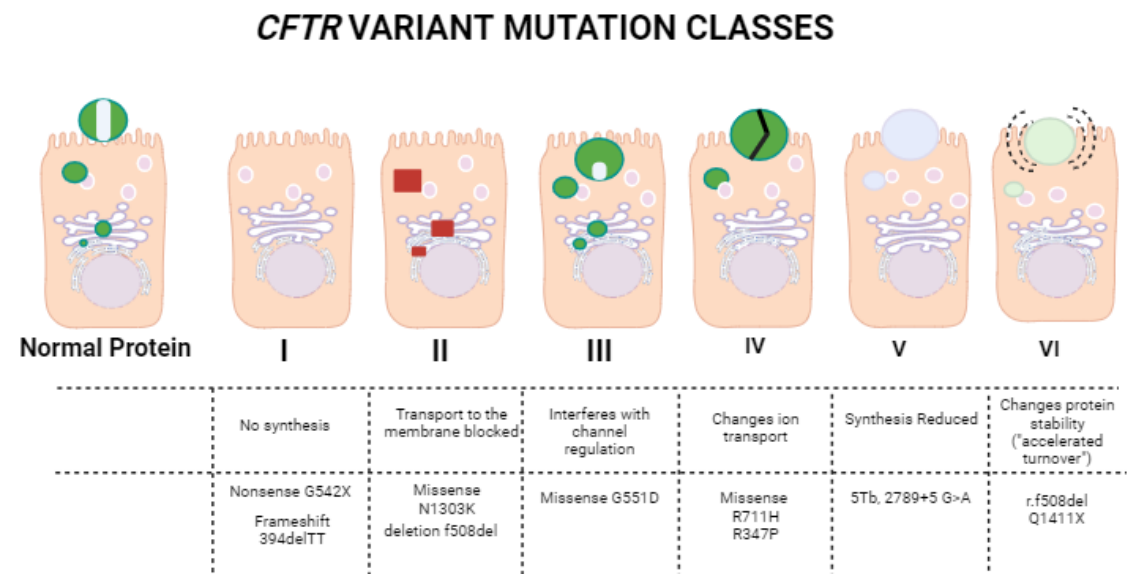


Figura 1 - Classes da mutação do gene CFTR conforme a variante específica e o tipo de mutação.

Dessa forma, a Fibrose Cística é resultante da presença de duas variantes patogênicas do *CFTR*, visto que o diagnóstico pode ser baseado na sintomatologia ou histórico familiar. De forma geral, por apresentar um padrão multissistêmico, as manifestações clínicas podem dividir-se em: respiratórias, como tosse produtiva insistente, obstrução das vias aéreas e infecções recorrentes nos pulmões, doença sinusal (congestão nasal crônica, dores de cabeça, distúrbios do sono e infecções sinusais) que podem, consequentemente, ocasionar exacerbações respiratórias.

Segundo Sharma e Cutting (2020), 2.065 variantes do *CFTR* foram identificadas e colocadas no inventário de dados CFMD (Cistic Fibrosis Mutation Database) em 2019, sendo que as mesmas possuem vários efeitos sobre a função molecular. Em torno de 40% das variantes *missense* (onde ocorre uma substituição de nucleotídeos que gera uma substituição de aminoácidos) identificadas alteram apenas um aminoácido, causando a produção da proteína de forma disfuncional ou dobrada.

Ainda com relação à gravidade de manifestações pelo genótipo do paciente, o estudo de Angyal et al. (2021) demonstrou que aproximadamente 95% dos pacientes avaliados possuíam mutações de classes I, II e III possuem a atividade pancreática reduzida, levando a uma atrofia das estruturas ductais e acinares, resultando assim em extensa insuficiência pancreática, já as mutações mais leves (classe IV e V) são pancreáticos suficientes.

Portanto, a fisiopatologia está relacionada a uma manifestação sistêmica da doença, a qual é comumente conhecida pelos sintomas no trato respiratório. Entretanto, um dos fatores geralmente associados ao diagnóstico de FC é a insuficiência pancreática, visto que o muco espesso bloqueia os ductos pancreáticos, impedindo a liberação de enzimas digestivas (amilase, lipase e protease) no intestino delgado. A obstrução destes ductos impede a passagem das enzimas pancreática e, consequentemente, causa hipovitaminose, além de má absorção das vitaminas lipossolúveis K, E, D e A, um dos fatores presentes na desnutrição (Rosa et al., 2008).

Nos pulmões, o acúmulo do muco na parte apical das células gera um fator propicia a infecção por bactérias, sendo *Pseudomonas aeruginosa* e *Staphylococcus aureus* as mais comuns e responsáveis pelo desencadeamento de infecções respiratórias recorrentes, gerando um acúmulo de dano tecidual nos pulmões devido a frequências

destas ocorrências, sendo a insuficiência respiratória a principal causa de mortalidade entre os pacientes de FC (Lichter et al., 2024; Ode et al., 2019).

Outro órgão afetado são as glândulas sudoríparas, visto que devido à alteração na reabsorção de cloreto e sódio (consequência da proteína disfuncional) é produzido um suor excessivamente salgado. O fígado pode ser afetado devido à obstrução dos ductos biliares, levando a uma hepatopatia biliar que pode evoluir, conseqüentemente, para a cirrose biliar. Quanto ao sistema reprodutor, em ambos os sexos, o acúmulo de muco nos ductos deferentes (homens) e o muco cervical espesso (mulher) progridem, geralmente, para casos de infertilidade (Ode et al., 2019).

Assim, a FC é uma doença na qual o desequilíbrio iônico provoca o aumento da viscosidade do muco produzido nas células epiteliais e, conseqüentemente, causa o congestionamento de ductos e propicia a proliferação de infecções por microrganismos (Jia; Taylor-Cousar, 2023). O desequilíbrio iônico principalmente do HCO_3^- prejudica a solubilização das mucinas, logo dificuldade de secreção extracelular tanto do Cl^- e do HCO_3^- geram uma desregulação no pH do lúmen celular, deixando o muco secretado além de hiper viscoso, com pH básico, propiciando a reprodução de bactérias.

EPIDEMIOLOGIA DA FIBROSE CÍSTICA

Em sua primeira descrição, no ano de 1938, feito pela patologista e pediatra Dorothy Andersen, a Fibrose Cística era conhecida predominantemente como uma doença pediátrica na qual os nascidos não ultrapassavam o primeiro ano de vida. Atualmente, nota-se um grande número pessoas com FC adultas em relação a crianças afetadas, evidenciando os avanços científicos com relação ao tratamento dos sintomas, prolongando a expectativa de vida (Scotet; L'Hostis; Férec, 2020) . Atualmente, os dados epidemiológicos apontam que a FC é prevalente em populações de origem ou descendência caucasiana, variando de 1:1.400 na Irlanda a 1: 3500 nos Estados Unidos, não restringindo a ocorrência em outros grupos étnicos. Na América Latina essa ocorrência varia entre 1: 1 600-14.000 recém-nascidos vivos e no México, 1: 1 850 nascidos-vivos (López-Valdez et al., 2021) .

No Brasil, é estimado a prevalência de FC entre 1:7 500 a 1:15 000 nascidos vivos a depender da região, indicando assim inúmeros casos ainda podem estar

subnotificações, evidenciando que muitos pacientes ainda não possuem um acompanhamento médico adequado (HASIAK SANTO; RIBEIRO FERREIRA DA SILVA-FILHO, 2021). Neste mesmo estudo, o levantamento de dados de mortes por FC no período de 1999 a 2017 evidenciou as taxas de mortalidade pela doença no Brasil. Entre este período, foram contabilizadas um total de 2854 mortes relacionadas a FC. Com relação à idade, verificou-se que em torno de 20%, 25%, 50% e 25% foram a óbito com menos de 1, 4, 22 e 65 anos, respectivamente.

Logo, esta tendência ao aumento da mortalidade observada conflui com a melhoria dos diagnósticos, principalmente com a implementação do teste do pezinho no Sistema Único de Saúde (SUS) entre os anos de 2000 e 2001 em estados como Santa Catarina, Paraná e Minas. Com isso, começou a ser possível identificar se o recém-nascido possui a doença e começar a terapêutica farmacológica, melhorando significativamente a qualidade de vida e, conseqüentemente, aumentando a expectativa de vida (Malaquias; Cardoso, 2023).

Em suma, as estimativas na literatura indicam uma sobrevida de aproximadamente 43 anos no país, mas apesar do acesso mais amplificado tanto de informações sobre a doença quanto de testes genéticos, não é possível mensurar a quantidade de casos subnotificados, tanto da doença quanto por mortalidade ocasionada pela mesma (Hasiak Santo; Ribeiro Ferreira Da Silva-Filho, 2021). As informações que estão disponibilizados atualmente estão concentrados em regiões específicas, como sul e sudeste, devido a fatores como concentração de renda, dificuldade no acesso ao sistema de saúde e a não inclusão, na maioria, de populações sub representadas.

A partir da ação do Registro Brasileiro de Fibrose Cística (REBRAFC) foi constatado que entre 2001 à 2021, o diagnóstico de indivíduos com FC foi de 3596 casos, sendo que 1.803 (53,2%) foram constatados por triagem neonatal. Em 2021 e 2022 o número de diagnósticos foi reduzido, provável consequência da pandemia de Covid-19 (Coronavirus Disease, 2019). Quanto a genotipagem, apesar de uma porcentagem significativa de 82,76% para indivíduos que realizaram este procedimento, o registro não especifica informações sobre qual tipo de teste genético foi realizado e qual a sua abrangência na que pode ir além da identificação de variante, como regiões codificantes e áreas intrônicas no gene *CFTR* (http://portalgbefc.org.br/ckfinder/userfiles/files/Rebrafc_2021_REV_fev24.pdf).

No estudo da população brasileira, a amostra populacional utilizada foi dos indivíduos das cinco regiões do país que possuíam diagnóstico ou estavam em

tratamento, considerando um total de $n = 6427$ para realizar o exame para diagnóstico da FC. Dentro desse total, 5319 indivíduos realizaram a pesquisa de genótipo, sendo que metade tem pelo menos uma cópia da variante f508del (2.682, 52,8% dos indivíduos), onde 1.244 (24,5%) são homozigotos e 1.438 (28,3%) são heterozigotos. Assim, a variante f508del é a mais comum entre os brasileiros, identificada em 44,3% dos alelos, sem contar, também, com os prováveis casos que possam estar subnotificados (http://portalgbefc.org.br/ckfinder/userfiles/files/Rebrafc_2021_REV_fev24.pdf).

Outra observação notável é a prevalência observada de casos de FC nas regiões sul e sudeste, estudos anteriores como de Araújo et al (2005), constatarem cerca de 1:9.600 nascimentos para a prevalência de FC, visto a prevalência de notificações da doença das regiões Sul e sudeste, que possui uma população majoritariamente caucasiana (Bezerra et al., 2023). No Rio Grande do Sul observou-se a frequências altas que variavam entre 1:1.600 e 1:1.700 nascimentos, ou seja, em cada um em cada 20 habitantes deste local é heterozigoto para a FC.

No estudo de Procianoy, Neto e Ribeiro (2023) é notória a heterogeneidade de distribuição dos Centros de Tratamento de Fibrose Cística no Brasil e, também, na concentração de pacientes por estado/região, sendo o Sudeste a região que mais apresenta polos específicos de tratamento para a FC (21 centros) equivalendo a 41%, enquanto a região Norte (2 centros), com 4%. Durante o período desse estudo, foi relatado que do total de 4.371 pacientes atendidos, 141 indivíduos eram da região norte.

Existem poucos estudos detalhados sobre a ocorrência de Fibrose Cística na região norte, que possui uma grande diversidade étnica resultante do processo de miscigenação. Além disso, ainda existem grupos que vivem de forma mais isolada — no caso, os povos indígenas — na qual a padronização de uma assistência em saúde, mas eficaz e, conseqüentemente, recursos para se realizar estudos são extremamente difíceis, o que torna ao nosso compêndio de informações mais escasso.

Araújo et al. (2005) sugerem que as baixas frequências encontradas com relação a variantes da FC podem estar relacionadas ao processo de formação da população da Amazônia Brasileira. No mesmo estudo, a amostra examinada de 33 pacientes (66 cromossomos no total) na cidade de Belém, localizada na região norte no estado do Pará, apresentou um percentual de 22,7% para a mutação f508del, sendo que 15 indivíduos eram heterozigotos e apenas dois homozigotos para a mesma mutação e apenas 3% apresentaram a mutação G551D.

INTEGRAÇÃO DO DIAGNÓSTICO GENÉTICO NO CONTEXTO HOSPITALAR UNIVERSITÁRIO DA REGIÃO NORTE

O Hospital Universitário João de Barros Barreto (HUJBB-UFPA), integrante da Rede Ebserh, consolidou-se como referência estadual no diagnóstico e tratamento da fibrose cística no Pará. O hospital é o único centro público do estado habilitado a realizar o teste do cloreto do suor, considerado o padrão-ouro para a confirmação clínica da doença, desempenhando papel central para o acesso ao diagnóstico precoce em crianças e adultos (HOSPITAL UNIVERSITÁRIO JOÃO DE BARROS BARRETO, 2022).

Além disso, o HUJBB oferece acompanhamento multiprofissional, com equipe composta por pediatras, fisioterapeutas, nutricionistas, psicólogos, assistentes sociais, enfermeiros, farmacêuticos, cardiologista e endocrinologista, o que possibilita um cuidado contínuo e integrado. Essa abordagem é destacada pelo próprio Complexo Hospitalar Universitário da UFPA, que reconhece o Barros Barreto como centro de referência regional em doenças raras, incluindo a fibrose cística (COMPLEXO HOSPITALAR UNIVERSITÁRIO DA UFPA, 2020.).

No contexto da saúde pública, a aplicabilidade do HUJBB vai além da assistência direta ao paciente. O acesso ao teste do suor, um exame de difícil disponibilidade e custo elevado, é assegurado gratuitamente pelo hospital, garantindo equidade no atendimento de populações historicamente vulneráveis da região Norte (HOSPITAL UNIVERSITÁRIO JOÃO DE BARROS BARRETO, 2022). A centralização desse serviço em uma instituição universitária também permite integrar ensino, pesquisa e assistência, além de formar recursos humanos qualificados para o manejo da fibrose cística e outras doenças raras (COMPLEXO HOSPITALAR UNIVERSITÁRIO DA UFPA, 2020).

Outro aspecto relevante é a contribuição do ambulatório especializado para a coleta de dados epidemiológicos regionais. Esses dados subsidiam políticas públicas mais eficazes e possibilitam a articulação com programas nacionais, como a distribuição de medicamentos de alto custo pelo Componente Especializado da Assistência Farmacêutica. Ao promover o acompanhamento sistemático e multidisciplinar, o HUJBB contribui para a redução de internações recorrentes e complicações graves, favorecendo a racionalização de recursos dentro do SUS (HOSPITAL UNIVERSITÁRIO JOÃO DE BARROS BARRETO, 2022).

Portanto, o Hospital Barros Barreto representa não apenas um centro de tratamento, mas também um pilar estratégico da saúde pública no Pará. Sua atuação evidencia como a presença de hospitais universitários de referência fortalece o SUS, ampliando o acesso a tecnologias diagnósticas e terapêuticas e assegurando maior equidade no cuidado à população amazônica (COMPLEXO HOSPITALAR UNIVERSITÁRIO DA UFPA, s.d.).

A IMPORTÂNCIA DO ESTUDO DO CFTR EM POPULAÇÕES INDÍGENAS

O estudo do gene *CFTR* em populações indígenas brasileiras tem uma relevância significativa na genética, ao ampliar o conhecimento sobre variantes genéticas específicas de grupos sub-representados. Embora a fibrose cística seja rara nessas populações, variantes específicas podem apresentar padrões únicos, não identificados em populações amplamente estudadas. Identificar essas variantes é crucial para compreender a diversidade genética e adaptar abordagens diagnósticas e terapêuticas que considerem as particularidades genômicas desses grupos.

O impacto dessas condições não se restringe à saúde individual, mas também afeta comunidades inteiras. Muitas vezes, doenças raras têm implicações socioculturais significativas, exacerbando vulnerabilidades econômicas e sociais já presentes em populações indígenas. Assim, compreender o efeito dessas doenças é crucial para desenvolver políticas de saúde pública inclusivas (Baynam et al., 2017). Isso requer abordagens éticas e colaborativas que respeitem a cultura indígena, promovendo a equidade no acesso ao cuidado médico e à pesquisa científica.

Na farmacogenética, essas pesquisas são essenciais para desenvolver tratamentos mais precisos e eficazes, considerando a variação na resposta a medicamentos. Populações indígenas podem apresentar variantes exclusivas do gene *CFTR* que afetam o metabolismo de fármacos usados no manejo da fibrose cística e outras condições relacionadas. Dessa forma, os dados genômicos gerados podem subsidiar estratégias de medicina personalizada, com impactos diretos na saúde pública e na equidade no acesso a terapias adequadas.

De acordo com (Rodrigues et al., 2022), ao tratarmos dos perfis ameríndios, por exemplo, outra estratégia deve ser aplicada visto que existem dois fatores importantes a serem considerados: 1) se um grupo for predominante, deve-se dar

atenção em como seu perfil genético se expressa; 2) se tratarmos a questão de um padrão genético único influenciado diretamente por fatores como endogamia, isolamento geográfico, seleção natural e deriva genética.

Dado a sua importância, o estudo realizado com indígenas por Rodrigues et al. (2022), verificou um total de 160 genes, dos quais foram encontradas um total de 3.311 variantes, sendo que 167 são exclusivas da população de ameríndios analisada. Sendo assim, a implantação dos genótipos indígenas é necessária para ampliar as possibilidades de aplicação farmacogenética; o estudo de determinadas variantes existentes especificamente nessas populações pode ajudar a elucidar possíveis interações com fármacos, outras substâncias com potencial farmacológico, patologias e suas variantes, além dos níveis de metabolização individual e toxicidade.

Mesmo com a sua limitada interação com povos externos, a singularidade apresentada nos sequenciamentos do genoma destes indivíduos nos chama a atenção para a interação com o metabolismo de fármacos expressas por variantes que, em alguns casos, são restritas a determinada população. Rodrigues et al. (2021) ratifica que apesar dessas populações indígenas da Amazônia apresentarem pouco cruzamento com grupos não indígenas, a alta variabilidade genética que apresentam pode ser uma “chave” para a compreensão de doenças hereditárias/genéticas, para fornecer informações que possam esquematizar uma previsão da resposta metabólica a drogas e outros xenobióticos. Logo, populações indígenas são distintas e necessitam de algoritmos específicos para dosagem de farmacológicas.

Como exemplo, a investigação conduzida por Rodrigues et al. (2019) analisou 26 genes relacionados ao mecanismo ADME (Absorção, Distribuição, Metabolismo e Excreção) de fármacos em uma amostra de população ameríndia da Amazônia. Esses dados foram comparados com cinco populações do Projeto 1000 Genomas: África (AFR), Europa (EUR), América (AMR), Leste Asiático (EAS) e Sul da Ásia (SAS). Embora o estudo tenha avaliado 32 variantes genéticas distribuídas nesses 26 genes, três deles se destacaram por apresentarem perfis significativamente distintos na população ameríndia: *ABCB1*, *CYP2A6* e *DPYD*. O gene *ABCB1*, que codifica a glicoproteína P (P-gp), pode influenciar a biodisponibilidade de fármacos como o Aliscireno, devido à redução da atividade dessa proteína transportadora. Já o *CYP2A6* está entre as principais enzimas hepáticas envolvidas na metabolização de medicamentos e xenobióticos, sendo que indivíduos com metabolismo lento — com base em polimorfismos específicos — estão mais sujeitos a efeitos tóxicos. Por fim,

variações no gene *DPYD*, que participa do metabolismo de fluoropirimidinas, também foram evidenciadas, reforçando a singularidade genética das populações ameríndias e sua importância para a farmacogenômica personalizada. Na população indígena estudada, verificou-se a frequência deste perfil em 46%; por último tem-se o gene *DPYD*, que codifica a enzima dihidropirimidina desidrogenase (DPD), responsável por aproximadamente 85% do catabolismo hepático do metabólito ativo 5-fluoracil para a forma inativa 5-fluoro-5,6-diidrofluoracil (fármaco fluoracil) e a baixa atividade desta enzima tem, por consequência, maior risco de toxicidade grave e até letal (Rodrigues et al., 2019).

Em um estudo posterior, a análise de nativos americanos realizado por Moreira et al. (2022) verifica se há predominância se o alelo PDE4B rs6683977 é realmente predominante nessas populações. O PDE4B (fosfodiesterase-4B) é responsável pela degradação do AMP cíclico (cAMP) e sua expressão também é relacionada a ação de corticoesteroides no tratamento de câncer, especificamente na leucemia linfoblástica aguda (LLA).

O mecanismo de ação ocorre após a ativação do receptor de células B e o cAMP modulando negativamente as vias sinalizadoras responsáveis pela proliferação celular. Portanto, a super expressão de PDE4B anula a inibição do cAMP da multiplicação celular e propicia uma sensibilidade e/ou resistência aos glicocorticoides. Tendo em vista esse efeito, o estudo destas variantes é altamente relevante para o tratamento farmacológico de LLA, visto que o alelo PDE4B-rs6683977.G está associado a recidiva da doença e elucidar seus mecanismos pode ser essencial nas formulações de tratamentos futuros direcionados a uma terapia precisa, levando os aspectos limitantes em consideração (Moreira et al., 2022).

Portanto, entende-se que o processo “chave-fechadura” é um dos momentos na farmacodinâmica que explica com detalhamento o acoplamento (ou encaixe) da substância terapêutica com o seu receptor orgânico. Na farmacogenética, visamos elucidar as possíveis interferências genéticas inerentes ao indivíduo com a terapia a ser administrada. A partir do sequenciamento e sistematização de genes e variantes presentes nas comunidades indígenas, pode-se estabelecer um protocolo de tratamento personalizado para cada indivíduo que possa apresentar determinado fenótipo, reduzindo significativamente o risco de reações adversas (Rodrigues et al., 2019).

Para indústria farmacêutica, principalmente na medicina de precisão, as novas variantes e a compreensão de como seus mecanismos afetam cada indivíduo

abrem oportunidades para novos testes, seja com dosagens de fármacos existentes e comercializados ou com moléculas com potencial efeito farmacológico, de modo a flexibilizar e ampliar o número de opções de terapêuticas.

Assim, a farmacogenética pode ser definida como a ciência aprofundada no estudo das interações genéticas com a ação dos fármacos no organismo, ou seja, como a variabilidade genética pode influenciar a ação de medicamentos em qualquer tratamento farmacoterapêutico, sendo aplicada para a prevenção de efeitos adversos e tornando as terapias farmacológicas mais eficazes, objetivando reduzir fracassos terapêuticos (Sainz De Medrano Sainz; Brunet Serra, 2023) .

Um dos maiores desafios na atualidade no tratamento de doenças são as Reações Adversas aos Medicamentos (RAMs), sendo responsáveis por casos de mortalidade e morbidade (Mota; Vigo; Kuchenbecker, 2021) . Segundo Pessoa, Nácul, e Noël (2006), um dos primeiros relatos farmacogenéticos associavam uma resposta a fármacos a certa com certa proporção ancestral, como, por exemplo, no período da Segunda Guerra Mundial em que aproximadamente 10\% dos soldados afro-americanos e apenas um pequeno número de soldados caucasianos desenvolviam uma crise hemolítica aguda quando administrada uma dose média de primaquina ou outro antimalárico.

Sendo assim, a influência genética individual pode interferir no processo farmacocinético dos fármacos, ocorrendo reações anormais devido a fatores farmacocinéticos ou farmacodinâmicos (Pessoa; Nácul; Noël, 2006) .Visto que estas substâncias estão sujeitas a mecanismos de fase I (oxidação, redução e hidrólise) e II (acetilação, glucoronidação, sulfatação e metilação) para a conversão em metabólitos mais solúveis e que posteriormente serão excretados com mais facilidade ou até em converter pró-fármacos em substâncias farmacologicamente ativas, qualquer alteração no gene pode afetar a velocidade de metabolização do indivíduo (muito lenta ou muito acelerada) ou aumentar a toxicidade a determinado fármaco, influenciando diretamente na resposta e eficácia de qualquer tratamento farmacológico (Garcia; Andrade; Palmero, 2022).

Tendo em vista que o organismo humano pode ser comparado a uma rede metabólica complexa, cada indivíduo apresenta individualidades específicas em seu código genético, e, portanto, as respostas a doses terapeuticamente recomendadas podem variar (Lauschke; Ingelman-Sundberg, 2020). Por exemplo, pode-se abordar

como exemplo a variabilidade específica a cada gene influenciando no conjunto de enzimas da família do citocromo p450.

Quanto a Fibrose cística, as investigações acerca do *CFTR* visam uma farmacoterapia que restaure a função da proteína conforme o mecanismo apresentado. A inserção de um gene saudável já foi cogitada, entretanto, os efeitos colaterais apresentados na coorte do estudo fizeram essa possibilidade ser, por enquanto, pausada (Perry et al., 2016). As variantes, classificadas em seis classes, possuem particularidades quanto ao tipo (deleções, inserções, *missense* e *non-sense*) que podem vir a interferir na farmacoterapia.

Com relação aos registros no Brasil, o Grupo Brasileiro de Estudos de Fibrose Cística (GBEFC) fundado no ano de 2003, é um dos principais meios de divulgação e rastreamento da FC no país. Anualmente são disponibilizados relatórios contendo dados demográficos, de diagnóstico do tratamento de pacientes com fibrose cística. O documento mais atual é referente ao ano de 2021, período pós-pandêmico, em que é relatada a aprovação do medicamento para indivíduos ao menos um alelo da mutação f508del, o Trikafta (elexacaftor/tezacaftor/ivacaftor) (http://portalgbefc.org.br/ckfinder/userfiles/files/Rebrafc_2021_REV_fev24.pdf).

Para isso, na medicina de precisão e consequentemente na farmacogenômica, o objetivo do tratamento dá-se na resposta modulada conforme o genótipo dos indivíduos. Medicamentos como o Ivacaftor não possuem resposta significativa, por exemplo, para homozigotos da mutação f508del, sendo mais eficiente, segundo o Clinical Pharmacogenetics Implementation Consortium (CPIC) em pacientes que possuem a mutação *missense* G551D (Clancy et al., 2014).

No contexto do gene *CFTR*, a farmacogenética durante muito tempo teve seu foco de pesquisa direcionado para o desenvolvimento de medicamentos que se concentravam na melhoria dos sintomas. Posteriormente, novos fármacos foram desenvolvidos segundo a modificação estrutural apresentada na proteína, conforme cada classe de variante. Atualmente, no CPIC, um dos medicamentos em destaque é o Ivacaftor.

O Ivacaftor (ou Kalydeco) é um potencializador do *CFTR* que tem indicação clínica — aprovado pela FDA — para pacientes com 1 mês de vida ou mais que apresentem ao menos uma mutação no gene. Logo, se o genótipo do indivíduo for desconhecido, um teste genético será necessário para detectar e quais são as mutações presentes no gene (Clancy et al., 2014). Na atualização do ano de 2019 no PharmGKB,

foram adicionadas cinco novas variantes onde o uso de Ivacaftor foi aprovado: 2789 + 5G→A (rs80224560), 3272-26A→G (rs76151804), 3849 + 10kbC →T (rs75039782), 711 + 1G →T (rs77188391) e E831X (rs78655421).

O Ivacaftor age aumentando a atividade do canal de cloreto da proteína CFTR, que para ficar aberto, precisa de duas condições para a efetivação da sua eficácia. Primeiramente, a proteína deve estar expressa na superfície epitelial das células, logo defeitos de dobramento ou tráfego não são reagentes ao tratamento com Kalydeco. Segundo, a proteína deve ser capaz de ser ativada por meio de mecanismos normais de sinalização intracelular, especialmente vias de adenosina monofosfato e proteína quinase (Van Goor et al., 2014). Entretanto, consoante as diretrizes do Clinical Pharmacogenetics Implementation Consortium (CPIC) para terapia com Ivacaftor ressalta que este fármaco não demonstrou eficácia para homozigotos da mutação f508del, provavelmente por conta da expressão desacertada da proteína.

Em suma, o tratamento com ivacaftor é fortemente indicado para heterozigotos da mutação G551D e não apresenta respostas clínicas satisfatórias em homozigotos para a f508del (Clancy et al., 2014). Segundo o trabalho de (Duckers et al., 2021), a função pulmonar em uma coorte tratada com Ivacaftor apresentou melhor função pulmonar preservada do que o grupo que não foi tratado com o fármaco. Além disso, foi observado a diminuição do uso de recursos hospitalares, como internações decorrentes de exacerbações, diminuição no uso e no tempo de antibioticoterapia para a contenção de infecções bacterianas.

Outro medicamento indicado para tratamento da FC é o Trikafta (Ivacaftor/ Tezacaftor/ Elexacaftor, que conforme a anotação do EMA (European Medicines Agency), é indicado para um regime combinado em pacientes que possuam ao menos uma mutação f508del no gene *CFTR*. Tanto a anotação do FDA (The Food and Drug Administration) quanto do EMA ressaltam que, caso o genótipo do paciente seja desconhecido, um teste deverá ser realizado para confirmação, sendo que no FDA são abordadas mais variantes do que na descrição do EMA (<https://www.pharmgkb.org/literature/15104482/overview>).

O mecanismo de ação se divide na ação concomitante do efeito dos três fármacos. Elexacaftor e tezacaftor ligam-se a locais diferentes da proteína, facilitando o processamento celular e o tráfego de f508del para aumentar a quantidade de proteína CFTR disponível na superfície celular. O ivacaftor, como mencionado anteriormente, potencializa a possibilidade da abertura do canal da proteína CFTR na superfície celular

(Figura 2). O resultado dessa combinação é o aumento da quantidade e da função de f508del na superfície celular, resultando no aumento da atividade do *CFTR* (<https://www.pharmgkb.org/literature/15104482/overview>).

POSSIBILIDADES DA INFLUÊNCIA DA ANCESTRALIDADE NA EXPRESSÃO DO GENE CFTR

O estudo da ancestralidade na população da América Latina é essencial visto que as características de miscigenação aparentam ser diferentes. Logo, compreender esses padrões considerando os vários eventos de relações interétnicas é essencial para a construção do passado evolutivo e mapeamento de informações genéticas (Suarez-Kurtz, 2005).

No estudo de Suarez-Kurtz (2005), foi identificado o efeito protetor dos polimorfismos do CYP1A1 contra o câncer de pulmão em fumantes, sendo que a presença de alelos variantes CYP1A1*2A ou CYP1A1*2C reduziu significativamente a razão de probabilidade de câncer de pulmão entre os americanos classificados como latinos, mas não entre os afro-americanos.

Já Mendes et al. (2020), em sua investigação na ancestralidade das populações indígenas, concluiu que os ameríndios são geneticamente mais semelhantes às populações americanas e do leste asiático e mais distantes das populações africanas. Concomitantemente, ele aborda sobre como a ancestralidade latino-americana resulta da mistura entre europeus, africanos e nativos americanos, tornando o genoma resultante em um mosaico constituído de fragmentos que podem, em investigações posteriores, servir como uma “trilha genética”.

Com relação à expressão do gene *CFTR* na fibrose cística, por exemplo, a predominância dos casos de FC estão concentrados nas regiões sul e sudeste em que geograficamente a apresenta maior proporcionalidade à ascendência europeia. Sabe-se que geograficamente essas regiões tiveram um grande fluxo de imigrantes europeus no período, principalmente de alemães (predominância no sul) e italianos (sul e sudeste) (Kehdy et al., 2015), persistindo nessa herança genética até os dias atuais.

Dados genome-wide revelam que nestas regiões há maior proporção de ancestralidade europeia no sul e sudeste. Na análise mtDNA (DNA mitocondrial) realizada por Kehdy et al. (2015) as populações de Salvador possuem maior proporção

de ancestralidade africana (50,8%), 42,9% e 6,4% de ascendência europeia e ameríndia, respectivamente, sendo explicado por apresentar menor índice de endogamia. Já em Bambuí e Pelotas, a predominância de ancestralidade europeia, visto que apresenta maior presença de endogamia na estrutura familiar. Essas evidências são reforçadas pelo estudo recente de Nunes et al. (2025), que analisou 2.723 genomas brasileiros e confirmou a distribuição assimétrica da ancestralidade genética no país — com maior proporção de ancestralidade africana nas regiões Norte e Nordeste, europeia no Sul e Sudeste, e indígena no Norte e Centro-Oeste. O estudo também demonstra como eventos históricos, padrões de acasalamento e pressões seletivas moldaram a diversidade genômica da população brasileira, revelando milhões de variantes previamente não descritas, incluindo 36.637 classificadas como potencialmente deletérias, com implicações relevantes para a saúde pública (Nunes et al., 2025).

Apesar desta afirmação, o processo de colonização em regiões predominantemente indígenas que ocorreram durante o período das expedições — principalmente na América Latina e do Sul — abrem espaço para a miscigenação intensa, levando a uma variabilidade genética significativa (Souza et al., 2019). Logo, é interessante abordar na investigação como essa miscigenação influencia na possível presença do gene CFTR nestas populações indígenas e como as prováveis variantes podem fornecer uma visão futura sobre como manejar o tratamento para a FC de uma forma mais eficaz e assertiva por meio da farmacogenômica (Rodrigues-Soares; Kehdy; Gouveia, 2021).

Organizar características genéticas em grupos populacionais facilita a identificação por meio de padrões gerais observáveis, mas deve ser estritamente como única forma de análise, visto que para a elaboração de uma prescrição personalizada, deve-se considerar os aspectos individuais. Ou seja, os fatores população/indivíduo devem ser complementares e não vistos de forma antagônica (Suarez-Kurtz, 2005).

1.2 PREDITORES DE PATOGENICIDADE

UTILIZAÇÃO E EFICÁCIA

O advento das tecnologias de sequenciamento de nova geração (NGS) gerou um volume exponencial de dados genômicos, exigindo ferramentas bioinformáticas robustas para análise e interpretação. Nesse contexto, os preditores de patogenicidade desempenham um papel central, pois possibilitam a classificação de variantes genéticas em termos de seu impacto potencial no fenótipo do indivíduo. Ao integrar dados estruturais e funcionais, essas ferramentas têm sido essenciais para decifrar as complexas interações entre genótipo e fenótipo em desordens genéticas (Grimm et al., 2015).

As ferramentas de predição de patogenicidade são amplamente baseadas em algoritmos de aprendizado de máquina, como *Random Forest*, *Support Vector Machines* e Redes Neurais, que aprendem padrões a partir de bases de dados previamente anotadas (Karim et al., 2023). Esses métodos utilizam características como conservância evolutiva, localização estrutural e funcional, e propriedades físico-químicas das variantes genéticas para inferir sua relevância patológica. Entretanto, a eficácia dessas ferramentas depende diretamente da qualidade e diversidade dos conjuntos de dados de treinamento.

Estudos como o de Gunning et al. (2021) destacam que a variabilidade aleatória nos dados de treinamento pode levar a problemas como overfitting, onde o modelo aprende peculiaridades específicas do conjunto de treinamento em detrimento de sua generalização para novos dados. Essa limitação não apenas reduz a precisão dos preditores, mas também gera preocupação sobre sua confiabilidade em contextos clínicos, onde decisões críticas dependem de análises precisas (Barreñada et al., 2024).

Dessa forma, a confiabilidade dos preditores de patogenicidade é um aspecto central para sua aplicação clínica e de pesquisa. Apesar dos avanços metodológicos, o desempenho de um modelo pode variar amplamente dependendo da qualidade e diversidade dos dados de treinamento utilizados. Bases de dados enviesadas, com representação insuficiente de variantes genéticas oriundas de diferentes populações, podem comprometer a validade do modelo ao restringir sua aplicabilidade. Em um contexto global, a falta de representatividade de populações sub-representadas nos bancos de dados genômicos pode levar a interpretações errôneas de variantes específicas, exacerbando disparidades em saúde.

Além disso, a seleção de recursos para treinamento de modelos preditivos é um ponto crítico. A incorporação de características genômicas, como a conservação evolutiva, o impacto funcional em proteínas e as interações estruturais, tem se mostrado eficaz na melhoria da performance de preditores. No entanto, a escolha inadequada ou o excesso de variáveis pode amplificar o risco de overfitting, reduzindo a capacidade do modelo de prever variantes desconhecidas com precisão. Estudos recentes têm explorado abordagens baseadas em validação cruzada e regularização para mitigar tais desafios, proporcionando modelos mais robustos e generalizáveis (Charilaou; Battat, 2022).

Outro aspecto relevante é a integração de dados multiômicos, como transcriptômica, proteômica e epigenômica, na predição de patogenicidade. Essas abordagens permitem uma análise mais abrangente do impacto funcional das variantes, indo além da sequência genômica e considerando efeitos em cascata nos níveis de expressão e regulação gênica (Wu; Xie, 2024). No entanto, a complexidade inerente à integração de múltiplos tipos de dados apresenta desafios técnicos, como o manejo de diferentes formatos, escalas e graus de confiabilidade das informações.

Por fim, o desenvolvimento de preditores mais explicáveis e transparentes, baseados em técnicas de inteligência artificial (IA), é uma área emergente de pesquisa. O conceito de "Inteligência Artificial Explicável" (*Explainable AI, XAI*) tem ganhado destaque por fornecer justificativas interpretáveis para as decisões dos modelos, aumentando sua aceitação em contextos clínicos (Sheu; Pardeshi, 2022). Ferramentas que combinam precisão com explicabilidade podem facilitar a adoção de preditores de patogenicidade na prática clínica, ao fornecer informações claras e confiáveis para apoiar a tomada de decisão médica.

Portanto, apesar dos avanços significativos no campo, há uma necessidade contínua de melhorar os preditores de patogenicidade, considerando aspectos como a representatividade populacional, a integração de dados multiômicos e a explicabilidade dos modelos. Investir nessas áreas é essencial para garantir que essas ferramentas atendam às exigências tanto da pesquisa biomédica quanto da prática clínica, contribuindo para a efetiva implementação da medicina personalizada.

UTILIZAÇÃO DOS PREDITORES NA DETECÇÃO DE VARIANTES RARAS EM POPULAÇÕES SUB-REPRESENTADAS

A identificação de variantes raras em subpopulações tem ganhado destaque na pesquisa genômica, especialmente devido ao potencial dessas variantes em elucidar mecanismos moleculares subjacentes a doenças hereditárias e raras, fornecendo assim insights sobre variações específicas que podem estar incluídas nestas populações (Fatumo; Chikowore; Kuchenbaecker, 2022).

Dessa forma, os preditores de patogenicidade desempenham um papel central nesse cenário, fornecendo ferramentas para a classificação de variantes com base em sua probabilidade de impacto funcional e clínico, principalmente em decorrência dos avanços acelerados na biologia molecular e, conseqüentemente, no grande volume de dados gerados (Garcia; Andrade; Palmero, 2022). Essa relevância é evidente em subpopulações, onde a presença de variantes específicas pode não ser representada adequadamente em bases de dados genômicas amplas, mas possui um impacto significativo na saúde local. O uso de preditores de patogenicidade em subpopulações enfrenta desafios únicos, visto que a limitada representação dessas populações em estudos genômicos globais resulta em lacunas importantes na interpretação de variantes, subestimando ou superestimando seus efeitos patogênicos.

Um dos principais desafios na interpretação de variantes genéticas raras é a escassez de evidências funcionais associadas a variantes *missense*, o que compromete a acurácia diagnóstica e dificulta o desenvolvimento de terapias direcionadas (Cheng et al., 2023). Nesse cenário, abordagens baseadas em aprendizado de máquina têm se mostrado promissoras ao suprir lacunas na predição da patogenicidade em dados genômicos. Ferramentas *in silico*, como o AlphaMissense, recebem sequências proteicas como entrada e estimam o impacto patogênico de todas as possíveis substituições de aminoácidos em uma dada posição (Jumper et al., 2021). Entre suas principais vantagens, destacam-se a incorporação de modelos altamente precisos da estrutura tridimensional das proteínas e a capacidade de aprender restrições evolutivas a partir de sequências homólogas, o que contribui para uma predição mais robusta e biologicamente fundamentada.

Outro preditor amplamente utilizado na avaliação de variantes *missense* é o PolyPhen-2, que combina dados de conservação evolutiva e características estruturais das proteínas para estimar o impacto funcional das substituições de aminoácidos. O

método emprega um classificador probabilístico do tipo Naïve Bayes, treinado com dois conjuntos de dados distintos — HumDiv e HumVar — permitindo sua adaptação tanto a contextos de doenças mendelianas quanto a estudos de variação poligênica (Adzhubei; Jordan; Sunyaev, 2013). Embora apresente sensibilidade geralmente satisfatória, sua baixa especificidade compromete a aplicação clínica de forma isolada, especialmente em cenários que exigem alta precisão na distinção entre variantes benignas e patogênicas (Ernst et al., 2018).

Ademais, o SIFT (Sorting Intolerant From Tolerant) constitui um método que prevê se uma substituição não sinônima de aminoácido comprometerá a função da proteína (Ng, 2003). Essa ferramenta baseia-se exclusivamente em dados de conservação evolutiva derivados de alinhamentos múltiplos de sequências homólogas para inferir o impacto funcional das mutações. Diferentemente de preditores como AlphaMissense e PolyPhen-2, o SIFT não requer modelos tridimensionais, o que possibilita sua aplicação em larga escala, desde que haja um número adequado de sequências homólogas disponíveis (Sim et al., 2012).

Assim, o desenvolvimento de modelos preditivos que incorporem características específicas de subpopulações, como frequências alélicas locais e contexto genético, é essencial para aumentar a precisão e relevância clínica dessas ferramentas. Estratégias baseadas em aprendizado de máquina têm sido amplamente utilizadas para abordar essas questões, mas sua eficácia depende da qualidade e diversidade dos dados de treinamento, que precisam incluir amostras representativas de diferentes origens geográficas e étnicas.

Além disso, variantes raras frequentemente têm efeitos mais pronunciados no fenótipo devido à sua menor exposição à seleção purificadora. Isso reforça a necessidade de métodos preditivos robustos que considerem não apenas aspectos funcionais, como a conservação evolutiva e a estrutura proteica, mas também o contexto populacional em que essas variantes ocorrem (Karczewski et.al., 2024). Logo, a integração de dados multiômicos tem se mostrado uma abordagem promissora para enriquecer essas análises, permitindo uma compreensão mais abrangente dos efeitos funcionais das variantes em diferentes níveis biológicos.

Por fim, a aplicação de preditores de patogenicidade para a análise de variantes em subpopulações não é apenas uma questão de avanço científico, mas também uma questão de equidade em saúde. Ferramentas que incorporam diversidade genética têm o potencial de reduzir disparidades, fornecendo diagnósticos e tratamentos

mais precisos para populações frequentemente negligenciadas. O investimento em iniciativas que promovam a representatividade genômica global, combinado com o desenvolvimento contínuo de preditores especializados, é essencial para alcançar os objetivos da medicina personalizada em escala global.

2. JUSTIFICATIVA

A fibrose cística (FC) é uma doença hereditária grave caracterizada pela disfunção da proteína CFTR (cystic fibrosis transmembrane conductance regulator), resultando em um desequilíbrio iônico que aumenta a viscosidade do muco produzido nas células epiteliais. Embora extensivamente estudada em populações de ascendência europeia, a FC e suas variantes associadas permanecem subexploradas em grupos sub-representados, como populações indígenas e latino-americanas. Isso gera lacunas significativas no entendimento da diversidade genética e das implicações clínicas do gene CFTR nesses grupos, limitando a aplicabilidade de ferramentas preditivas e abordagens terapêuticas baseadas em farmacogenômica (Fernandes et al., 2021).

No contexto das variantes do gene *CFTR*, as ferramentas preditivas como AlphaMissense, PolyPhen-2 e SIFT têm sido amplamente utilizadas para prever a patogenicidade de mutações *missense*. Apesar de seu potencial, esses preditores apresentam limitações, especialmente no caso de variantes raras, que são sub-representadas nos conjuntos de treinamento dessas ferramentas. Estudos recentes indicam que a integração dessas ferramentas por meio de modelos avançados, como Random Forest, não tem demonstrado melhora substancial na precisão das predições para variantes causadoras de FC (Siddik; Badal; Islam, 2024). Além disso, erros frequentes na classificação de variantes não causadoras de FC destacam a necessidade de desenvolver modelos especializados que incorporem dados específicos do CFTR, integrando informações funcionais e estruturais (Gunning et al., 2021). Tais falhas reforçam a importância de validações experimentais e abordagens complementares, como análises multiômicas, para melhorar a robustez das predições e minimizar os riscos de classificações incorretas.

Portanto, a integração das duas abordagens abordadas é essencial. Por um lado, a avaliação crítica do desempenho de preditores de patogenicidade sublinha a necessidade de melhorar as ferramentas existentes, considerando as limitações observadas em variantes raras e não causadoras de FC. Por outro, estudos voltados para populações sub-representadas revelam a importância de explorar a diversidade genética para enriquecer o contexto de aplicações clínicas (Rodrigues et al., 2022). A relação entre esses aspectos justifica este trabalho, cujo objetivo é abordar de forma integrada a análise da precisão dos preditores e a identificação de variantes específicas em populações sub-representadas, contribuindo para uma visão mais abrangente da genética

do *CFTR* e suas implicações clínicas. Essa abordagem multidimensional busca preencher lacunas críticas no conhecimento, promovendo uma base mais sólida para a interpretação de variantes genéticas e a tomada de decisões clínicas robustas.

3. OBJETIVOS

3.1 OBJETIVO GERAL

Investigar a variabilidade genética e a funcionalidade de variantes *missense* do gene *CFTR* em populações indígenas da Amazônia, avaliando seu impacto funcional por meio de ferramentas de predição, com foco em aplicações farmacogenéticas.

3.2 OBJETIVOS ESPECÍFICOS

- Caracterizar a variabilidade genética de variantes do gene *CFTR* em uma amostra de populações indígenas da Amazônia.
- Avaliar a funcionalidade e a patogenicidade de variantes *missense* do gene *CFTR* utilizando ferramentas de predição em exomas de populações indígenas da região Norte.
- Identificar o perfil farmacogenético associado às variantes do gene *CFTR* em populações indígenas, explorando potenciais implicações terapêuticas e de medicina personalizada.
- Comparar o desempenho e a acurácia preditiva de diferentes ferramentas (como SIFT, PolyPhen-2 e AlphaMissense) na classificação das variantes observadas, com foco na sensibilidade, especificidade e correlação com parâmetros estruturais e funcionais da proteína.

4. CAPÍTULO I

Este capítulo é referente ao manuscrito intitulado “Functional Impact of Variants of the *CFTR* Gene in Indigenous Populations of the Amazon” em preparação para submissão no periódico.

Functional Impact of Variants of the CFTR Gene in Indigenous Populations of the Amazon

Ana Katarina Campos Nunes ^{1†}, Camila Sinimbú Forte^{2†}, Arthur Felipe Vasconcelos Ferreira Reis ^{1†}, Jorge Santana³, André Ribeiro dos Santos¹, Sidney Santos^{1,5}, Ândrea Ribeiro-dos-Santos^{1*}, Gilderlanio Santana de Araújo^{1*}

†These authors contributed equally to this work.

***These authors contributed equally to this work and share last authorship.**

Short Title: *CFTR* Variants in Amazonian Indigenous Populations

Author affiliations:

¹ Laboratory of Human and Medical Genetics, Institute of Biological Science, Federal University of Pará, Belém, Pará, Brazil.

² Institute of Technology, Federal University of Pará, Belém, Pará, Brazil.

³ IMD - Instituto Metr pole Digital, Universidade Federal do Rio Grande do Norte

⁵NPO – N cleo de Pesquisas Oncol gicas, Universidade Federal do Par  (UFPA), Bel m (PA).

Correspondence to: Gilderlanio Santana de Ara jo

ORCID:

Gilderlanio Santana de Ara jo - 0000-0001-9199-9419

ABSTRACT: Cystic Fibrosis (CF) is caused by dysfunction of the CFTR protein, leading to ionic imbalance and thickened mucus in epithelial tissues. In this study, we analyzed exome data from indigenous Amazonian populations and compared CFTR gene variants with those listed in the CFTR2 and ABraOM public databases. The data underwent quality filtering, alignment to the GRCh38 reference genome, and variant identification. Variant pathogenicity was predicted using AlphaMissense, and functional and genetic epidemiological analyses were performed with SNP Nexus. We found only one variant in common with CFTR2 and four with ABraOM, underscoring the genetic

uniqueness of our sample. Four missense variants of clinical interest were identified based on AlphaMissense scores for the P13569 CFTR isoform: c.3728T>C and c.2932A>G showed high pathogenicity (0.87 and 0.89), while c.2588T>A and c.4045G>A had moderate scores (0.67 and 0.61). Structurally, c.3728T>C and c.4045G>A are located in the Nucleotide Binding Domain 2 (NBD2), while c.2932A>G and c.2588T>A are in Transmembrane Domain 2 (TMD2). These findings contribute to the understanding of *CFTR* genetic diversity in underrepresented populations and may inform future pharmacogenomic research in Indigenous and mixed Latin American groups.

Keywords: Amerindian population; *CFTR* gene; Genetic Variability; Cystic Fibrosis; Missense Variants.

INTRODUCTION

The *CFTR* gene encodes the CFTR protein, which consists of 1,480 amino acids. It is located on the long arm of chromosome 7, specifically in the 7q31 region, comprising 27 coding exons and spanning approximately 190 kb of human genomic DNA (Dickinson; Collaco, 2021). The CFTR protein is expressed in various tissues, including the lungs, pancreas, sweat glands, intestines, liver, nasal mucosa, salivary glands, and reproductive tract. This protein forms an ion channel that transports chloride ions. Its interaction with the epithelial sodium channel (ENaC) primarily regulates fluid secretion and absorption, although the complete mechanism has not yet been fully elucidated. Physiologically, CFTR function depends on phosphorylation of its regulatory (R) domain by protein kinase A. ATP binding opens the channel pores, allowing ion transport, and ATP hydrolysis at the catalytically active “consensus” site subsequently closes the channel (Levring et al., 2023).

When studying Indigenous populations, several genetic and demographic factors—such as limited population size, frequent consanguinity, genetic drift, and geographical isolation—contribute to their genetic uniqueness. However, the lack of genomic data for these groups poses a significant barrier to pharmacogenomic research (Cohen-Paes et al., 2022). Furthermore, settlement patterns and ancestral distribution are often heterogeneous and dispersed. The scarcity of genomic information

complicates efforts to develop a comprehensive pharmacogenetic panel tailored to the Brazilian population.

Due to this genetic distinctiveness, particularly in Indigenous populations from the Brazilian Amazon, potentially pathogenic variants may not be represented in clinically relevant databases such as ClinVar (Landrum et al., 2018) or pharmacogenomic resources like PharmGKB (Whirl-Carrillo et al., 2012). This highlights the underrepresentation of these populations in genetic studies. Investigating their genetic variability can therefore enhance our understanding of hereditary diseases such as cystic fibrosis (CF), their distribution among these groups, and their potential functional impacts on protein structure.

MATERIAL AND METHODS

EXOME FROM INDIGENOUS POPULATIONS

Samples were collected from adult individuals (18 to 50 years old) from the following indigenous populations: Arara/Arara do Iriri (ARA); Asurini do Koatinemo (AKW); Asurini do Trocará (AST); Araweté (ARW); Kayapó/Xikrin (KAY); Zo'é (ZOE); Wayãpy (WPI); and Awa-Guajá (AWA) (Figure 1). The exome data was pre-processed for quality filtering and stored by the Laboratory of Human and Medical Genetics (LGHM), and alignment was carried out with the GRCh38 genome and variant calling with GATK (v 4.0.0). The sequencing data generated here is available on the European Nucleotide Archive database (ENA) (<https://www.ebi.ac.uk/ena>) under the access number PRJEB35045 (Ribeiro-dos-Santos et al., 2020).

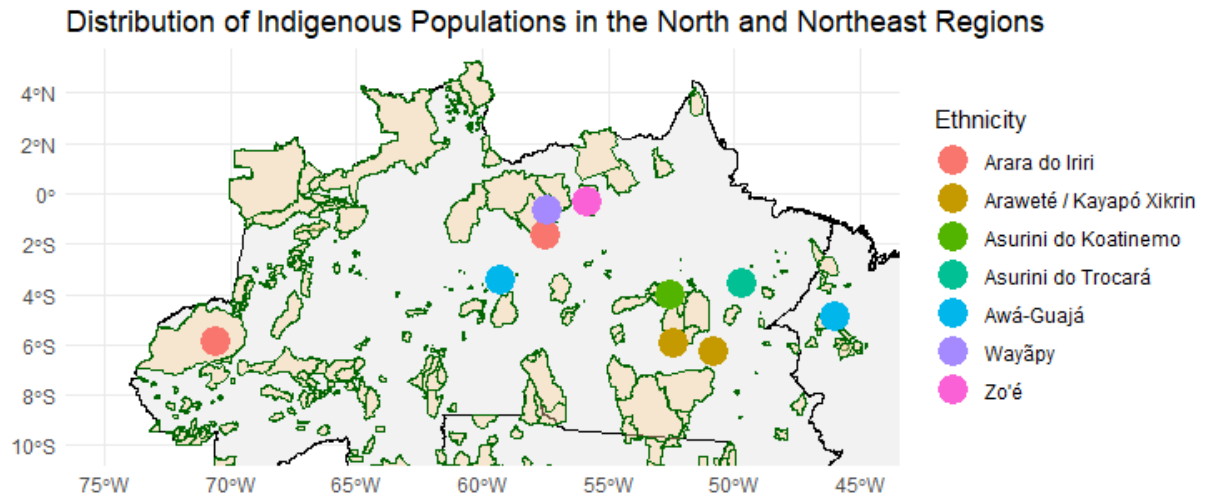


Figure 1. Location of the indigenous populations mentioned in the study: Arara (Irirí), Asurini (Koatinemo and Trocará), Araweté, Xikrin (Cateté / Bacajá), Zo'é, Wayãpy (Wajãpi) and Awá-Guajá. Source of geospatial data: FUNAI — Geoprocessing Coordination (Indigenous Lands and Villages layers), download on [date]. Map prepared by the author.

SAMPLE COLLECTION AND LIBRARY PREPARATION

As described by Ribeiro-dos-Santos et al. (2020), blood samples were obtained from participants using EDTA-containing vacutainer tubes. Genomic DNA was extracted with the phenol–chloroform method, quantified with a NanoDrop fluorometer (Thermo Fisher Scientific), and checked for integrity by 2% agarose gel electrophoresis.

Exome libraries were prepared with three commercial kits, according to the manufacturers' protocols: Nextera Rapid Capture Exome (Illumina), SureSelect Human All Exon V6 (Agilent Technologies), and TruSeq Exome (Illumina). Sequencing was performed on the Illumina NextSeq 550 platform in four runs with the NextSeq 500/550 High Output kit, processing approximately 16 samples per run (Ribeiro-dos-Santos et al., 2020).

READING, PROCESSING, AND CALLING VARIANTS

Following Barros et al. (2024), sequencing quality and coverage were first assessed with FASTQC. Reads of poor quality were trimmed with Trimmomatic v0.36 (Bolger; Lohse; Usadel, 2014), removing adapters, sequences shorter than 50 bp, and those with a PHRED score <20. Only pairs with both ends passing the criteria were kept for alignment.

The filtered reads were aligned to the GRCh38 genome using BWA-MEM v0.7.12, producing SAM files that were then sorted by genomic coordinates with SAMtools v1.7 (Li et al., 2009). Duplicate reads were marked using Picard-tools MarkDuplicates v2.27.3. Subsequently, BAM files were generated and subjected to local realignment around indels with GATK RealignerTargetCreator and IndelRealigner. To correct for systematic sequencing errors, base quality scores were recalibrated with GATK BaseRecalibrator.

Variant calling was then performed with DeepVariant v1.15.0, applying the Whole Exome Sequencing (WES) model. This pipeline identifies and genotypes positions differing from the GRCh38 reference (excluding ALT contigs).

The resulting variants were annotated with SnpEff v5.1 and dbNSFP v4.2. Clinical annotations were complemented with ClinVar, and population allele frequencies were checked in ExAC and gnomAD. Detailed sequencing metrics are presented in Supplementary File 1.

For pathogenicity classification, ACMG/AMP guidelines were followed, combined with predictions from FATHMM, MutationAssessor, MutationTaster, SIFT, MetaSVM, and PROVEAN. Variants considered deleterious by at least two predictors and confirmed in ClinVar were retained for downstream analyses.

SAMPLES FROM PUBLIC DATABASES

CFTR2

The Clinical and Functional Translation of CFTR (CFTR2) (<https://cftr2.org/>) provides curated information on CFTR variants and their clinical interpretation in cystic fibrosis. We used the most recent dataset (April 7, 2023), which includes 804 annotated variants from 89,052 patients (Castellani, 2013).

BRAZILIAN ONLINE MUTATION ARCHIVE

To search for overlap with CFTR2 variants, we also consulted the ABraOM database (<https://abraom.ib.usp.br/search.php>), which contains data on 3,608 variants from 1,171 unrelated elderly individuals from São Paulo, sequenced under the Health, Well-being and Ageing (SABE) study (Naslavsky et al., 2022).

DATA PROCESSING

To extract variants from the CFTR gene, we used VCFTools v0.1.16, filtering the genomic region corresponding to CFTR (chr7:117287120–117715971) based on gnomAD v4.1.0 coordinates. Hardy–Weinberg Equilibrium (HWE) was applied ($p = 0.01$) to exclude variants under selective pressure (Hao; Storey, 2019), and missing data were filtered with the “max-missing 0.9” parameter. Allele frequencies were calculated with PLINK v1.9, and plots were generated with the ggpubr and ggarrange packages in R.

After processing the data, we performed the allele frequency using Plink (v1.9). The “ggpubr” and “ggarrange” packages were used to plot the graphs.

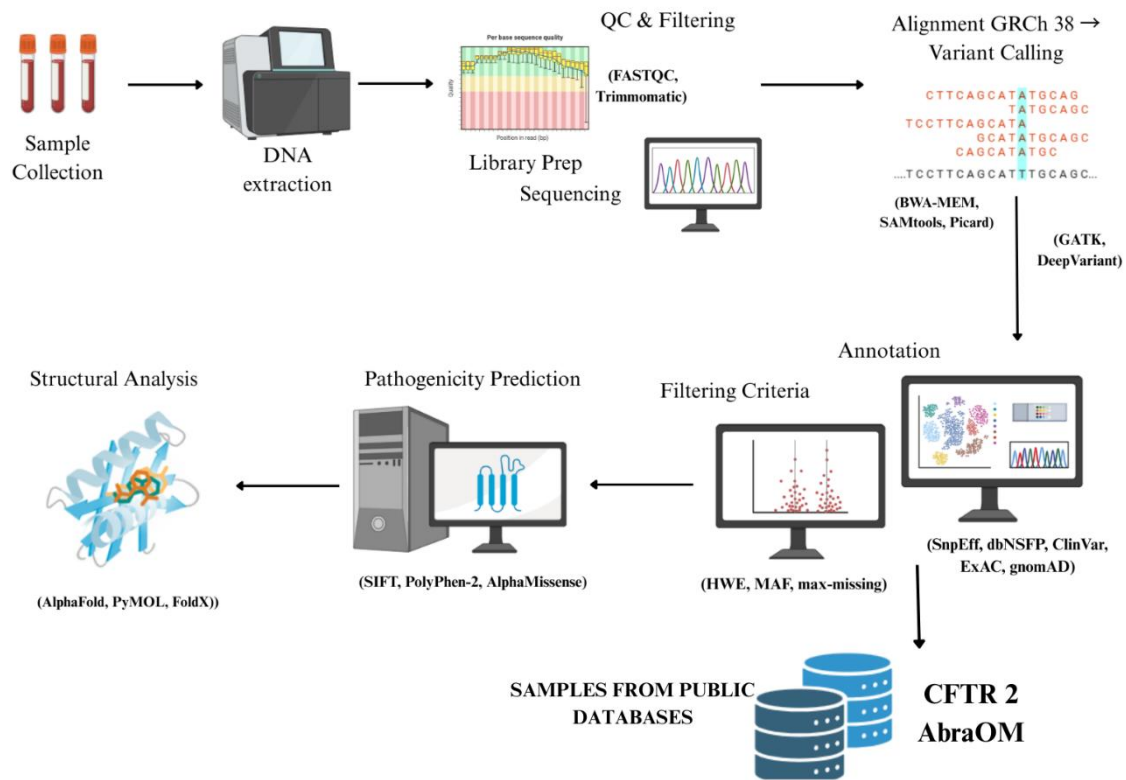


Figure 1 - Illustration of summarized methodology.

GENOMIC ANNOTATION

SNPNexus

Variants with $MAF > 0.05$ were annotated in SNPNexus (<https://www.snpx.org/v4/>), using the GRCh38/hg38 reference, Ensembl Variation 95, dbSNP 151, PolyPhen-2, and ClinVar (Oscanoa et al., 2020).

The Variant Effect Predictor (VEP)

Pathogenicity predictions were generated with the Ensembl Variant Effect Predictor (v112) using the AlphaMissense plugin.

For structural analyses, we used the AlphaFold model corresponding to the canonical CFTR isoform (UniProt ID: P13569-1, AlphaFold ID: AF-P13569-F1). The canonical isoform was chosen because it represents the full-length, clinically relevant protein sequence, which is consistently used in functional and clinical databases such as ClinVar and CFTR2. This choice ensures comparability of structural and functional interpretations across datasets.

The protein model was visualized in PyMOL v3.0.2, and surface coloring was performed with the coloram.py plugin (<https://alphamissense.hegelab.org/coloram.py>), mapping AlphaMissense pathogenicity scores onto the B-factors of the PDB file. Residues were classified as: 0–0.34 (blue, likely benign), 0.34–0.564 (colorless, ambiguous), 0.564–0.78 (light red, likely pathogenic), and 0.78–1.0 (dark red, pathogenic). Missense variants of interest were specifically highlighted on the model.

RESULTS

DATA PROCESSING AND ANALYSIS POPULATION GENETICS

A total of 309 variants were initially identified in the CFTR gene through the analysis of 110 individuals from indigenous populations in the Amazon. After excluding missing data (NA), the dataset was reduced to 238 variants with valid genotypic information.

Most variants (216) were observed only once in the entire cohort, with each detected in a single distinct individual. These were therefore classified as singletons. The classification was based on the number of individuals carrying each variant, regardless of the minor allele frequency (MAF). After removing variants with missing data in the reference allele, 205 singletons remained. Of these, 199 had an MAF of 0.0 (absence of the alternative allele, indicating homozygosity for the reference allele) (Supplementary Table 4), and 17 had an MAF of 0.5, consistent with singleton heterozygous genotypes (Table 1).

Since variants found in only one individual may have artificially inflated MAF values—especially in small sample sizes—all singletons were excluded from further analysis, regardless of their allele frequency. This filtering step aimed to minimize the impact of sequencing artifacts, low coverage, or sampling noise, which can compromise interpretation at the population level.

After this filtering, 21 variants remained in the dataset because they were observed in more than one individual (Table 3). Among them, 10 variants had a MAF > 1% and were classified as common variants (Table 2).

Although variants with MAF = 0.0 were numerically predominant, this value indicates the absence of the alternative allele in all individuals analyzed, reflecting

homozygosity for the reference allele. Therefore, these variants do not segregate in the sample and were not considered in the final classification of rare or common variants.

It is important to note that, in all analytical steps, variants that did not have the reference allele were excluded, as such cases represent missing or inconsistent data for allele frequency analysis.

Position	Ref(A1)	Alt(A2)	MAF	N_CHR	Individuals
117297583	A	G	0,5	2	1
117363427	T	A	0,5	2	1
117380104	G	C	0,5	2	1
117380116	G	A	0,5	2	1
117381177	G	A	0,5	2	1
117383007	T	C	0,5	2	1
117479942	T	A	0,5	2	1
117531141	G	GGC	0,5	2	1
117536684	T	C	0,5	2	1
117559655	T	TGG	0,5	2	1
117594980	A	C	0,5	2	1
117603748	A	C	0,5	2	1
117606754	G	A	0,5	2	1
117610677	AGT	A	0,5	2	1
117642590	A	C	0,5	2	1
117666815	T	G	0,5	2	1
117668519	A	C	0,5	2	1

Table 1 – Singleton variants identified in the *CFTR* gene in a sample of 110 indigenous individuals from the Amazon. Each variant presented was observed in only one individual from the cohort. The position refers to the genomic location on chromosome 7; “Ref (A1)” represents the reference allele and “Alt (A2)” the alternative allele. MAF (Minor Allele Frequency) indicates the frequency of the alternative allele in the total sample, and N_CHR represents the total number of chromosomes analyzed. The “Individuals” column indicates the number of individuals carrying the variant, set at 1 for all cases.

Position	Ref(A1)	Alt(A2)	MAF	N	CHR	Individuals
117363485	T	A	0,230769	78		39
117363641	G	T	0,333333	6		3
117380100	T	A	0,5	10		5
117426986	T	C	0,423077	26		13
117536514	AGAT	TA	0,315789	152		76
117548628	ATG	ATGTG	0,25	4		2
117559422	CTT	CT	0,5	10		5
117652642	C	T	0,4	20		10
117666997	A	C	0,5	4		2
117668359	T	A	0,5	4		2

Table 2 - Table of common variants - Common variants identified in the CFTR gene among indigenous individuals from the Amazon. Variants with an allele frequency (MAF) greater than 1% and observed in more than one individual, after exclusion of singletons, were considered “common.”

Position	Ref(A1)	Ref(A2)	MAF	N	CHR	Individuals
117315116	A	ATCCTT	0	24		12
117323098	G	T	0	4		2
117323255	G	T	0	4		2
117363485	T	A	0,230769	78		39
117363641	G	T	0,333333	6		3
117380100	T	A	0,5	10		5
117426809	A	C	0	4		2
117426986	T	C	0,423077	26		13
117427487	C	A	0	4		2
117479956	C	A	0	6		3
117536514	AGAT	TA	0,315789	152		76
117548628	ATG	ATGTG	0,25	4		2
117548682	T	G	0	72		36
117548786	A	C	0	22		11
117548796	G	T	0	4		2
117559422	CTT	CT	0,5	10		5
117610489	G	A	0	6		3
117652642	C	T	0,4	20		10
117666997	A	C	0,5	4		2
117668359	T	A	0,5	4		2
117711627	A	G	0	10		5

Table 3 - *CFTR* gene variants observed in more than one indigenous individual from the Amazon. The selection includes all variants shared by at least two individuals, regardless of minor allele frequency (MAF).

DATA ANALYSIS

We analyzed the data collected separately in order to identify classifications and total numbers of variants. For each result, we listed the number of variants found in tables according to the parameters provided by each database. For CFTR2, the variants were divided according to “CF-causing”, “Non-CF Causing”, “Unknown significance” and “Varying Clinical Consequence”.

Using the sample of indigenous Amazonian populations, we searched for common variants that could be present in CFTR2. We found only one variant of the *missense* type (p.Ser912Leu / rs121909034), which was of uncertain significance according to CFTR2, but was classified as benign according to the Alpha Missense pathogenicity score. We also checked for variants in the ABraOM sample that could be found in our data on Amerindians, thus finding three variants in common (Table 3).

Position	Variant	Classification	Frequency ABraOM	Frequency Indigenous	N_Individuals
117642590	rs1800130	synonymous_variant	0,050811	0,5	1
117652642	rs214164	intron_variant	0,18702	0,4	10
117668359	rs1042180	3_prime_UTR_variant	0,218617	0,5	2

Table 3 - Variants in common between ABraOM and the sample of Indigenous Amazonians

GENOMIC ANNOTATION USING SNPNEXUS

After processing the variants in SNP Nexus, of the 27 indigenous variants, we observed nine distributed variants in terms of clinical functionality, the majority of which were benign in intronic regions. We found only one pathogenic variant (MAF = 0.5) related to cystic fibrosis and pancreatic insufficiency (PI) located in the intronic region, at a splice site in only 1 individual. According to the Clinical and Functional Translation of CFTR (CFTR2), the variant identified as c.2988+1G>A causes both CF and PI when associated with another variant that also causes the same phenotypes (Table 4).

Position	Variation	Clinical Significance	Predicted Function
117536684	C/T	Benign/Likely benign	Intronic
117559422	G/A	Benign	Intronic
117559655	G/A	Benign/Likely benign	Coding
117606754	G/A	Pathogenic	Intronic (Splice Site)
117610677	A/G	Conflicting interpretations of pathogenicity	Intronic
117642590	A/G	Benign/Likely benign	Coding
117652642	G/A	Benign	Intronic
117666997	C/T	Conflicting interpretations of pathogenicity	Coding
117668359	CT	Likely Benign	3 UTR

Table 4 - Characterization of classic and variant functions by SNP Nexus.

Despite the prediction of localization, the literature already reports the presence of intronic variants and how these are responsible for causing serious diseases. Since non-coding region variants that cause perturbations in non-coding elements cause severe diseases, thus affecting *splicing*, transcription, translation, RNA processing and stability and chromatin interactions (Ellingford, 2022) .

VARIANTS IDENTIFIED BY ALPHAMISSENSE

We found four singletons variants of the missense type of clinical interest according to the alphamissense pathogenicity prediction score for the P13569 isoform of the CFTR. Two variants, c.3728T>C and c.2932A>G, had high pathogenicity scores, 0.87 and 0.89, respectively, while c.2588T>A and c.4045G>A had moderate scores, 0.67 and 0.61. The high pathogenicity *score* and the localization in the protein domains are relevant, as evidenced both by the frequencies observed in gnomAD and in the sample itself, verifying the rarity of two of these variants and that they, with specific changes in specific amino acids, may be exclusive to indigenous peoples (Figure 2).

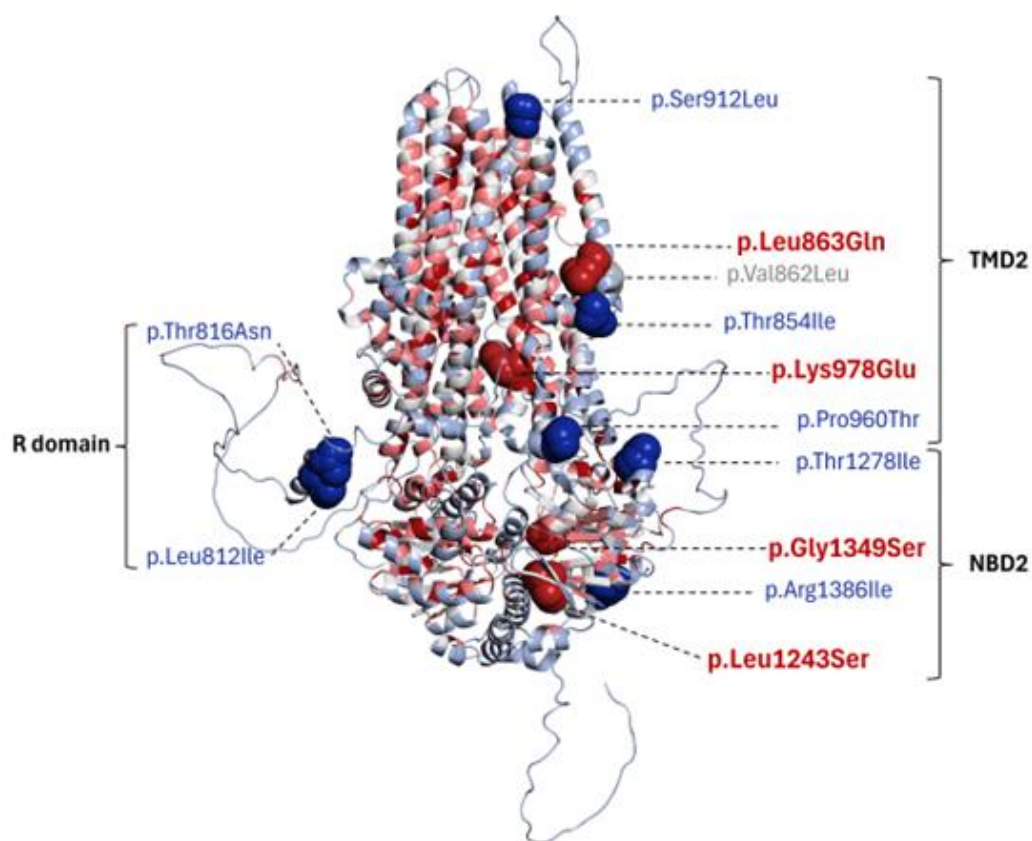


Figure 2 – AlphaMissense predictions of possible missense variants mapped onto the three-dimensional structure of the human CFTR protein, as predicted by AlphaFold. The structural model corresponds to the canonical isoform (UniProt ID: P13569-1), which is considered the principal functional form of CFTR. Predicted pathogenicity scores are represented on the canonical protein structure, highlighting regions more susceptible to destabilization or functional impairment.

DISCUSSION

POPULATION GENETIC ANALYSIS

Proteins like CFTR are essential for biological processes, and alterations in their structure can impair functions such as ion transport. The location of CFTR gene variants directly influences their pathogenic potential. According to Gerasimavicius, Livesey & Marsh (2023), although mutations with structural impact have been identified, their phenotypic consequences are often undercharacterized.

The identification of rare variants—particularly those absent from major databases such as ClinVar and dbSNP—may have direct implications for diagnosis, pharmacogenetics, and the study of diseases like cystic fibrosis in specific populations. Variants with minor allele frequencies (MAF) between 0.2% and 0.4% exemplify the importance of examining rare alleles in underrepresented groups, such as Indigenous populations (Oddsson et al., 2023). The presence of rs34906874, although classified as benign, highlights that intronic variants documented in clinical databases (ClinVar) can be present in asymptomatic individuals, underscoring the necessity for careful interpretation grounded in both population-specific and functional contexts.

The use of three different kits for preparing exome libraries (Nextera Rapid Capture Exome, SureSelect Human All Exon V6, and TruSeq Exome) was a methodological factor that likely had a significant influence on the results obtained. Each kit has particularities in terms of probe design, capture efficiency, and coverage of target regions, which can result in differences in the number of variants detected, the average reading depth, and the representation of critical genomic regions, such as CFTR gene exons (Li et al., 2023). This methodological variability may explain discrepancies in the detection rate of rare variants or in regions that are difficult to capture, directly reflecting on the robustness of subsequent analyses (López et al., 2025). Therefore, the adoption of multiple kits not only expanded the scope of sequencing but also brought the need for greater caution in comparing detected variants, highlighting the influence of experimental design on genetic interpretation, especially in studies focused on populations in the Northern Region of Brazil.

Importantly, the absence of annotation in reference databases does not necessarily indicate functional neutrality. On the contrary, as emphasized by Petrovski *et al.* (2022), ultra-rare variants are frequently associated with deleterious effects and are enriched in genes subject to purifying selection—particularly when located in conserved coding regions or within critical protein domains.

The rs1432807327 variant, though not clinically annotated, has emerged in studies investigating the role of non-pathogenic CFTR alleles in modulating respiratory phenotypes. This suggests a potential gene–environment interaction, which may be particularly relevant in populations with unique environmental exposures or in complex conditions such as chronic obstructive pulmonary disease (COPD) (Saferali *et al.*, 2022).

In the case of the 211 variants with MAF = 0.0, the absence of heterozygous and homozygous individuals for the alternative allele may reflect negative selective pressure, where the presence of the variant is lethal or highly deleterious, preventing its propagation in the population. Alternatively, technical calling errors or low sequencing depth cannot be ruled out, which highlights the need for validation through complementary methods such as Sanger sequencing or reanalysis using higher-coverage data (Karczewski, [S.d.]).

Given the specific population context of this study, it is plausible that some of these variants are associated with local adaptive processes or genetic drift events, especially in subgroups with reproductive isolation or a history of endogamy. This scenario reinforces the importance of expanding genomic databases with information from underrepresented populations — such as Indigenous peoples, people of African descent, and traditional communities from the Amazon — to enable fairer and more accurate interpretation of genetic variants (Sirugo; Williams; Tishkoff, 2019).

Other variants identified in this study, lacking functional or clinical annotations, may represent neutral polymorphisms or novel loci yet to be functionally characterized. Collectively, these findings underscore the relevance of population-based studies in broadening our understanding of *CFTR* genomic diversity—particularly within underrepresented populations such as Indigenous groups of the Amazon region.

We initially consulted the CFTR2 and ABraOM public databases. We found 46 shared variants, but only one was classified as pathogenic (rs12908794). When compared with data from Amazonian Indigenous individuals, only one variant overlapped with CFTR2 and four with ABraOM, indicating strong genetic uniqueness in this population.

Of the nine variants identified in our sample with SNPnexus, six were benign, two had uncertain significance, and one was pathogenic associated with cystic fibrosis and pancreatic insufficiency. The intronic variant c.2988+1G>A may affect splicing and influence response to pharmacological modulators (Deletang, 2022; Joynt; Cutting; Sharma, 2022).

This genetic variability reflects Indigenous ancestry, shaped by relationships with Asian populations and later admixture with Europeans. Such heterogeneity emphasizes the importance of personalized pharmacogenomics for underrepresented populations (Roden et al., 2019; Rodrigues et al., 2022).

Four exclusive missense variants showed high pathogenic potential, affecting domains like NBDs and TMD2, suggesting impaired chloride channel function. Even with common alleles like f508del, dosage adjustments are needed to avoid toxicity (Luzum et al., 2021).

Although some variants appeared in single individuals, their potential effects demand attention. These findings highlight the need to expand Indigenous representation in genomic databases and tailor pharmacotherapy to their unique genetic backgrounds.

Results of Alpha Missense

Current genetic panels are largely based on European populations, underrepresenting groups like Latinos and Indigenous peoples. Studying these populations helps clarify genetic and pharmacogenetic factors associated with disease susceptibility and treatment response. Rare variants and high genetic diversity in Indigenous populations remain poorly represented in precision medicine efforts (Rodrigues et al., 2022).

In our study, two missense variants were found in the NBD2 domain: c.3728T>C (p.Leu1243Ser) and c.4045G>A (p.Gly1349Ser). Both substitutions increase polarity and hydrophilicity, potentially affecting CFTR function through altered hydrogen bonding and steric interactions. G1349S has been previously described in Japanese patients with congenital absence of the vas deferens (Anzai et al., 2003), and shows low frequency in East Asians (gnomAD EAS: 0.000218), consistent with the East Asian ancestry of Native Americans (Ribeiro-dos-Santos et al., 2020; Yu et al., 2020).

Two additional missense variants were found in TMD2: c.2588T>A (p.Leu863Gln) and c.2932A>G (p.Lys978Glu). These involve substitutions to polar or charged residues, possibly destabilizing alpha-helices or disrupting the chloride channel environment through electrostatic repulsion, potentially impairing ion transport and protein folding.

All four variants, identified via AlphaMissense scores, likely alter intramolecular interactions and folding, supporting their potential pathogenicity (Tordai et al. 2024). Mutations in NBDs can impair ATP-driven channel activation, consistent with CFTR class III mutations (Farinha; Callebaut, 2022).

Though CF is most prevalent in European populations, Indigenous populations may harbor unique CFTR variants not captured by standard panels. De Azevedo et al. (2023) noted the lack of reliable prevalence estimates in Native groups. Moreover, CF in Black and Indigenous patients often presents worse outcomes due to rare, undetected variants (Wu et al., 2024), underscoring the urgency of population-specific genomic research (Rodrigues et al., 2022).

CONCLUSION

In view of the above, we found that despite the low occurrence of cystic fibrosis in indigenous Brazilian populations, there are specific pathogenic variants that can directly interfere with the response to existing therapeutic modulators. We also highlight the need for more detailed clinical studies on the effect of these variants, both splicing and *missense* on the protein structure of the CFTR in order to expand the

pharmacogenetic information of these groups with the aim of formulating personalized strategies for treating the disease.

Acknowledgements

We recognize the relevance of studies that expand genetic understanding in groups historically underrepresented in biomedical research. This investigation reinforces the importance of considering specific variants that may directly impact therapeutic efficacy in indigenous patients with cystic fibrosis. We thank the communities involved and the scientific collaborators who made this work possible. The results presented here serve as a basis for the development of more inclusive and personalized clinical approaches, contributing to health equity and the advancement of pharmacogenetics in Brazil.

Declarations

Ethics approval

The study was reviewed and approved by Brazilian National Committee on Research Ethics – CONEP (identified by Nos. 1062/2006 and 123/98). All participants signed a free-informed consent as well as the tribe leaders and when necessary a translator explained the project and the importance of the research.

Availability of data and materials

The data obtained from the public domain are available at the European Nucleotide Archive (ENA <https://www.ebi.ac.uk/ena>) under the accession numbers: PRJEB9586, PRJNA393593, PRJEB24629, PRJNA229448, PRJEB25445, PRJEB29074, PRJEB28961, and PRJEB12437 and the sequencing data generated here are available at the ENA database under the accession number PRJEB35045.

Competing interests

The authors report no competing interests.

Funding

This research was funded by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq); Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) – Biocomputacional Protocol no. 3381/2013/CAPES (Rede de Pesquisa em Genômica Populacional Humana), and Pró-Reitoria de Pesquisa e Pós-Graduação da Universidade Federal do Pará (PROPESP/UFPA). A.R.S. was supported by CNPq/Productivity (312916/2021-3). The funders had no role in study design, data collection, analysis, interpretation or writing of the report.

Author Contributions

A.K.C.N was responsible for data processing, methodology, results and writing of the discussion. C.S.F participated in data processing. A.F.V.F.R participated in the methodology and analysis of the variants found and in the discussion. J.S, A.R.S, S.S and A.R.S developed the analysis and availability of data. G.S.A wrote, analyzed data, structured and designed the project and developed the article.

Supplementary Materials

Table S1 – Rare CFTR variants identified in Amazonian Indigenous individuals. List of rare missense and non-coding variants found in the CFTR gene across Indigenous samples, annotated with genomic position, variant type, and predicted pathogenicity.

Table S2 – Minor allele frequencies (MAF) of CFTR variants in Indigenous populations. Distribution of allele frequencies for CFTR variants detected in the study sample, including comparison with reference populations (e.g., gnomAD, ABraOM) when available.

REFERENCES

- ABDELLAOUI, Abdel *et al.* 15 years of GWAS discovery: Realizing the promise. **The American Journal of Human Genetics**, v. 110, n. 2, p. 179–194, fev. 2023.
- ADZHUBEI, Ivan; JORDAN, Daniel M.; SUNYAEV, Shamil R. Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2. **Current Protocols in Human Genetics**, v. 76, n. 1, jan. 2013.
- ANZAI, Chieko *et al.* CFTR gene mutations in Japanese individuals with congenital bilateral absence of the vas deferens. **Journal of Cystic Fibrosis**, v. 2, n. 1, p. 14–18, mar. 2003.
- BACALHAU, Mafalda *et al.* Elexacaftor-Tezacaftor-Ivacaftor: A Life-Changing Triple Combination of CFTR Modulator Drugs for Cystic Fibrosis. **Pharmaceuticals**, v. 16, n. 3, p. 410, 8 mar. 2023.
- BARREÑADA, Lasai *et al.* Understanding overfitting in random forest for probability estimation: a visualization and simulation study. **Diagnostic and Prognostic Research**, v. 8, n. 1, p. 14, 27 set. 2024.
- BARROS, Maria Clara *et al.* Unraveling the protective genetic architecture of COVID-19 in the Brazilian Amazon. **Scientific Reports**, v. 14, n. 1, 9 nov. 2024.
- BAYNAM, Gareth *et al.* Indigenous Genetics and Rare Diseases: Harmony, Diversity and Equity. In: POSADA DE LA PAZ, Manuel; TARUSCIO, Domenica; GROFT, Stephen C. (Orgs.). **Rare Diseases Epidemiology: Update and Overview**. Advances in Experimental Medicine and Biology. Cham: Springer International Publishing, 2017. v. 1031 p. 511–520.
- BEZERRA, Vanessa Lima *et al.* ANÁLISE DA EPIDEMIOLOGIA DA FIBROSE CÍSTICA E A APLICAÇÃO DO DIAGNÓSTICO MOLECULAR NO BRASIL DE 2009 A 2020. **Revista Interfaces: Saúde, Humanas e Tecnologia**, v. 11, n. 4, 31 dez. 2023.
- BIENVENU, Thierry; LOPEZ, Maureen; GIRODON, Emmanuelle. Molecular Diagnosis and Genetic Counseling of Cystic Fibrosis and Related Disorders: New Challenges. **Genes**, v. 11, n. 6, p. 619, 4 jun. 2020.
- BOLGER, Anthony M.; LOHSE, Marc; USADEL, Bjoern. Trimmomatic: a flexible trimmer for Illumina sequence data. **Bioinformatics**, v. 30, n. 15, p. 2114–2120, 1 ago. 2014.
- BONETTI, Emanuele; TINI, Giulia; MAZZARELLA, Luca. Accuracy of renovo predictions on variants reclassified over time. **Journal of Translational Medicine**, v. 22, n. 1, p. 713, 31 jul. 2024.
- BORKENHAGEN, Beatrice; PREHM, Peter. Recovery of $\Delta F508$ -CFTR Function by Citrate. **Nutrients**, v. 14, n. 20, p. 4283, 14 out. 2022.

- BUSS, Oliver; RUDAT, Jens; OCHSENREITHER, Katrin. FoldX as Protein Engineering Tool: Better Than Random Based Approaches? **Computational and Structural Biotechnology Journal**, v. 16, p. 25–33, 2018.
- CASTELLANI, Carlo. CFTR2: How will it help care? **Paediatric Respiratory Reviews**, Royal Society of Medicine – The 26th symposium: Cystic fibrosis in children and adults, 20th November 2012. v. 14, p. 2–5, 1 maio 2013.
- CHARILAOU, Paris; BATTAT, Robert. Machine learning models and over-fitting considerations. **World Journal of Gastroenterology**, v. 28, n. 5, p. 605–607, 7 fev. 2022.
- CHENG, Jun *et al.* Accurate proteome-wide missense variant effect prediction with AlphaMissense. **Science**, v. 381, n. 6664, p. eadg7492, 22 set. 2023.
- CLANCY, J. P. *et al.* Clinical Pharmacogenetics Implementation Consortium (CPIC) Guidelines for Ivacaftor Therapy in the Context of CFTR Genotype. **Clinical Pharmacology & Therapeutics**, v. 95, n. 6, p. 592–597, jun. 2014.
- CROMWELL, Elizabeth A. *et al.* Cystic fibrosis prevalence in the United States and participation in the Cystic Fibrosis Foundation Patient Registry in 2020. **Journal of Cystic Fibrosis**, v. 22, n. 3, p. 436–442, maio 2023.
- CURTIS, David. Assessment of ability of AlphaMissense to identify variants affecting susceptibility to common disease. **European Journal of Human Genetics**, v. 32, n. 11, p. 1419–1427, nov. 2024.
- DELETANG, Karine. Splicing mutations in the CFTR gene as therapeutic targets. **Gene Therapy**, 2022.
- DELGADO, Javier *et al.* FoldX force field revisited, an improved version. **Bioinformatics**, v. 41, n. 2, p. btaf064, 4 fev. 2025.
- DICKINSON, Kimberly M.; COLLACO, Joseph M. Cystic Fibrosis. **Pediatrics In Review**, v. 42, n. 2, p. 55–67, fev. 2021.
- DUCKERS, Jamie *et al.* Real-World Outcomes of Ivacaftor Treatment in People with Cystic Fibrosis: A Systematic Review. **Journal of Clinical Medicine**, v. 10, n. 7, p. 1527, 6 abr. 2021.
- ELLINGFORD, Jamie M. Recommendations for clinical interpretation of variants found in non-coding regions of the genome. 2022.
- ERNST, Corinna *et al.* Performance of in silico prediction tools for the classification of rare BRCA1/2 missense variants in clinical diagnostics. **BMC Medical Genomics**, v. 11, n. 1, dez. 2018.
- FARINHA, Carlos M.; CALLEBAUT, Isabelle. Molecular mechanisms of cystic fibrosis – how mutations lead to misfunction and guide therapy. **Bioscience Reports**, 2022.

FATUMO, Segun; CHIKOWORE, Tinashe; KUCHENBAECKER, Karoline. Editorial: Genetics of Complex Traits and Diseases From Under-Represented Populations. **Frontiers in Genetics**, v. 12, p. 817683, 17 jan. 2022.

FERNANDES, Marianne Rodrigues *et al.* Genetic Diversity of Drug-Related Genes in Native Americans of the Brazilian Amazon. **Pharmacogenomics and Personalized Medicine**, v. Volume 14, p. 117–133, jan. 2021.

GARCIA, Felipe Antonio De Oliveira; ANDRADE, Edilene Santos De; PALMERO, Edenir Inez. Insights on variant analysis in silico tools for pathogenicity prediction. **Frontiers in Genetics**, v. 13, p. 1010327, 29 nov. 2022.

GERASIMAVICIUS, Lukas; LIVESEY, Benjamin J.; MARSH, Joseph A. Loss-of-function, gain-of-function and dominant-negative mutations have profoundly different effects on protein structure. **Nature Communications**, v. 13, n. 1, p. 3895, 6 jul. 2022.

GRIMM, Dominik G. *et al.* The Evaluation of Tools Used to Predict the Impact of Missense Variants Is Hindered by Two Types of Circularity. **Human Mutation**, v. 36, n. 5, p. 513–523, maio 2015.

GUNNING, Adam C. *et al.* Assessing performance of pathogenicity predictors using clinically relevant variant datasets. **Journal of Medical Genetics**, v. 58, n. 8, p. 547–555, ago. 2021.

HAO, Wei; STOREY, John D. Extending Tests of Hardy–Weinberg Equilibrium to Structured Populations. ago. 2019.

HAREL, Tom *et al.* Predicting Phenotypic Diversity from Molecular and Genetic Data. **Genetics**, v. 213, n. 1, p. 297–311, 1 set. 2019.

HASIAK SANTO¹, Augusto; RIBEIRO FERREIRA DA SILVA-FILHO^{2,3}, Luiz Vicente. Cystic fibrosis-related mortality trends in Brazil for the 1999-2017 period: a multiple-cause-of-death study. **Jornal Brasileiro de Pneumologia**, p. e20200166, 30 abr. 2021.

JÄNES, Jürgen *et al.* **Predicted mechanistic impacts of human protein missense variants**. Cold Spring Harbor Laboratory, , 29 maio 2024. Disponível em: <<http://biorxiv.org/lookup/doi/10.1101/2024.05.29.596373>>. Acesso em: 19 jul. 2025

JIA, Shijing; TAYLOR-COUSAR, Jennifer L. Cystic Fibrosis Modulator Therapies. **Annual Review of Medicine**, v. 74, n. 1, p. 413–426, 27 jan. 2023.

JOYNT, Anya T.; CUTTING, Garry R.; SHARMA, Neeraj. Genetics of Cystic Fibrosis. **Clinics in Chest Medicine**, v. 43, n. 4, p. 591–602, dez. 2022.

JUMPER, John *et al.* Highly accurate protein structure prediction with AlphaFold. **Nature**, v. 596, n. 7873, p. 583–589, 26 ago. 2021.

KARCZEWSKI, Konrad J. The mutational constraint spectrum quantified from variation in 141,456 humans. *[S.d.]*.

KARIM, Md Rezaul *et al.* Explainable AI for Bioinformatics: Methods, Tools and Applications. **Briefings in Bioinformatics**, v. 24, n. 5, p. bbad236, 20 set. 2023.

- KEHDY, Fernanda S. G. *et al.* Origin and dynamics of admixture in Brazilians and its effect on the pattern of deleterious mutations. **Proceedings of the National Academy of Sciences**, v. 112, n. 28, p. 8696–8701, 14 jul. 2015.
- KUMAR, Prateek; HENIKOFF, Steven; NG, Pauline C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. **Nature Protocols**, v. 4, n. 7, p. 1073–1081, jul. 2009.
- LANDRUM, Melissa J. *et al.* ClinVar: improving access to variant interpretations and supporting evidence. **Nucleic Acids Research**, v. 46, n. D1, p. D1062–D1067, 4 jan. 2018.
- LAUSCHKE, Volker M.; INGELMAN-SUNDBERG, Magnus. Emerging strategies to bridge the gap between pharmacogenomic research and its clinical implementation. **npj Genomic Medicine**, v. 5, n. 1, p. 9, 5 mar. 2020.
- LEVRING, Jesper *et al.* CFTR function, pathology and pharmacology at single-molecule resolution. **Nature**, v. 616, n. 7957, p. 606–614, 20 abr. 2023.
- LI, Heng *et al.* The Sequence Alignment/Map format and SAMtools. **Bioinformatics**, v. 25, n. 16, p. 2078–2079, 15 ago. 2009.
- LI, W.; LIU, H.; ZHANG, H.; WANG, J. Performance comparison of whole-exome sequencing kits for clinical diagnostics. *Human Genomics*, v. 17, n. 1, p. 1-12, 2023. doi: 10.1186/s40246-023-00482-1.
- LICHTER, David José Duque *et al.* Fibrose cística: uma revisão. **Brazilian Journal of Health Review**, v. 7, n. 5, p. 73568, 11 out. 2024.
- LIVESEY, Benjamin J.; MARSH, Joseph A. Updated benchmarking of variant effect predictors using deep mutational scanning. **Molecular Systems Biology**, v. 19, n. 8, p. e11474, 8 ago. 2023.
- LOPES-PACHECO, Miquéias. CFTR Modulators: The Changing Face of Cystic Fibrosis in the Era of Precision Medicine. **Frontiers in Pharmacology**, v. 10, p. 1662, 21 fev. 2020.
- LÓPEZ-VALDEZ, Jaime A. *et al.* Cystic fibrosis: current concepts. **Boletín Médico del Hospital Infantil de México**, v. 78, n. 6, p. 6536, 20 dez. 2021.
- LÓPEZ, F. V.; ASHTON, J. J.; CHENG, G.; ENNIS, S. A systematic analysis of contemporary whole exome sequencing capture kits to optimise high-coverage capture of CCDS regions. *NAR Genomics and Bioinformatics*, v. 7, n. 3, 1 set. 2025. doi: 10.1093/nargab/lqaf115.
- LUZUM, Jasmine A. *et al.* Moving Pharmacogenetics Into Practice: It's All About the Evidence! **Clinical Pharmacology & Therapeutics**, v. 110, n. 3, p. 649–661, set. 2021.
- MALAQUIAS, Soraya Beatriz Paula; CARDOSO, Alessandra Marques. Desafios no manejo da fibrose cística. **REVISTA BRASILEIRA MILITAR DE CIÊNCIAS**, v. 9, n. 23, 5 abr. 2023.

- MCDONALD, Eli Fritz *et al.* Benchmarking AlphaMissense Pathogenicity Predictions Against Cystic Fibrosis Variants. **bioRxiv**, p. 2023.10.05.561147, 4 jan. 2024.
- MCLAREN, William *et al.* The Ensembl Variant Effect Predictor. **Genome Biology**, v. 17, n. 1, p. 122, dez. 2016.
- MICHELS, Marcus *et al.* Determining the pathogenicity of CFTR missense variants: Multiple comparisons of in silico predictors and variant annotation databases. **Genetics and Molecular Biology**, v. 42, n. 3, p. 560–570, set. 2019.
- MOREIRA, Rennan Garcias *et al.* Population genetics of *PDE4B* (phosphodiesterase-4B) in neglected Native Americans: Implications for cancer pharmacogenetics. **Clinical and Translational Science**, v. 15, n. 6, p. 1400–1405, jun. 2022.
- MOTA, Daniel Marques; VIGO, Álvaro; KUCHENBECKER, Ricardo De Souza. Formulários de notificação de reações adversas a medicamento nos sistemas de farmacovigilância do Brasil e outros doze países latino-americanos: análise comparativa. **Ciência & Saúde Coletiva**, v. 26, n. 4, p. 1245–1257, abr. 2021.
- NASLAVSKY, Michel S. *et al.* Whole-genome sequencing of 1,171 elderly admixed individuals from Brazil. **Nature Communications**, v. 13, n. 1, p. 1004, 4 mar. 2022.
- NG, P. C. SIFT: predicting amino acid changes that affect protein function. **Nucleic Acids Research**, v. 31, n. 13, p. 3812–3814, 1 jul. 2003.
- NUNES, Kelly *et al.* Admixture's impact on Brazilian population evolution and health. **Science**, v. 388, n. 6748, 15 maio 2025.
- ODDSSON, Asmundur *et al.* Deficit of homozygosity among 1.52 million individuals and genetic causes of recessive lethality. **Nature Communications**, v. 14, n. 1, p. 3453, 10 jun. 2023.
- ODE, Katie Larson *et al.* Cystic fibrosis related diabetes: Medical management. **Journal of Cystic Fibrosis**, v. 18, p. S10–S18, out. 2019.
- OSCANOA, Jorge *et al.* SNPnexus: a web server for functional annotation of human genome sequence variation (2020 update). **Nucleic Acids Research**, v. 48, n. W1, p. W185–W192, 2 jul. 2020.
- PERRY, Luke A. *et al.* Topical cystic fibrosis transmembrane conductance regulator gene replacement for cystic fibrosis-related lung disease. **Cochrane Database of Systematic Reviews**, v. 2016, n. 7, 17 jun. 2016.
- PESSÔA, Renata F.; NÁCUL, Flávio E.; NOËL, François. FARMACOGENÉTICA E FARMACOGENÔMICA. EVIDÊNCIAS DE COMO A GENÉTICA PODE INFLUENCIAR A EFICÁCIA DE FÁRMACOS E A BUSCA POR NOVOS ALVOS FARMACOLÓGICOS. 2006.
- PLASSCHAERT, Lindsey W.; MACDONALD, Kelvin D.; MOFFIT, Jeffrey S. Current landscape of cystic fibrosis gene therapy. **Frontiers in Pharmacology**, v. 15, p. 1476331, 8 out. 2024.

- PREM, Pranjali *et al.* The Impact of Increased Homozygosity on Human Fertility: A Comprehensive Review. **Cureus**, 18 nov. 2023.
- RASTOGI, Ruchir *et al.* Critical assessment of missense variant effect predictors on disease-relevant variant data. **Human Genetics**, v. 144, n. 2–3, p. 281–293, mar. 2025.
- RIBEIRO-DOS-SANTOS, André M. *et al.* Exome Sequencing of Native Populations From the Amazon Reveals Patterns on the Peopling of South America. **Frontiers in Genetics**, v. 11, p. 548507, 29 out. 2020.
- RODEN, Dan M. *et al.* Pharmacogenomics. **The Lancet**, v. 394, n. 10197, p. 521–532, ago. 2019.
- RODRIGUES, Juliana Carla Gomes *et al.* Polymorphisms of ADME-related genes and their implications for drug safety and efficacy in Amazonian Amerindians. **Scientific Reports**, v. 9, n. 1, p. 7201, 10 maio 2019.
- RODRIGUES, Juliana Carla Gomes *et al.* Pharmacogenomic Profile of Amazonian Amerindians. **Journal of Personalized Medicine**, v. 12, n. 6, p. 952, 10 jun. 2022.
- RODRIGUES-SOARES, Fernanda; KEHDY, Fernanda S. G.; GOUVEIA, Mateus H. Editorial: Genomic Ancestry and Biological Traits. **Frontiers in Genetics**, v. 12, p. 754725, 6 set. 2021.
- ROSA, Fernanda Ribeiro *et al.* Fibrose cística: uma abordagem clínica e nutricional. **Revista de Nutrição**, v. 21, n. 6, p. 725–737, dez. 2008.
- SAFERALI, Aabida *et al.* *C FTR* variants are associated with chronic bronchitis in smokers. **European Respiratory Journal**, v. 60, n. 2, p. 2101994, ago. 2022.
- SAINZ DE MEDRANO SAINZ, Jaime I.; BRUNET SERRA, Mercè. Influencia de la farmacogenética en la diversidad de respuesta a las estatinas asociada a las reacciones adversas. **Advances in Laboratory Medicine / Avances en Medicina de Laboratorio**, v. 4, n. 4, p. 353–364, 18 dez. 2023.
- SANAVIA, Tiziana *et al.* Limitations and challenges in protein stability prediction upon genome variations: towards future applications in precision medicine. **Computational and Structural Biotechnology Journal**, v. 18, p. 1968–1979, 2020.
- SAPOZHNIKOV, Yesol *et al.* Statistical modeling to quantify the uncertainty of FoldX-predicted protein folding and binding stability. **BMC Bioinformatics**, v. 24, n. 1, 12 nov. 2023.
- SCOTET, Virginie; L’HOSTIS, Carine; FÉREC, Claude. The Changing Epidemiology of Cystic Fibrosis: Incidence, Survival and Impact of the CFTR Gene Discovery. **Genes**, v. 11, n. 6, p. 589, 26 maio 2020.
- SHARMA, N.; CUTTING, G. R. The genetics and genomics of cystic fibrosis. **Journal of Cystic Fibrosis**, v. 19, p. S5–S9, mar. 2020.
- SHEU, Ruey-Kai; PARDESHI, Mayuresh Sunil. A Survey on Medical Explainable AI (XAI): Recent Progress, Explainability Approach, Human Interaction and Scoring System. **Sensors**, v. 22, n. 20, p. 8068, 21 out. 2022.

- SIDDIK, Abu Bakar; BADAL, Faisal R.; ISLAM, Afroza. **Comparative Performance of Machine Learning Algorithms for Early Genetic Disorder and Subclass Classification**. arXiv, , 3 dez. 2024. Disponível em: <<http://arxiv.org/abs/2412.02189>>. Acesso em: 16 jul. 2025
- SIM, Ngak-Leng *et al.* SIFT web server: predicting effects of amino acid substitutions on proteins. **Nucleic Acids Research**, v. 40, n. W1, p. W452–W457, 1 jul. 2012.
- SIRUGO, Giorgio; WILLIAMS, Scott M.; TISHKOFF, Sarah A. The Missing Diversity in Human Genetic Studies. **Cell**, v. 177, n. 1, p. 26–31, mar. 2019.
- SOBITAN, Adebisi *et al.* Prediction of the Effects of Missense Mutations on Human Myeloperoxidase Protein Stability Using In Silico Saturation Mutagenesis. **Genes**, v. 13, n. 8, p. 1412, 8 ago. 2022.
- SOUZA, Aracele Maria De *et al.* A systematic scoping review of the genetic ancestry of the Brazilian population. **Genetics and Molecular Biology**, v. 42, n. 3, p. 495–508, set. 2019.
- STOLLAR, Elliott J.; SMITH, David P. Uncovering protein structure. **Essays in Biochemistry**, v. 64, n. 4, p. 649–680, 8 out. 2020.
- SUAREZKURTZ, G. Pharmacogenomics in admixed populations. **Trends in Pharmacological Sciences**, v. 26, n. 4, p. 196–201, abr. 2005.
- TORDAI, Hedvig *et al.* Analysis of AlphaMissense data in different protein groups and structural context. **Scientific Data**, v. 11, n. 1, p. 495, 14 maio 2024.
- VAN GOOR, Fredrick *et al.* Effect of ivacaftor on CFTR forms with missense mutations associated with defects in protein processing or function. **Journal of Cystic Fibrosis**, v. 13, n. 1, p. 29–36, jan. 2014.
- WALSH, Nicola *et al.* Variant reclassification and clinical implications. **Journal of Medical Genetics**, v. 61, n. 3, p. 207–211, mar. 2024.
- WANG, Juexin *et al.* Exploring Human Diseases and Biological Mechanisms by Protein Structure Prediction and Modeling. In: SHEN, Bairong; TANG, Haixu; JIANG, Xiaoqian (Orgs.). **Translational Biomedical Informatics**. Advances in Experimental Medicine and Biology. Singapore: Springer Singapore, 2016. v. 939 p. 39–61.
- WHIRL-CARRILLO, M. *et al.* Pharmacogenomics Knowledge for Personalized Medicine. **Clinical Pharmacology & Therapeutics**, v. 92, n. 4, p. 414–417, out. 2012.
- WU, Malinda *et al.* Racial inequities and rare CFTR variants: Impact on cystic fibrosis diagnosis and treatment. **Journal of Clinical & Translational Endocrinology**, v. 36, p. 100344, jun. 2024.
- WU, You; XIE, Lei. **AI-driven multi-omics integration for multi-scale predictive modeling of causal genotype-environment-phenotype relationships**. arXiv, , 8 jul. 2024. Disponível em: <<http://arxiv.org/abs/2407.06405>>. Acesso em: 15 jan. 2025
- YU, He *et al.* Paleolithic to Bronze Age Siberians Reveal Connections with First Americans and across Eurasia. **Cell**, v. 181, n. 6, p. 1232–1245.e20, jun. 2020.

5. CAPITULO II

Este capítulo é referente ao manuscrito intitulado “Dissecting Predictive Accuracy in *CFTR* Variant Interpretation: A Comparative Analysis of Pathogenicity Predictors and Free Energy Approaches” em preparação para submissão no periódico.

Dissecting Predictive Accuracy in *CFTR* Variant Interpretation: A Comparative Analysis of Pathogenicity Predictors and Free Energy Approaches

Ana Katarina Campos Nunes¹, Camila Sinimbú Forte², Arthur Felipe Vasconcelos Ferreira Reis¹ and Gilderlanio Santana de Araújo¹

¹Human and Medical Genetics Lab, Institute of Biological Sciences, Federal University of Pará, Belém, Pará, Brazil

²Institute of Biological Sciences, Federal University of Minas Gerais, Belo Horizonte Minas Gerais, Brazil

Corresponding Author:

Email: gildelanio@ufpa.br

Abstract:

Accurate interpretation of *CFTR* variants is essential for understanding cystic fibrosis (CF) pathogenesis. We analyzed 1,098 variants from the CFTR2 database – an expert-curated database, identifying 92.5% as CF-causing, 2.5% as non CF-causing, and 5.0% as having varying clinical consequences. Most variants were rare (allele frequency ≤ 0.009), highlighting the genetic heterogeneity of CF. Loss-of-function variants, particularly frameshift (40.9%) and stop-gained (20.7%) predominated, followed by missense (15.2%) and splice site alterations. We evaluated 166 missense variants using three computational predictors: SIFT, PolyPhen-2, and AlphaMissense. AlphaMissense outperformed others, achieving 98.5% F1-score and the highest concordance with clinical classifications. It correctly identified CF-causing variants as likely pathogenic and distinguished non CF-causing better than SIFT and PolyPhen. Furthermore, AlphaMissense predictions showed the strongest correlation with FoldX-estimated structural destabilization ($\Delta\Delta G$, $r = 0.38$), surpassing SIFT ($r = -0.22$) and PolyPhen-2 ($r = 0.33$). These results highlight the value of AlphaMissense as a reliable tool for predicting pathogenic *CFTR* variants and suggest its scores partially reflect underlying biophysical impacts. Our findings support integrating advanced AI-based predictors into clinical variant interpretation frameworks to enhance precision medicine efforts in CF.

Introduction

The identification of causal variants underlying complex diseases has advanced significantly through candidate gene analyses and genome-wide association studies, enabled by next-generation sequencing technologies (Abdellaoui et al., 2023). These approaches have revealed both novel and population-specific variants, with rare genetic disorders such as cystic fibrosis (CF) serving as critical models for exploring genotype–phenotype relationships. CF is the most common autosomal recessive disease, affecting ~80,000 individuals worldwide (Cromwell et al., 2023; Plasschaert; MacDonald; Moffit, 2024), and is caused by pathogenic variants in the *CFTR* gene. Initially characterized by exonic mutations, *CFTR* variants have since been implicated in broader phenotypes, including gastrointestinal cancers through splicing defects (Deletang, 2022).

The *CFTR* gene encodes a phosphorylation-regulated, ATP-gated ion channel of the ABC transporter family (Bacalhau *et al.*, 2023), expressed at the apical surface of epithelial cells. Over 2,000 variants have been identified, many of which disrupt protein expression or function. The F508del variant remains the most prevalent disease-causing allele (Bareil & Bergounoux, 2020). Curated resources such as the CFTR2 database provide clinically annotated variant data from over 122,000 individuals, categorizing mutations as CF-causing, non-CF-causing, or of uncertain significance.

Accurate interpretation of missense variants is essential but remains challenging. Tools such as AlphaMissense, SIFT, and PolyPhen-2 are widely used for predicting the pathogenicity of amino acid substitutions. However, benchmarking studies have shown that these models can exhibit biases and misclassify neutral variants as pathogenic (McDonald et al., 2024; Michels et al., 2019), particularly in the context of class imbalance.

Here, we systematically assess the predictive accuracy of AlphaMissense, SIFT, and PolyPhen-2, along with bioenergetic models, in classifying missense variants in the *CFTR* gene. By examining tool-specific biases and their susceptibility to class imbalance, our study aims to enhance the robustness of *in silico* pathogenicity predictions and support their application in clinical variant interpretation for CF and other Mendelian disorders. In parallel, we perform a population-level screening of

expert-curated *CFTR* variants using large-scale genomic databases , including the 1000 Genomes Project, All of Us, Brazilian Genomes, and gnomAD, aiming to characterize their allele frequencies and population distribution of CF-causing variants to fill a gap on genetic studies of underrepresented populations.

Materials and Methods

Clinically Curated Catalogue of CFTR Variants

The Clinical and Functional Translation of CFTR (CFTR2) database (<https://cftr2.org/>) stores CFTR variants associated with Cystic Fibrosis, allele frequency, and clinical annotation of causality. For this study, we pre-processed CFTR2 (September 24, 2024), which includes 1,167 functionally annotated variants found from a total of 122,935 patients (Castellani, 2013). Among these, 1,085 variants are classified as CF-causing, 27 non CF-causing, and 55 variants have varying clinical consequences (Supplementary Table 1), for the reference transcript for the CFTR gene (NM_000492.4). The CFTR2 data was preprocessed for removing duplicates, handling missing values, and standardizing formats to ensure data consistency and accuracy. For each variant, we extract the following attributes: the variant legacy, frequency, chromosome, position, reference allele, alternative allele, and clinical significance. None of these attributes may present missing data.

Computational assessment of CFTR variant pathogenicity

Following data preprocessing and curation, all CFTR variants were functionally annotated using the Ensembl Variant Effect Predictor (VEP) via automated batch queries (McLaren et al., 2016). VEP integrates multiple computational tools for the prediction of variant impact, including AlphaMissense, SIFT, and PolyPhen-2. AlphaMissense provides probabilistic pathogenicity scores for missense variants based on a deep learning model trained on amino acid distributions within sequence context, augmented with structural features derived from AlphaFold predictions (Cheng et al., 2023). The model leverages unsupervised training on large-scale population frequency data, avoiding reliance on manually curated annotations. Variants are categorized as likely pathogenic, ambiguous, or likely benign. SIFT (Sorting Intolerant From Tolerant) predicts the effect of amino acid substitutions on protein function based on sequence homology and the degree of evolutionary conservation at each position variants

(Kumar; Henikoff; Ng, 2009). Variants are classified as deleterious or tolerated, providing insight into the potential disruption of protein function. PolyPhen-2 (Polymorphism Phenotyping v2) estimates the impact of missense substitutions by integrating multiple features, including sequence conservation, protein structure, and functional domain annotations (Adzhubei; Jordan; Sunyaev, 2013). Variants are assigned to benign, possibly damaging, or probably damaging categories based on the predicted probability of pathogenicity.

CFTR stability predictions with FoldX

To estimate the impact of missense variants on *CFTR* protein stability, we performed $\Delta\Delta G$ calculations using the FoldX suite (Delgado et al., 2025; Gerasimavicius; Livesey; Marsh, 2022), a structure-based tool for predicting changes in protein free energy upon mutation (Buss; Rudat; Ochsenreither, 2018; Sobitan et al., 2022). Analyses were conducted locally with default parameters.

We employed the AlphaFold-predicted *CFTR* structure (AF-P13569-F1-v4) as the input template. Prior to mutation modeling, the structure was preprocessed using the RepairPDB function to resolve steric clashes and optimize atomic coordinates. Missense variants were individually introduced using the BuildModel command, which computes folding energy changes by comparing mutant and wild-type structures.

The resulting folding energy difference ($\Delta\Delta G = \Delta G_{MUT} - \Delta G_{WT}$) was used to classify the predicted effects on stability: stabilizing ($\Delta\Delta G < -0.46$ kcal/mol), neutral ($-0.46 \leq \Delta\Delta G \leq +0.46$ kcal/mol), or destabilizing ($\Delta\Delta G > +0.46$ kcal/mol), as previously described (Sanavia et al., 2020; Sobitan et al., 2022). Positive $\Delta\Delta G$ values indicate destabilizing variants, while negative values suggest stabilization.

Statistical analysis

To examine the relationship between pathogenicity prediction scores and protein stability, we computed Pearson correlation coefficients between the outputs of AlphaMissense, SIFT, and PolyPhen-2 and the corresponding FoldX-derived $\Delta\Delta G$ values.

To assess whether AlphaMissense scores correlate with predicted destabilization, we performed a simple linear regression using $\Delta\Delta G$ as the dependent variable and AlphaMissense score as the independent variable.

We further evaluated whether protein stability predictions could discriminate clinically relevant variants. A binary logistic regression was conducted using $\Delta\Delta G$ as a predictor of clinical classification (CF-causing vs. non-CF-causing), with odds ratios (ORs) and 95% confidence intervals (CIs) reported to quantify the strength and precision of association.

Results

CF-causing and non CF-causing variants

A total of 1,098 CFTR variants were identified following rigorous data extraction and curation from the CFTR2 database. Among these, 1,016 variants (92.5%) were classified as CF-causing, while 27 (2.5%) were non-CF-causing and 55 (5.0%) were of varying clinical significance. Frequency data referenced a cohort of 122,935 patients, revealing that F508del remains the most prevalent variant, accounting for 65.1% of all alleles observed. Other recurrent CF-causing variants included G542X (2.7%), G551D (1.8%), N1303K (1.7%), and W1282X (1.2%). The R117H variant, despite its variable clinical impact, was also among the most frequent, with an allele frequency of 1.1%. Notably, the vast majority of variants occurred at low allele frequencies (≤ 0.009), indicating their rare occurrence within the CF patient population and underscoring the genetic heterogeneity of the disease (Figure 1A and Figure 1B, Supplementary Table 1).

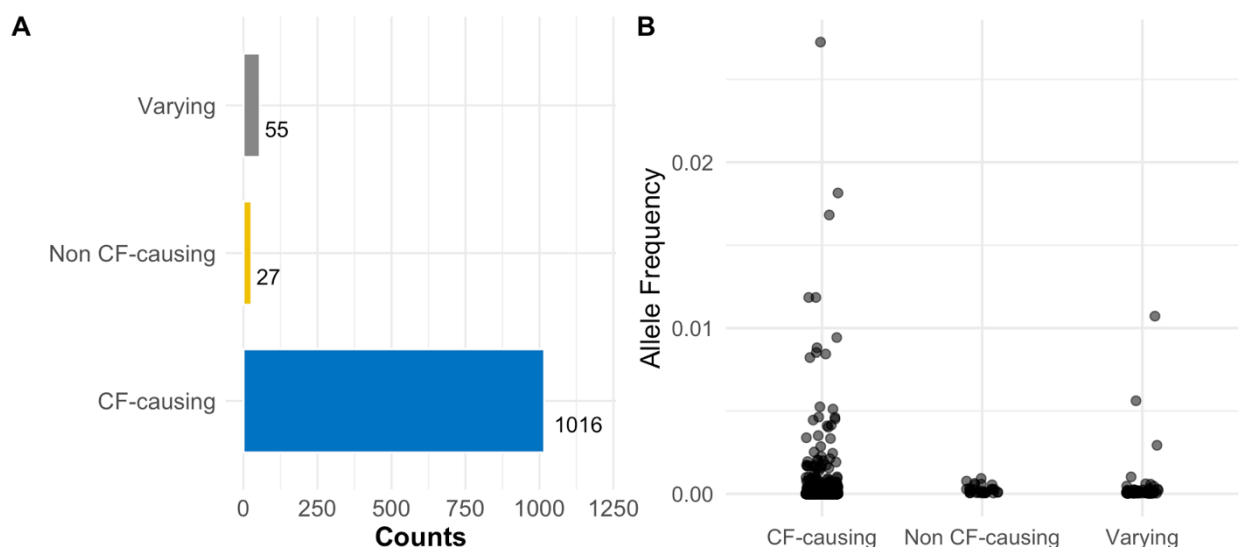


Figure 1 | Distribution and allele frequency of CFTR2 variants across clinical classifications. 1a, Number of variants categorized as CF-causing, non CF-causing, or varying clinical consequence in the CFTR2.org database. 1b, Rare allele frequencies of variants grouped by clinical classification, showing the distribution across CF-causing, non CF-causing, and varying effect groups.

Among the curated CFTR variants, frameshift variants (40.9%) and stop-gained mutations (20.7%) were the most frequent consequence types, together accounting for over 60% of all annotated variants. Missense variants represented 15.2% of the dataset, while splice donor (7.0%) and splice acceptor (6.0%) variants were also prevalent, reflecting the disruption of essential splicing elements in CFTR pathology. Other consequence types, including synonymous, inframe, and UTR variants, were comparatively rare (<1% each), consistent with their typically lower impact on protein function (Supplementary Table 2 – Consequence Table).

Assessment of Genetic Variant Effects Prediction

The analysis of 166 missense genetic variants shows a strong trend toward potentially harmful effects. Based on SIFT predictions, most variants (159) were labeled as *deleterious_low_confidence*, while only 7 were considered *tolerated_low_confidence*. PolyPhen results support this finding: 17 variants were classified as *benign*, but a much larger number were predicted to be *possibly damaging* (23) or *probably damaging* (126). Similarly, AlphaMissense categorized 127 variants as *likely pathogenic*, compared to just 25 as *likely benign* and 14 as *ambiguous*. Overall, these results suggest that many of the variants analyzed may negatively impact protein function, highlighting the need for further investigation.

AlphaMissense Classifier Shows Strong Concordance with Clinical Impact of CFTR Variants

To evaluate the predictive value of computational tools in assessing the clinical relevance of CFTR missense variants, we compared annotations from SIFT, PolyPhen-2, and AlphaMissense against expert-curated clinical impact classifications (CF-causing, Non CF-causing, and Varying).

SIFT predictions indicated that a majority of CF-causing variants (n = 103) were labeled as deleterious, with fewer assigned to the non CF-causing (n = 13) and Varying (n = 43) categories. Tolerated predictions were rare and primarily associated with non CF-causing variants (n = 5), with only one CF-causing variant misclassified. Although the low confidence designation associated with these categories warrants caution, deleterious predictions captured a substantial proportion of disease-associated variants, supporting the tool's relevance for identifying potentially pathogenic alleles.

PolyPhen-2 annotations showed a similar pattern, with probably damaging variants enriched in the CF-causing group (n = 90), and less frequent among Non CF-causing (n = 7) and Varying (n = 29) classifications. However, possibly damaging and benign predictions exhibited substantial overlap across clinical impact groups, including five CF-causing variants predicted as benign — suggesting a measurable false negative rate.

The AlphaMissense model, trained on large-scale protein and genomic datasets, demonstrated the highest concordance with clinical labels. Among variants classified as likely pathogenic, 100 were CF-causing, with only two and 25 assigned to the Non CF-causing and Varying groups, respectively. Likely benign predictions were largely confined to Non CF-causing (n = 15) and Varying (n = 9) variants, with just one misclassified CF-causing variant. Ambiguous predictions were primarily linked to variants with varying impact (n = 10).

For benchmarking, a binary classification scheme was adopted in which CF-causing variants were treated as true positives (TP), while all others (Non CF-causing and Varying) were considered negatives. Predictions of deleterious (SIFT), probably damaging (PolyPhen-2), and likely pathogenic (AlphaMissense) were treated as positives. Predictions labeled as tolerated, benign, and likely benign were treated as negatives. Intermediate categories such as ambiguous and possibly damaging were excluded from metric calculations to ensure comparability across tools.

Under these assumptions, AlphaMissense achieved the highest performance, with an accuracy of 97.5%, precision of 98.0%, recall of 99.0%, and an F1-score of 98.5%. In comparison, SIFT yielded a similar recall (99.0%) but lower precision (88.9%) and accuracy (88.5%), resulting in an F1-score of 93.7%. PolyPhen-2 showed balanced yet slightly lower performance (F1 = 93.8%), with a greater number of misclassified variants reflected in its reduced specificity. These findings highlight the superior

discriminative capacity of AlphaMissense in identifying pathogenic CFTR variants and support its integration into clinical variant interpretation workflows (Figure 2, Figure 3).

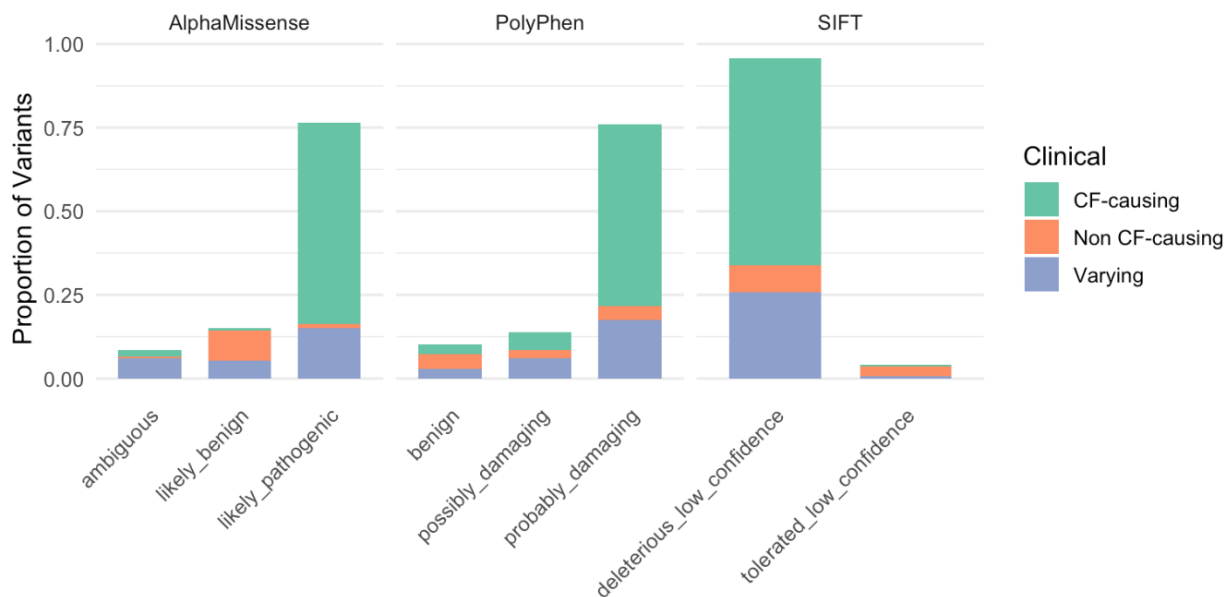


Figure 2 | Proportion of variants predicted by AlphaMissense, PolyPhen-2, and SIFT across clinical classifications. Variants from CFTR2.org were stratified by expert-curated clinical impact (*CF-causing*, *non CF-causing*, *varying clinical consequence*) and evaluated using AlphaMissense, PolyPhen-2, and SIFT annotations.

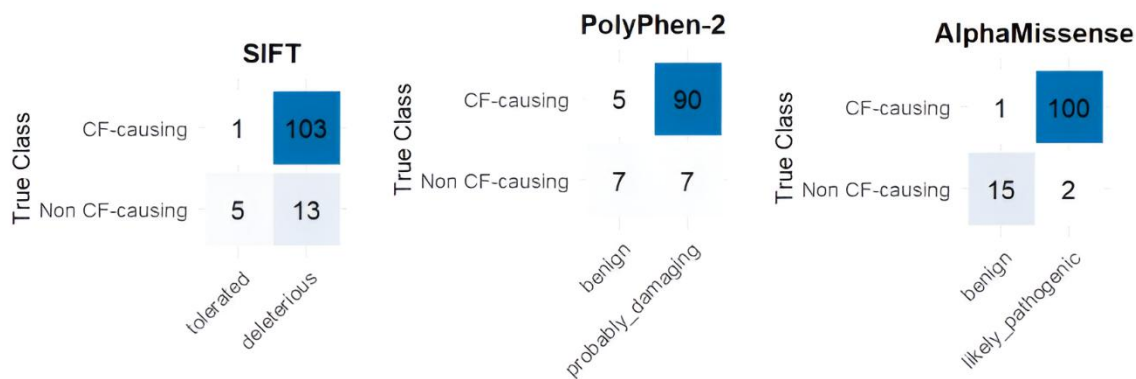


Figure 3 | Confusion matrix of computational predictions versus clinical classifications of CFTR variants. Variants from CFTR2.org were stratified by expert-curated clinical impact (*CF-causing* and *non CF-causing*) and compared to functional effect predictions from AlphaMissense, PolyPhen-2, and SIFT. The matrix illustrates

concordance and discordance patterns between clinical and computational classifications.

AlphaMissense scores correlates with Gibbs Free Energy estimated by FoldX

Among the tools evaluated, AlphaMissense showed the strongest correlation with the $\Delta\Delta G$ values calculated by FoldX ($r = 0.38$; $p \leq 3.8 \times 10^{-7}$), suggesting greater agreement between its pathogenicity predictions and the degree of structural destabilization predicted for the variants. This result indicates that, although AlphaMissense is not directly based on thermodynamic parameters, its scores partly reflect relevant structural changes. In comparison, SIFT showed a weaker correlation ($r = -0.22$), and PolyPhen-2 showed an intermediate correlation ($r = 0.33$), demonstrating less alignment with FoldX stability predictions (Figure 4). These data reinforce that, among the predictors analyzed, AlphaMissense is the closest to the structural behavior of variants, even using different prediction approaches (Figure 4).

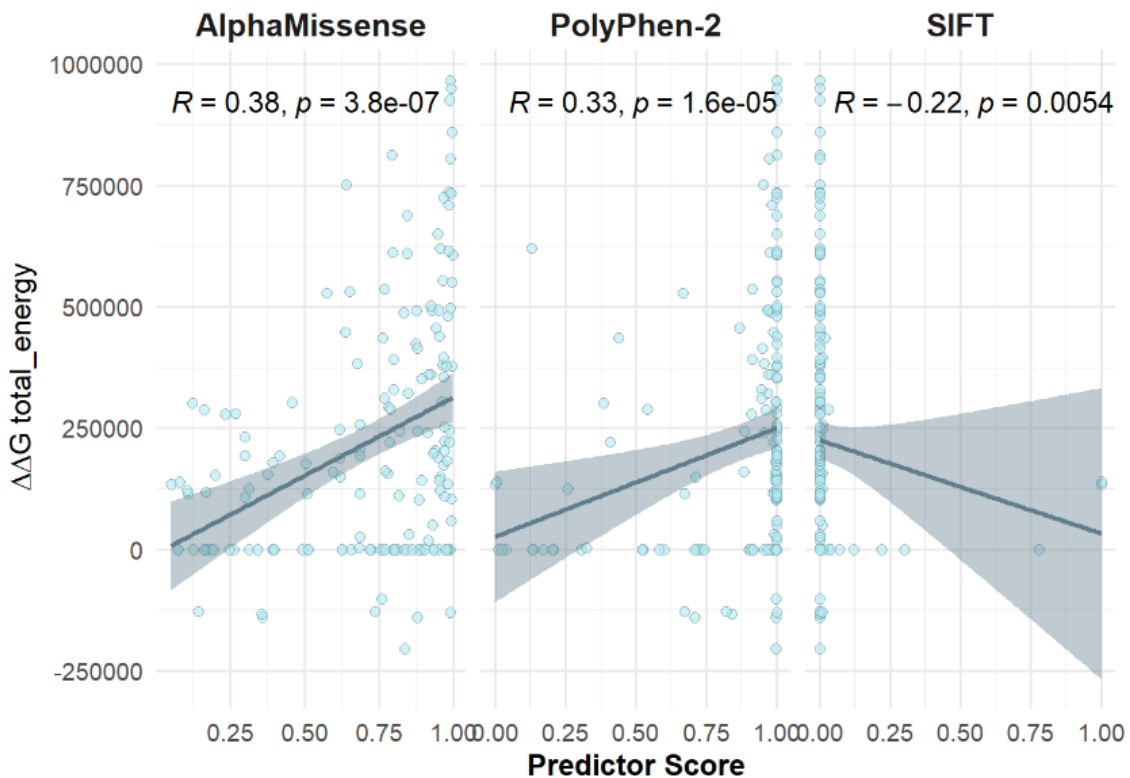


Figure 4 | Correlation between pathogenicity predictor scores and $\Delta\Delta G$ total_energy values for CFTR missense variants. Shaded areas around the regression lines represent 95% confidence intervals. Spearman's correlation coefficient (R) and p-

values are displayed at the top of each facet. A positive correlation was observed for AlphaMissense ($R = 0.38$, $p = 3.8\text{e-}07$) and PolyPhen-2 ($R = 0.33$, $p = 1.6\text{e-}05$), indicating that higher predictor scores tend to be associated with greater structural destabilization. In contrast, SIFT showed a negative correlation ($R = -0.22$, $p = 0.0054$), consistent with its inverse scoring logic (lower scores = more deleterious).

Discussion

The increasing resolution of next-generation sequencing has revealed extensive mutational heterogeneity in the CFTR gene. While resources like CFTR2.org have catalogued numerous clinically relevant variants, particularly intronic and exonic missense changes, accurate interpretation remains a bottleneck. Computational tools offer a scalable solution, but their reliability varies, especially for non-CF-causing or uncertain variants.

Comparing AlphaMissense, SIFT, and PolyPhen-2 against CFTR2 data, all tools reliably identified CF-causing variants. However, discrepancies emerged in the classification of benign or uncertain variants. AlphaMissense and PolyPhen-2, despite their utility, frequently overestimated pathogenicity, echoing findings by McDonald et al. (2024). SIFT achieved the highest accuracy for pathogenic variants, while AlphaMissense performed better for non-pathogenic calls. Such inconsistencies stress the importance of integrating multiple prediction sources with empirical validation (Gerasimavicius et al., 2023).

McDonald et.al, 2024 conducted an isolated assessment of AlphaMissense, focusing solely on its performance, without exploring its interaction or combined use with other predictive tools. By comparing AlphaMissense to SIFT and PolyPhen-2 in the context of an ensemble approach, our study provides a broader perspective, revealing how these tools complement or conflict with each other in variant classification. This integrated analysis highlights the nuances of their performance and emphasizes the importance of using multiple approaches to improve prediction accuracy, particularly for challenging categories like benign or uncertain variants.

The integration of AlphaMissense with FoldX provided deeper insight into the biophysical consequences of missense mutations. AlphaMissense showed the highest

correlation with $\Delta\Delta G$ values ($r = 0.38$), suggesting it better captures mutational impacts on protein stability. Logistic regression confirmed $\Delta\Delta G$ as an independent predictor of pathogenicity ($OR = 1.376$, $p < 0.01$), supporting the utility of combining thermodynamic and deep learning-based predictors.

However, only 43% of AlphaMissense-classified pathogenic variants showed significant destabilization ($\Delta\Delta G > 1.5$ kcal/mol), consistent with Jänes et al. (2024). This indicates many variants exert pathogenicity through mechanisms not captured by FoldX, highlighting the model's limitations in classifying benign variants. AlphaMissense, by integrating structural and evolutionary context, remains more robust for functional discrimination (Jänes et al., 2024; Sapozhnikov et al., 2023).

Our study demonstrates that while individual predictors provide partial insights, their integration, especially with structural tools like FoldX, enhances the interpretability of ambiguous *CFTR* variants. This approach may inform precision diagnostics and genetic counseling, particularly when dealing with variants of uncertain significance. Nonetheless, further refinement, personalization of models, and integration with transcriptomic/proteomic data are needed to mitigate bias and improve prediction fidelity.

References

- Abdellaoui, A., Yengo, L., Verweij, K. J. H., & Visscher, P. M. (2023). 15 years of GWAS discovery: Realizing the promise. *American journal of human genetics*, 110(2), 179–194. <https://doi.org/10.1016/j.ajhg.2022.12.011>
- Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S., & Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nature methods*, 7(4), 248–249. <https://doi.org/10.1038/nmeth0410-248>
- Bacalhau, M., Camargo, M., Magalhães-Ghiotto, G. A. V., Drumond, S., Castelletti, C. H. M., & Lopes-Pacheco, M. (2023). Elexacaftor-Tezacaftor-Ivacaftor: A Life-Changing Triple Combination of CFTR Modulator Drugs for Cystic Fibrosis. *Pharmaceuticals (Basel, Switzerland)*, 16(3), 410. <https://doi.org/10.3390/ph16030410>
- Bareil, C., & Bergougnoux, A. (2020). CFTR gene variants, epidemiology and molecular pathology. *Archives de pediatrie : organe officiel de la Societe francaise de pediatrie*, 27 Suppl 1, eS8–eS12. [https://doi.org/10.1016/S0929-693X\(20\)30044-0](https://doi.org/10.1016/S0929-693X(20)30044-0)
- Bienvenu, T., Lopez, M., & Girodon, E. (2020). Molecular Diagnosis and Genetic Counseling of Cystic Fibrosis and Related Disorders: New Challenges. *Genes*, 11(6), 619. <https://doi.org/10.3390/genes11060619>
- Bonetti, E., Tini, G., & Mazzarella, L. (2024). Accuracy of renovo predictions on variants reclassified over time. *Journal of translational medicine*, 22(1), 713. <https://doi.org/10.1186/s12967-024-05508-w>
- Buß O, Rudat J, Ochsenreither K. FoldX as Protein Engineering Tool: Better Than Random Based Approaches? Comput Struct Biotechnol J. 2018;16:25-33.doi:10.1016/j.csbj.2018.01.002
- Castellani, C., & CFTR2 team (2013). CFTR2: How will it help care?. *Paediatric respiratory reviews*, 14 Suppl 1, 2–5. <https://doi.org/10.1016/j.prrv.2013.01.006>
- Cheng, J., Novati, G., Pan, J., Bycroft, C., Žemgulytė, A., Applebaum, T., Pritzel, A., Wong, L. H., Zielinski, M., Sargeant, T., Schneider, R. G., Senior, A. W., Jumper, J., Hassabis, D., Kohli, P., & Avsec, Ž. (2023). Accurate proteome-wide missense variant

effect prediction with AlphaMissense. *Science (New York, N.Y.)*, 381(6664), eadg7492.
<https://doi.org/10.1126/science.adg7492>

Cromwell, E. A., Ostrenga, J. S., Todd, J. V., Elbert, A., Brown, A. W., Faro, A., Goss, C. H., & Marshall, B. C. (2023). Cystic fibrosis prevalence in the United States and participation in the Cystic Fibrosis Foundation Patient Registry in 2020. *Journal of cystic fibrosis : official journal of the European Cystic Fibrosis Society*, 22(3), 436–442.
<https://doi.org/10.1016/j.jcf.2023.02.009>

Curtis D. (2024). Assessment of ability of AlphaMissense to identify variants affecting susceptibility to common disease. *European journal of human genetics : EJHG*, 32(11), 1419–1427. <https://doi.org/10.1038/s41431-024-01675-y>

Deletang, K., & Taulan-Cadars, M. (2022). Splicing mutations in the CFTR gene as therapeutic targets. *Gene therapy*, 29(7-8), 399–406. <https://doi.org/10.1038/s41434-022-00347-0>

Delgado J, Radusky LG, Cianferoni D, Serrano L. FoldX 5.0: working with RNA, small molecules and a new graphical interface. *Bioinforma Oxf Engl*. 2019;35(20):4168-4169.doi:10.1093/bioinformatics/btz184

Gerasimavicius, L., Livesey, B. J., & Marsh, J. A. (2023). Correspondence between functional scores from deep mutational scans and predicted effects on protein stability. *Protein science : a publication of the Protein Society*, 32(7), e4688.
<https://doi.org/10.1002/pro.4688>

Gerasimavicius L, Livesey BJ, Marsh JA. Loss-of-function, gain-of-function and dominant-negative mutations have profoundly different effects on protein structure. *NatCommun*. 2022;13(1):3895. doi:10.1038/s41467-022-31686-6

Gunning, A. C., Fryer, V., Fasham, J., Crosby, A. H., Ellard, S., Baple, E. L., & Wright, C. F. (2021). Assessing performance of pathogenicity predictors using clinically relevant variant datasets. *Journal of medical genetics*, 58(8), 547–555.
<https://doi.org/10.1136/jmedgenet-2020-107003>

Harel, T., Peshes-Yaloz, N., Bacharach, E., & Gat-Viks, I. (2019). Predicting Phenotypic Diversity from Molecular and Genetic Data. *Genetics*, 213(1), 297–311.
<https://doi.org/10.1534/genetics.119.302463>

- Jänes, J., Müller, M., Selvaraj, S., Manoel, D., Stephenson, J., Gonçalves, C., Lafita, A., Polacco, B., Obernier, K., Alasoo, K., Lemos, M. C., Krogan, N., Martin, M., Saraiva, L. R., Burke, D., & Beltrao, P. (2024). Predicted mechanistic impacts of human protein missense variants. *bioRxiv : the preprint server for biology*, 2024.05.29.596373. <https://doi.org/10.1101/2024.05.29.596373>
- Kumar, P., Henikoff, S., & Ng, P. C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature protocols*, 4(7), 1073–1081. <https://doi.org/10.1038/nprot.2009.86>
- Livesey, B. J., & Marsh, J. A. (2023). Updated benchmarking of variant effect predictors using deep mutational scanning. *Molecular systems biology*, 19(8), e11474. <https://doi.org/10.15252/msb.202211474>
- McDonald, E. F., Oliver, K. E., Schleich, J. P., Meiler, J., & Plate, L. (2024). Benchmarking AlphaMissense pathogenicity predictions against cystic fibrosis variants. *PloS one*, 19(1), e0297560. <https://doi.org/10.1371/journal.pone.0297560>
- McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R., Thormann, A., Flicek, P., & Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome biology*, 17(1), 122. <https://doi.org/10.1186/s13059-016-0974-4>
- Michels, M., Matte, U., Fraga, L. R., Mancuso, A. C. B., Ligabue-Braun, R., Berneira, E. F. R., Siebert, M., & Sanseverino, M. T. V. (2019). Determining the pathogenicity of CFTR missense variants: Multiple comparisons of in silico predictors and variant annotation databases. *Genetics and molecular biology*, 42(3), 560–570. <https://doi.org/10.1590/1678-4685-GMB-2018-0148>
- Palheta, H. G. A., Gonçalves, W. G., Brito, L. M., Dos Santos, A. R., Dos Reis Matsumoto, M., Ribeiro-Dos-Santos, Â., & de Araújo, G. S. (2022). AmazonForest: In Silico Metaprediction of Pathogenic Variants. *Biology*, 11(4), 538. <https://doi.org/10.3390/biology11040538>
- Plasschaert, L. W., MacDonald, K. D., & Moffit, J. S. (2024). Current landscape of cystic fibrosis gene therapy. *Frontiers in pharmacology*, 15, 1476331. <https://doi.org/10.3389/fphar.2024.1476331>

Rastogi, R., Chung, R., Li, S. *et al.* Critical assessment of missense variant effect predictors on disease-relevant variant data. *Hum. Genet.* **144**, 281–293 (2025).
<https://doi.org/10.1007/s00439-025-02732-2>.

Sanavia T, Birolo G, Montanucci L, Turina P, Capriotti E, Fariselli P. Limitations and challenges in protein stability prediction upon genome variations: towards future applications in precision medicine. *Comput Struct Biotechnol J.* 2020;18:1968-1979.doi:10.1016/j.csbj.2020.07.011

Sobitan A, Edwards W, Jalal MS, et al. Prediction of the Effects of Missense Mutations on Human Myeloperoxidase Protein Stability Using In Silico Saturation Mutagenesis. *Genes.* 2022;13(8):1412. doi:10.3390/genes13081412

Walsh, N., Cooper, A., Dockery, A., & O'Byrne, J. J. (2024). Variant reclassification and clinical implications. *Journal of medical genetics*, 61(3), 207–211.
<https://doi.org/10.1136/jmg-2023-109488>

6. DISCUSSÃO GERAL

Dessa forma, o primeiro capítulo deste trabalho apresenta uma análise recente do gene CFTR em populações indígenas amazônicas, na qual foram identificadas nove variantes — sendo três delas exclusivas da amostra. Entre essas, seis foram classificadas como benignas ou provavelmente benignas, duas de significado incerto e uma como patogênica, associada à fibrose cística e à insuficiência pancreática (CFTR2). Destaca-se a variante intrônica c.2988+1G>A, que afeta o splicing e dificulta a resposta aos tratamentos disponíveis. Vale ressaltar que, do ponto de vista farmacológico, pacientes com essa mutação geralmente não respondem aos moduladores atualmente disponíveis (Deletang, 2022). Essa limitação evidencia a importância de se aprimorar a classificação de variantes, a fim de evitar falhas terapêuticas (Joynt; Cutting; Sharma, 2022).

Além disso, o estudo enfatiza a importância de investigar populações subrepresentadas. Na região amazônica, muitos povos originários já tiveram contato com populações externas, enquanto outros permanecem mais isolados. Em ambos os casos, observa-se uma significativa escassez de estudos genômicos, o que resulta em uma limitada base de dados genéticos. Isso reforça a urgência de desenvolver ações de pesquisa e saúde que contemplem essas comunidades, ampliando o conhecimento genético e contribuindo diretamente para a epidemiologia e para a farmacogenômica (Sirugo; Williams; Tishkoff, 2019).

Entre as variantes identificadas, duas localizadas no domínio nucleotídico 2 (NBD2) merecem destaque. A c.3728T>C (Leu1243Ser) altera a polaridade do aminoácido, o que pode comprometer suas interações moleculares. Já a c.4045G>A (Gly1349Ser), já descrita na literatura, possui frequência elevada no leste asiático, refletindo a ancestralidade asiática presente nos povos indígenas americanos (Anzai et al., 2003; Ribeiro-dos-Santos et al., 2020).

No domínio transmembranar 2 (TMD2), foram identificadas as variantes c.2588T>A (Leu863Gln) e c.2932A>G (Lys978Glu). A primeira pode interferir nas interações com a membrana, enquanto a segunda envolve a substituição de um aminoácido básico por outro de caráter ácido, potencialmente desestabilizando o canal iônico. Considerando a proximidade do TMD2 com a região apolar da membrana, é possível que a maior polaridade da glutamina gere conflitos estruturais com a membrana celular.

Especificamente na variante c.2932A>G (Lys978Glu), observa-se que a lisina, um aminoácido básico com cadeia lateral hidrofóbica e longa, interage com outras moléculas por meio de forças de van der Waals, ligações de hidrogênio e interações eletrostáticas. Ao ser substituída por ácido glutâmico — que possui cadeia lateral semelhante, porém com carga negativa — há possibilidade de mudanças significativas na dinâmica molecular (Stollar; Smith, 2020).

A análise dessas variantes em populações isoladas revelou uma baixa frequência de homozigotos para FC, o que pode estar associado ao isolamento geográfico e à consanguinidade. Mesmo assim, a presença de variantes exclusivas reforça a necessidade de ampliar os estudos genéticos em populações não europeias, ainda sub-representadas nos grandes bancos de dados genômicos (Sirugo; Williams; Tishkoff, 2019).

Nesse cenário, a modelagem estrutural de proteínas mostrou-se uma ferramenta essencial para avaliar as interações moleculares e os impactos funcionais das variantes sobre a CFTR. Os achados deste estudo reforçam a importância de incluir populações miscigenadas e sub-representadas, como a brasileira, em pesquisas genéticas e farmacogenômicas, com vistas à redução de vieses e promoção de maior equidade nos cuidados de saúde (Wang et al., 2016).

A localização da variante no canal iônico, mais especificamente no TMD2, indica que a substituição de um resíduo básico por um ácido pode causar repulsão eletrostática durante a passagem do íon cloreto (Cl^-). Além disso, a presença do ácido glutâmico pode formar pontes de hidrogênio devido à sua eletronegatividade, alterando negativamente a conformação do canal. Supõe-se também a possibilidade de desnaturação da estrutura da proteína, promovendo sua instabilidade.

Com base nas quatro variantes identificadas e nos escores de patogenicidade do AlphaMissense, infere-se que as substituições de aminoácidos analisadas podem afetar tanto a estrutura quanto as interações intermoleculares da proteína, com impacto direto sobre sua funcionalidade e, por consequência, sua patogenicidade (Tordai et al., 2024).

No que se refere aos domínios nucleotídicos, é importante observar que as mutações nesses locais afetam a formação dos sítios de ligação entre NBD1 e NBD2, essenciais para o processo de hidrólise do ATP. Alterações que comprometem essa interação ou a própria hidrólise provavelmente impedem a ativação do canal iônico, caracterizando-se como mutações de classe III no gene (Farinha; Callebaut, 2022).

Observa-se que, na amostra analisada, a maioria das variantes apresentou frequência alélica menor (MAF) igual a zero, o que sugere alta homozigosidade. Esse achado pode ser explicado pelo isolamento geográfico e pelo baixo ou inexistente contato com populações externas ou até pela qualidade de cobertura dos dados. Em grupos pequenos, com gargalos populacionais ou práticas consanguíneas, a homozigosidade tende a ser mais comum (Prem et al., 2023; Sirugo; Williams; Tishkoff, 2019). Ainda assim, foram identificadas variantes missense que, mesmo sem estarem registradas em bancos como o CFTR2, indicam possíveis alterações estruturais na proteína.

A modelagem proteica permitiu a visualização de possíveis interações inter e intramoleculares que podem comprometer a estabilidade da CFTR (Wang et al., 2016). Apesar das limitações inerentes aos dados disponíveis, este estudo oferece uma abordagem relevante para a identificação de novas variantes e destaca a necessidade de ampliação dos painéis genéticos voltados tanto à FC quanto a outras doenças genéticas raras. As três variantes candidatas a exclusivas reforçam que muitas informações permanecem invisíveis em análises genéticas centradas no padrão europeu. Conforme demonstrado por Sirugo, Williams e Tishkoff (2019), a falta de diversidade nos estudos genéticos leva a vieses analíticos: nos estudos de GWAS, 78% dos participantes são europeus, 10% asiáticos, 2% africanos, 1% hispânicos e todas as demais etnias somadas representam menos de 1%.

Populações miscigenadas, como as latino-americanas — e, em especial, o Brasil, cuja formação é baseada na mistura entre povos indígenas, europeus e africanos — ainda não são devidamente contempladas pelos atuais modelos genéticos e farmacogenômicos. O estudo dessas populações é fundamental para compreender melhor os perfis metabólicos e expandir os horizontes da medicina de precisão, sobretudo em contextos de sistemas públicos de saúde, como o brasileiro.

Diante disso, o uso de ferramentas de predição de patogenicidade surge como estratégia para otimizar a descoberta de variantes em genomas e exomas, oferecendo resultados mais ágeis. Entender a relação genótipo-fenótipo torna-se imprescindível para propor medidas terapêuticas personalizadas. Com esse propósito, os preditores de patogenicidade devem buscar o maior grau de precisão possível, especialmente diante do avanço das técnicas de sequenciamento, que hoje permitem detectar variantes patogênicas inclusive em regiões não codificantes, como as intrônicas e de splicing.

Assim, o segundo capítulo deste trabalho tem como objetivo realizar uma análise comparativa para verificar a eficiência e precisão de três preditores específicos — AlphaMissense, SIFT e PolyPhen-2 — sendo o primeiro também utilizado na análise do capítulo anterior. O foco recai, posteriormente, na comparação entre os resultados obtidos pelo AlphaMissense e os valores de $\Delta\Delta G$ gerados pelo software FoldX.

Desta forma, entende-se que advento das tecnologias de sequenciamento de última geração avançou substancialmente a detecção de variantes no gene *CFTR* em todas as populações. A quantidade de dados de anotação genética e funcional representa um progresso que ressalta a necessidade urgente de ferramentas computacionais robustas, capazes de prever com precisão a patogenicidade de variantes de *CFTR* recém-identificadas (McDonald et al., 2024).

Para doenças complexas, especialmente para a FC, um grande conjunto de variantes foi catalogado em bancos de dados públicos, como o banco de dados CFTR2, que revela muitas variantes causadoras de FC, predominantemente intrônicas, com um subconjunto menor classificado como variantes *missense* exônicas, acrescentando uma camada de complexidade ao contexto clínico e de pesquisa. O desenvolvimento de modelos computacionais precisos para prever a patogenicidade dessas variantes é essencial para melhorar o diagnóstico e o aconselhamento genético, dada a extensa diversidade mutacional entre as populações (Bienvenu; Lopez; Girodon, 2020) .

6.1 PREVISÃO DE VARIANTES DE *CFTR* POR ALPHAMISSENSE, SIFT E POLYPHEN-2

Nossa avaliação comparativa do AlphaMissense, SIFT e PolyPhen-2 revelou diferenças significativas em seu desempenho ao classificar variantes *missense* no banco de dados CFTR.org. Utilizando dados clínicos e com curadoria do banco de dados CFTR2 como padrão de referência, avaliamos sistematicamente a precisão preditiva de cada ferramenta. Para as variantes causadoras de FC, os três modelos classificaram-nas consistentemente como patogênicas, com o SIFT demonstrando precisão semelhante ao AlphaMissense e ao PolyPhen-2. Essa consistência sugere que essas ferramentas podem identificar variantes com patogenicidade estabelecida. No entanto, ao prever variantes não causadoras de FC, surgiram disparidades notáveis no desempenho, indicando uma lacuna a ser preenchida para lidar com variantes benignas. Essas descobertas se alinham

com observações anteriores de alta taxa de falsos positivos e desempenho de classificação razoável em variantes de FC do CFTR2.org pelo AlphaMissense, relatadas por McDonald et al., 2024. As implicações de tais vieses são críticas, pois podem resultar em uma superestimação do impacto patogênico de variantes cujos efeitos clínicos não são bem definidos.

Isso ressalta a necessidade emergente de cautela ao interpretar os resultados das previsões *in silico*, especialmente em contextos clínicos nos quais as consequências da classificação incorreta podem ter impactos de longo alcance. Para abordar essas limitações, torna-se essencial integrar as previsões computacionais à validação experimental e a outros métodos de avaliação robustos, conforme relatado anteriormente por Gerasimavicius, Livesey e Marsh (2023).

A dependência exclusiva de ferramentas computacionais pode levar a imprecisões e interpretações errôneas, enfatizando a necessidade de uma abordagem multifacetada para a análise de variantes. Um único método de anotação pode resultar em associações perdidas, sugerindo que as características específicas do gene influenciam o desempenho de diferentes ferramentas. Ao integrar várias estratégias de anotação ou adaptar métodos com base em propriedades específicas do gene, a precisão das previsões de impacto das variantes pode ser significativamente aprimorada. Além disso, o emprego de uma estratégia abrangente que incorpore diversas técnicas de validação pode melhorar a precisão da classificação de variantes, mitigando os riscos associados à classificação incorreta e fortalecendo a confiança na interpretação clínica dos dados genéticos (Curtis, 2024).

Portanto, entende-se que esse cenário é particularmente desafiador em modelos preditivos de variantes genéticas, como no gene *CFTR*, em que variantes ambíguas podem exibir características mais sutis do que aquelas aproximadas por variantes benignas ou patogênicas (Bonetti; Tini; Mazarella, 2024). Isso pode dificultar a distinção entre essas classes, especialmente quando os dados de treinamento não capturam a variabilidade biológica das variantes ambíguas. Além disso, a presença de dados incompletos ou a falta de contextualização fenotípica pode contribuir para o desempenho insatisfatório do classificador. Essa dificuldade em classificar com precisão as variantes ambíguas é uma importante área de aprimoramento para modelos preditivos, que podem ser refinados para aumentar a especificidade e a sensibilidade na detecção de variantes de significado incerto (Harel et al., 2019).

Esse viés observado nos preditores aponta para um desafio fundamental no desenvolvimento e na aplicação de modelos de classificação de variantes genéticas. O erro na distinção entre variantes incertas verdadeiramente benignas e deletérias não apenas compromete o valor preditivo dos modelos, mas também pode afetar negativamente decisões clínicas, como escolhas terapêuticas e aconselhamento genético (Livesey; Marsh, 2023). Portanto, aprimorar os preditores para distinguir melhor as variantes clinicamente relevantes daquelas que não oferecem risco é uma etapa essencial para aumentar a confiabilidade das ferramentas de previsão. Isso requer estratégias como a personalização de modelos, o uso de bancos de dados mais abrangentes e a integração de métodos complementares que minimizem os vieses inerentes às classificações automatizadas (Walsh et al., 2024).

6.2 PREVISÃO DE PATOGENICIDADE DO CONJUNTO

Nosso estudo se baseia nessas percepções, oferecendo uma caracterização exclusiva de três ferramentas preditivas — SIFT, PolyPhen-2 e AlphaMissense — por meio de uma análise abrangente e integrativa. Embora Michels et al., 2019, tenham fornecido diretrizes essenciais para a avaliação de ferramentas preditivas, seu trabalho não incluiu o AlphaMissense, que surgiu como uma ferramenta amplamente utilizada com recursos de detecção aprimorados (Cheng et al., 2023). A incorporação do AlphaMissense em nossa análise introduz uma nova dimensão na avaliação das ferramentas de previsão de variantes, permitindo uma compreensão mais profunda de seus pontos fortes e fracos em comparação com os métodos amplamente utilizados no cenário de pesquisa.

McDonald et al., 2024, realizaram uma avaliação isolada do AlphaMissense, concentrando-se apenas em seu desempenho, sem explorar sua interação ou uso combinado com outras ferramentas de previsão. Ao comparar o AlphaMissense com o SIFT e o PolyPhen-2 no contexto de uma abordagem de conjunto, nosso estudo oferece uma perspectiva mais ampla, revelando como essas ferramentas se complementam ou entram em conflito umas com as outras na classificação de variantes. Essa análise integrada destaca as nuances de seu desempenho e enfatiza a importância do uso de múltiplas abordagens para melhorar a precisão da previsão, principalmente para categorias desafiadoras, como variantes benignas ou incertas.

Abordamos as lacunas na interpretação patogênica das variantes *CFTR* avaliando o desempenho de SIFT, PolyPhen-2 e AlphaMissense utilizando o banco de dados CFTR2, um recurso com curadoria de variantes *missense* clinicamente significativas. Além disso, desenvolvemos um modelo de conjunto para avaliar conjuntamente o potencial preditivo e as tendências. Essa abordagem integrada oferece novos insights sobre seus recursos combinados e destaca sua utilidade para a classificação de variantes em contextos de diagnóstico e prognóstico relacionados a variantes causadoras de FC.

Por fim, recomendamos cautela na interpretação clínica e no relato de pesquisa dos resultados derivados dessas ferramentas computacionais para o *CFTR*. Os vieses observados, especialmente na classificação errônea de variantes não causadoras de FC, ressaltam a importância de complementar os modelos preditivos com validação experimental e métodos de avaliação alternativos. Uma abordagem integrativa, incluindo dados multiômicos (como transcriptoma ou proteômica), pode apoiar e melhorar as decisões clínicas. Isso garante que tais decisões possam ser respaldadas por um processo de avaliação robusto, minimizando o risco de superestimar a patogenicidade de mutações no *CFTR* e otimizando a confiabilidade dos diagnósticos e prognósticos em diferentes grupos populacionais.

6.3 ANÁLISE DOS PREDITORES COM O FOLDX

Além dos modelos de predição baseados em aprendizado de máquina, a integração entre ferramentas estruturais e funcionais tem se mostrado uma estratégia eficaz para ampliar a acurácia na interpretação de variantes. Portanto, a classificação precisa de variantes *missense* ainda representa um dos principais desafios na genômica médica, sobretudo quando se trata de distinguir entre mutações patogênicas e variantes benignas em doenças monogênicas, como a fibrose cística (FC). Essa dificuldade reforça a necessidade de desenvolver e aplicar abordagens computacionais e funcionais mais robustas, com o objetivo de aprimorar a precisão dos diagnósticos genéticos, especialmente em contextos clínicos (Rastogi et al., 2025).

Dentre os preditores utilizados atualmente, ferramentas como SIFT, PolyPhen-2 e, mais recentemente, o AlphaMissense (AM), têm sido amplamente aplicadas. No entanto, a utilidade desses métodos varia, sobretudo quando avaliamos

diferentes tipos de variantes. O AlphaMissense se destaca por ser uma ferramenta baseada em aprendizado profundo, que integra informações de sequência, estrutura tridimensional e conservação evolutiva. Por isso, apresenta um desempenho superior ao de métodos que se baseiam exclusivamente em sequências (Cheng et al., 2023).

Neste estudo, buscamos avaliar comparativamente o desempenho dessas três ferramentas — AlphaMissense, SIFT e PolyPhen-2 — na classificação de variantes missense do gene CFTR com relevância clínica, utilizando como referência 164 variantes disponibilizadas pelo banco de dados CFTR2. Os resultados mostraram que o Um aspecto particularmente interessante foi a relação entre os scores de patogenicidade gerados pelo AlphaMissense e os valores de $\Delta\Delta G$ estimados com o FoldX, uma ferramenta que calcula a variação na energia livre da proteína em função das mutações. Observou-se que o AlphaMissense teve a correlação mais alta com os valores de $\Delta\Delta G$ ($r = 0,38$; $p \leq 3.8 \times 10^{-7}$), superando tanto o SIFT ($r = -0,22$) quanto o PolyPhen-2 ($r = 0,33$). Essa correlação sugere que as predições do AlphaMissense estão mais alinhadas com os efeitos estruturais e energéticos das variantes sobre a proteína CFTR.

As análises de regressão linear múltipla reforçaram a superioridade do AlphaMissense em relação aos demais preditores. Observou-se que, a cada unidade de aumento na pontuação do AlphaMissense, houve um aumento médio de aproximadamente 310 mil unidades nos valores de $\Delta\Delta G$ total_energy ($p \leq 8,08e-05$). Em contrapartida, os escores gerados pelo SIFT e pelo PolyPhen-2 não apresentaram associações estatisticamente significativas com os valores de $\Delta\Delta G$ ($p = 0,414$ e $p = 0,426$, respectivamente), sugerindo menor sensibilidade dessas ferramentas aos efeitos estruturais das variantes. Esses achados reforçam que, entre os preditores analisados, o AlphaMissense é o que melhor se alinha às variações energéticas previstas pelo FoldX, capturando com maior precisão o impacto funcional das mutações missense na proteína CFTR. Esses resultados sugerem que o AlphaMissense não apenas capta sinais de desestabilização proteica, mas também é capaz de refletir impactos funcionais mais amplos das variantes missense. Assim, a combinação entre as predições do AlphaMissense e os cálculos de energia livre fornecidos pelo FoldX se apresenta como uma abordagem complementar promissora.

Logo, por meio desta análise em geral verificamos que esta integração na análise e interpretação de variantes oferece maior sensibilidade, principalmente quando estamos tratando aquelas com significado clínico incerto, o que pode beneficiar diretamente a prática clínica e o aconselhamento genético em pacientes com fibrose

cística, visto que pode ser uma junção efetiva para facilitar a velocidade dos diagnósticos e o mecanismo da mutação, já que os efeitos da desestabilização da proteína divergem minimamente, principalmente tratando-se de variantes missense.

7. CONCLUSÃO GERAL

De modo geral, nossas análises demonstram que a identificação de possíveis novas variantes genéticas em populações indígenas contribui significativamente para a valorização e compreensão da diversidade genômica desses grupos, frequentemente sub-representados em estudos populacionais de larga escala. No contexto de doenças monogênicas como a fibrose cística, marcada por amplo espectro fenotípico e complexidade clínica, esses achados reforçam a importância de incluir grupos geneticamente diversos em investigações genômicas, sobretudo quando se busca precisão diagnóstica e desenvolvimento de abordagens terapêuticas mais individualizadas.

Nesse sentido, os preditores computacionais de patogenicidade se mostram ferramentas valiosas dentro da medicina de precisão, não apenas por sua aplicabilidade em larga escala e custo relativamente reduzido, mas também por sua capacidade de auxiliar na priorização de variantes com maior potencial de impacto funcional. No entanto, é importante destacar que esses algoritmos, quando utilizados de forma isolada, apresentam limitações que podem comprometer a interpretação clínica das variantes. Assim, nossos resultados indicam que o uso combinado de diferentes preditores — especialmente quando aliado a abordagens estruturais, como cálculos de energia livre — é fundamental para fortalecer a acurácia das análises *in silico* e aumentar a confiabilidade das inferências funcionais.

Portanto, conclui-se que a integração de dados preditivos e estruturais representa uma estratégia promissora para o refinamento da análise de variantes *missense*, com aplicações diretas no diagnóstico genético, aconselhamento familiar e desenvolvimento de terapias personalizadas. Tal abordagem se mostra ainda mais relevante quando aplicada a populações geneticamente distintas, como as indígenas, cujos perfis genéticos únicos podem revelar mecanismos ainda pouco explorados na fisiopatologia de doenças raras como a fibrose cística.

8. REFERÊNCIAS BIBLIOGRÁFICAS

- ABDELLAOUI, Abdel *et al.* 15 years of GWAS discovery: Realizing the promise. **The American Journal of Human Genetics**, v. 110, n. 2, p. 179–194, fev. 2023.
- ADZHUBEI, Ivan; JORDAN, Daniel M.; SUNYAEV, Shamil R. Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2. **Current Protocols in Human Genetics**, v. 76, n. 1, jan. 2013.
- ANZAI, Chieko *et al.* CFTR gene mutations in Japanese individuals with congenital bilateral absence of the vas deferens. **Journal of Cystic Fibrosis**, v. 2, n. 1, p. 14–18, mar. 2003.
- BACALHAU, Mafalda *et al.* Elexacaftor-Tezacaftor-Ivacaftor: A Life-Changing Triple Combination of CFTR Modulator Drugs for Cystic Fibrosis. **Pharmaceuticals**, v. 16, n. 3, p. 410, 8 mar. 2023.
- BARREÑADA, Lasai *et al.* Understanding overfitting in random forest for probability estimation: a visualization and simulation study. **Diagnostic and Prognostic Research**, v. 8, n. 1, p. 14, 27 set. 2024.
- BARROS, Maria Clara *et al.* Unraveling the protective genetic architecture of COVID-19 in the Brazilian Amazon. **Scientific Reports**, v. 14, n. 1, 9 nov. 2024.
- BAYNAM, Gareth *et al.* Indigenous Genetics and Rare Diseases: Harmony, Diversity and Equity. In: POSADA DE LA PAZ, Manuel; TARUSCIO, Domenica; GROFT, Stephen C. (Orgs.). **Rare Diseases Epidemiology: Update and Overview**. Advances in Experimental Medicine and Biology. Cham: Springer International Publishing, 2017. v. 1031 p. 511–520.
- BEZERRA, Vanessa Lima *et al.* ANÁLISE DA EPIDEMIOLOGIA DA FIBROSE CÍSTICA E A APLICAÇÃO DO DIAGNÓSTICO MOLECULAR NO BRASIL DE 2009 A 2020. **Revista Interfaces: Saúde, Humanas e Tecnologia**, v. 11, n. 4, 31 dez. 2023.
- BIENVENU, Thierry; LOPEZ, Maureen; GIRODON, Emmanuelle. Molecular Diagnosis and Genetic Counseling of Cystic Fibrosis and Related Disorders: New Challenges. **Genes**, v. 11, n. 6, p. 619, 4 jun. 2020.
- BOLGER, Anthony M.; LOHSE, Marc; USADEL, Bjoern. Trimmomatic: a flexible trimmer for Illumina sequence data. **Bioinformatics**, v. 30, n. 15, p. 2114–2120, 1 ago. 2014.
- BONETTI, Emanuele; TINI, Giulia; MAZZARELLA, Luca. Accuracy of renovo predictions on variants reclassified over time. **Journal of Translational Medicine**, v. 22, n. 1, p. 713, 31 jul. 2024.
- BORKENHAGEN, Beatrice; PREHM, Peter. Recovery of $\Delta F508$ -CFTR Function by Citrate. **Nutrients**, v. 14, n. 20, p. 4283, 14 out. 2022.

- BUSS, Oliver; RUDAT, Jens; OCHSENREITHER, Katrin. FoldX as Protein Engineering Tool: Better Than Random Based Approaches? **Computational and Structural Biotechnology Journal**, v. 16, p. 25–33, 2018.
- CASTELLANI, Carlo. CFTR2: How will it help care? **Paediatric Respiratory Reviews**, Royal Society of Medicine – The 26th symposium: Cystic fibrosis in children and adults, 20th November 2012. v. 14, p. 2–5, 1 maio 2013.
- CHARILAOU, Paris; BATTAT, Robert. Machine learning models and over-fitting considerations. **World Journal of Gastroenterology**, v. 28, n. 5, p. 605–607, 7 fev. 2022.
- CHENG, Jun *et al.* Accurate proteome-wide missense variant effect prediction with AlphaMissense. **Science**, v. 381, n. 6664, p. eadg7492, 22 set. 2023.
- CLANCY, J. P. *et al.* Clinical Pharmacogenetics Implementation Consortium (CPIC) Guidelines for Ivacaftor Therapy in the Context of CFTR Genotype. **Clinical Pharmacology & Therapeutics**, v. 95, n. 6, p. 592–597, jun. 2014.
- CROMWELL, Elizabeth A. *et al.* Cystic fibrosis prevalence in the United States and participation in the Cystic Fibrosis Foundation Patient Registry in 2020. **Journal of Cystic Fibrosis**, v. 22, n. 3, p. 436–442, maio 2023.
- COMPLEXO HOSPITALAR UNIVERSITÁRIO DA UFPA. Hospitais universitários – região Norte – Barros Barreto. Brasília, DF: Ebserh. Disponível em: <https://www.gov.br/ebserh/pt-br/hospitais-universitarios/regiao-norte/chu-ufpa/saude/especialidades/barros-barreto>. Acesso em: 30 de agosto de 2025.
- CURTIS, David. Assessment of ability of AlphaMissense to identify variants affecting susceptibility to common disease. **European Journal of Human Genetics**, v. 32, n. 11, p. 1419–1427, nov. 2024.
- DELETANG, Karine. Splicing mutations in the CFTR gene as therapeutic targets. **Gene Therapy**, 2022.
- DELGADO, Javier *et al.* FoldX force field revisited, an improved version. **Bioinformatics**, v. 41, n. 2, p. btaf064, 4 fev. 2025.
- DICKINSON, Kimberly M.; COLLACO, Joseph M. Cystic Fibrosis. **Pediatrics In Review**, v. 42, n. 2, p. 55–67, fev. 2021.
- DUCKERS, Jamie *et al.* Real-World Outcomes of Ivacaftor Treatment in People with Cystic Fibrosis: A Systematic Review. **Journal of Clinical Medicine**, v. 10, n. 7, p. 1527, 6 abr. 2021.
- ELLINGFORD, Jamie M. Recommendations for clinical interpretation of variants found in non-coding regions of the genome. 2022.
- ERNST, Corinna *et al.* Performance of in silico prediction tools for the classification of rare BRCA1/2 missense variants in clinical diagnostics. **BMC Medical Genomics**, v. 11, n. 1, dez. 2018.

- FARINHA, Carlos M.; CALLEBAUT, Isabelle. Molecular mechanisms of cystic fibrosis – how mutations lead to misfunction and guide therapy. **Bioscience Reports**, 2022.
- FATUMO, Segun; CHIKOWORE, Tinashe; KUCHENBAECKER, Karoline. Editorial: Genetics of Complex Traits and Diseases From Under-Represented Populations. **Frontiers in Genetics**, v. 12, p. 817683, 17 jan. 2022.
- FERNANDES, Marianne Rodrigues *et al.* Genetic Diversity of Drug-Related Genes in Native Americans of the Brazilian Amazon. **Pharmacogenomics and Personalized Medicine**, v. Volume 14, p. 117–133, jan. 2021.
- GARCIA, Felipe Antonio De Oliveira; ANDRADE, Edilene Santos De; PALMERO, Edenir Inez. Insights on variant analysis in silico tools for pathogenicity prediction. **Frontiers in Genetics**, v. 13, p. 1010327, 29 nov. 2022.
- GERASIMAVICIUS, Lukas; LIVESEY, Benjamin J.; MARSH, Joseph A. Loss-of-function, gain-of-function and dominant-negative mutations have profoundly different effects on protein structure. **Nature Communications**, v. 13, n. 1, p. 3895, 6 jul. 2022.
- GRIMM, Dominik G. *et al.* The Evaluation of Tools Used to Predict the Impact of Missense Variants Is Hindered by Two Types of Circularity. **Human Mutation**, v. 36, n. 5, p. 513–523, maio 2015.
- GUNNING, Adam C. *et al.* Assessing performance of pathogenicity predictors using clinically relevant variant datasets. **Journal of Medical Genetics**, v. 58, n. 8, p. 547–555, ago. 2021.
- HAO, Wei; STOREY, John D. Extending Tests of Hardy–Weinberg Equilibrium to Structured Populations. ago. 2019.
- HAREL, Tom *et al.* Predicting Phenotypic Diversity from Molecular and Genetic Data. **Genetics**, v. 213, n. 1, p. 297–311, 1 set. 2019.
- HASIAK SANTO¹, Augusto; RIBEIRO FERREIRA DA SILVA-FILHO^{2,3}, Luiz Vicente. Cystic fibrosis-related mortality trends in Brazil for the 1999-2017 period: a multiple-cause-of-death study. **Jornal Brasileiro de Pneumologia**, p. e20200166, 30 abr. 2021.
- HOSPITAL UNIVERSITÁRIO JOÃO DE BARROS BARRETO (HUIBB-UFPA). Teste do suor: HU Barros Barreto dispõe de exame para identificar a Fibrose Cística. Brasília, DF: Ebserh, 01 set. 2022 (atualizado em 08 dez. 2023). Disponível em: <https://www.gov.br/ebserh/pt-br/comunicacao/noticias/teste-do-suor-hospital-ebserh-mec-no-para-dispoe-de-exame-para-identificar-a-fibrose-cistica>. Acesso em: 30/08/2025.
- JÄNES, Jürgen *et al.* **Predicted mechanistic impacts of human protein missense variants**. Cold Spring Harbor Laboratory, , 29 maio 2024. Disponível em: <<http://biorxiv.org/lookup/doi/10.1101/2024.05.29.596373>>. Acesso em: 19 jul. 2025
- JIA, Shijing; TAYLOR-COUSAR, Jennifer L. Cystic Fibrosis Modulator Therapies. **Annual Review of Medicine**, v. 74, n. 1, p. 413–426, 27 jan. 2023.

JOYNT, Anya T.; CUTTING, Garry R.; SHARMA, Neeraj. Genetics of Cystic Fibrosis. **Clinics in Chest Medicine**, v. 43, n. 4, p. 591–602, dez. 2022.

JUMPER, John *et al.* Highly accurate protein structure prediction with AlphaFold. **Nature**, v. 596, n. 7873, p. 583–589, 26 ago. 2021.

KARCZEWSKI, Konrad J. The mutational constraint spectrum quantified from variation in 141,456 humans. *[S.d.]*.

KARIM, Md Rezaul *et al.* Explainable AI for Bioinformatics: Methods, Tools and Applications. **Briefings in Bioinformatics**, v. 24, n. 5, p. bbad236, 20 set. 2023.

KEHDY, Fernanda S. G. *et al.* Origin and dynamics of admixture in Brazilians and its effect on the pattern of deleterious mutations. **Proceedings of the National Academy of Sciences**, v. 112, n. 28, p. 8696–8701, 14 jul. 2015.

KUMAR, Prateek; HENIKOFF, Steven; NG, Pauline C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. **Nature Protocols**, v. 4, n. 7, p. 1073–1081, jul. 2009.

LANDRUM, Melissa J. *et al.* ClinVar: improving access to variant interpretations and supporting evidence. **Nucleic Acids Research**, v. 46, n. D1, p. D1062–D1067, 4 jan. 2018.

LAUSCHKE, Volker M.; INGELMAN-SUNDBERG, Magnus. Emerging strategies to bridge the gap between pharmacogenomic research and its clinical implementation. **npj Genomic Medicine**, v. 5, n. 1, p. 9, 5 mar. 2020.

LEVRING, Jesper *et al.* CFTR function, pathology and pharmacology at single-molecule resolution. **Nature**, v. 616, n. 7957, p. 606–614, 20 abr. 2023.

LI, Heng *et al.* The Sequence Alignment/Map format and SAMtools. **Bioinformatics**, v. 25, n. 16, p. 2078–2079, 15 ago. 2009.

LICHTER, David José Duque *et al.* Fibrose cística: uma revisão. **Brazilian Journal of Health Review**, v. 7, n. 5, p. 73568, 11 out. 2024.

LIVESEY, Benjamin J.; MARSH, Joseph A. Updated benchmarking of variant effect predictors using deep mutational scanning. **Molecular Systems Biology**, v. 19, n. 8, p. e11474, 8 ago. 2023.

LOPES-PACHECO, Miquéias. CFTR Modulators: The Changing Face of Cystic Fibrosis in the Era of Precision Medicine. **Frontiers in Pharmacology**, v. 10, p. 1662, 21 fev. 2020.

LÓPEZ-VALDEZ, Jaime A. *et al.* Cystic fibrosis: current concepts. **Boletín Médico del Hospital Infantil de México**, v. 78, n. 6, p. 6536, 20 dez. 2021.

LUZUM, Jasmine A. *et al.* Moving Pharmacogenetics Into Practice: It's All About the Evidence! **Clinical Pharmacology & Therapeutics**, v. 110, n. 3, p. 649–661, set. 2021.

MALAGUIAS, Soraya Beatriz Paula; CARDOSO, Alessandra Marques. Desafios no manejo da fibrose cística. **REVISTA BRASILEIRA MILITAR DE CIÊNCIAS**, v. 9, n. 23, 5 abr. 2023.

MCDONALD, Eli Fritz *et al.* Benchmarking AlphaMissense Pathogenicity Predictions Against Cystic Fibrosis Variants. **bioRxiv**, p. 2023.10.05.561147, 4 jan. 2024.

MCLAREN, William *et al.* The Ensembl Variant Effect Predictor. **Genome Biology**, v. 17, n. 1, p. 122, dez. 2016.

MICHELS, Marcus *et al.* Determining the pathogenicity of CFTR missense variants: Multiple comparisons of in silico predictors and variant annotation databases. **Genetics and Molecular Biology**, v. 42, n. 3, p. 560–570, set. 2019.

MOREIRA, Rennan Garcias *et al.* Population genetics of *PDE4B* (phosphodiesterase-4B) in neglected Native Americans: Implications for cancer pharmacogenetics. **Clinical and Translational Science**, v. 15, n. 6, p. 1400–1405, jun. 2022.

MOTA, Daniel Marques; VIGO, Álvaro; KUCHENBECKER, Ricardo De Souza. Formulários de notificação de reações adversas a medicamento nos sistemas de farmacovigilância do Brasil e outros doze países latino-americanos: análise comparativa. **Ciência & Saúde Coletiva**, v. 26, n. 4, p. 1245–1257, abr. 2021.

NASLAVSKY, Michel S. *et al.* Whole-genome sequencing of 1,171 elderly admixed individuals from Brazil. **Nature Communications**, v. 13, n. 1, p. 1004, 4 mar. 2022.

NG, P. C. SIFT: predicting amino acid changes that affect protein function. **Nucleic Acids Research**, v. 31, n. 13, p. 3812–3814, 1 jul. 2003.

NUNES, Kelly *et al.* Admixture's impact on Brazilian population evolution and health. **Science**, v. 388, n. 6748, 15 maio 2025.

ODDSSON, Asmundur *et al.* Deficit of homozygosity among 1.52 million individuals and genetic causes of recessive lethality. **Nature Communications**, v. 14, n. 1, p. 3453, 10 jun. 2023.

ODE, Katie Larson *et al.* Cystic fibrosis related diabetes: Medical management. **Journal of Cystic Fibrosis**, v. 18, p. S10–S18, out. 2019.

OSCANOA, Jorge *et al.* SNPnexus: a web server for functional annotation of human genome sequence variation (2020 update). **Nucleic Acids Research**, v. 48, n. W1, p. W185–W192, 2 jul. 2020.

PERRY, Luke A. *et al.* Topical cystic fibrosis transmembrane conductance regulator gene replacement for cystic fibrosis-related lung disease. **Cochrane Database of Systematic Reviews**, v. 2016, n. 7, 17 jun. 2016.

PESSÔA, Renata F.; NÁCUL, Flávio E.; NOËL, François. FARMACOGENÉTICA E FARMACOGENÔMICA. EVIDÊNCIAS DE COMO A GENÉTICA PODE INFLUENCIAR A EFICÁCIA DE FÁRMACOS E A BUSCA POR NOVOS ALVOS FARMACOLÓGICOS. 2006.

PLASSCHAERT, Lindsey W.; MACDONALD, Kelvin D.; MOFFIT, Jeffrey S. Current landscape of cystic fibrosis gene therapy. **Frontiers in Pharmacology**, v. 15, p. 1476331, 8 out. 2024.

- PREM, Pranjali *et al.* The Impact of Increased Homozygosity on Human Fertility: A Comprehensive Review. **Cureus**, 18 nov. 2023.
- RASTOGI, Ruchir *et al.* Critical assessment of missense variant effect predictors on disease-relevant variant data. **Human Genetics**, v. 144, n. 2–3, p. 281–293, mar. 2025.
- RIBEIRO-DOS-SANTOS, André M. *et al.* Exome Sequencing of Native Populations From the Amazon Reveals Patterns on the Peopling of South America. **Frontiers in Genetics**, v. 11, p. 548507, 29 out. 2020.
- RODEN, Dan M. *et al.* Pharmacogenomics. **The Lancet**, v. 394, n. 10197, p. 521–532, ago. 2019.
- RODRIGUES, Juliana Carla Gomes *et al.* Polymorphisms of ADME-related genes and their implications for drug safety and efficacy in Amazonian Amerindians. **Scientific Reports**, v. 9, n. 1, p. 7201, 10 maio 2019.
- RODRIGUES, Juliana Carla Gomes *et al.* Pharmacogenomic Profile of Amazonian Amerindians. **Journal of Personalized Medicine**, v. 12, n. 6, p. 952, 10 jun. 2022.
- RODRIGUES-SOARES, Fernanda; KEHDY, Fernanda S. G.; GOUVEIA, Mateus H. Editorial: Genomic Ancestry and Biological Traits. **Frontiers in Genetics**, v. 12, p. 754725, 6 set. 2021.
- ROSA, Fernanda Ribeiro *et al.* Fibrose cística: uma abordagem clínica e nutricional. **Revista de Nutrição**, v. 21, n. 6, p. 725–737, dez. 2008.
- SAFERALI, Aabida *et al.* *C FTR* variants are associated with chronic bronchitis in smokers. **European Respiratory Journal**, v. 60, n. 2, p. 2101994, ago. 2022.
- SAINZ DE MEDRANO SAINZ, Jaime I.; BRUNET SERRA, Mercè. Influencia de la farmacogenética en la diversidad de respuesta a las estatinas asociada a las reacciones adversas. **Advances in Laboratory Medicine / Avances en Medicina de Laboratorio**, v. 4, n. 4, p. 353–364, 18 dez. 2023.
- SANAVIA, Tiziana *et al.* Limitations and challenges in protein stability prediction upon genome variations: towards future applications in precision medicine. **Computational and Structural Biotechnology Journal**, v. 18, p. 1968–1979, 2020.
- SAPOZHNIKOV, Yesol *et al.* Statistical modeling to quantify the uncertainty of FoldX-predicted protein folding and binding stability. **BMC Bioinformatics**, v. 24, n. 1, 12 nov. 2023.
- SCOTET, Virginie; L’HOSTIS, Carine; FÉREC, Claude. The Changing Epidemiology of Cystic Fibrosis: Incidence, Survival and Impact of the CFTR Gene Discovery. **Genes**, v. 11, n. 6, p. 589, 26 maio 2020.
- SHARMA, N.; CUTTING, G. R. The genetics and genomics of cystic fibrosis. **Journal of Cystic Fibrosis**, v. 19, p. S5–S9, mar. 2020.
- SHEU, Ruey-Kai; PARDESHI, Mayuresh Sunil. A Survey on Medical Explainable AI (XAI): Recent Progress, Explainability Approach, Human Interaction and Scoring System. **Sensors**, v. 22, n. 20, p. 8068, 21 out. 2022.

- SIDDIK, Abu Bakar; BADAL, Faisal R.; ISLAM, Afroza. **Comparative Performance of Machine Learning Algorithms for Early Genetic Disorder and Subclass Classification**. arXiv, , 3 dez. 2024. Disponível em: <<http://arxiv.org/abs/2412.02189>>. Acesso em: 16 jul. 2025
- SIM, Ngak-Leng *et al.* SIFT web server: predicting effects of amino acid substitutions on proteins. **Nucleic Acids Research**, v. 40, n. W1, p. W452–W457, 1 jul. 2012.
- SIRUGO, Giorgio; WILLIAMS, Scott M.; TISHKOFF, Sarah A. The Missing Diversity in Human Genetic Studies. **Cell**, v. 177, n. 1, p. 26–31, mar. 2019.
- SOBITAN, Adebisi *et al.* Prediction of the Effects of Missense Mutations on Human Myeloperoxidase Protein Stability Using In Silico Saturation Mutagenesis. **Genes**, v. 13, n. 8, p. 1412, 8 ago. 2022.
- SOUZA, Aracele Maria De *et al.* A systematic scoping review of the genetic ancestry of the Brazilian population. **Genetics and Molecular Biology**, v. 42, n. 3, p. 495–508, set. 2019.
- STOLLAR, Elliott J.; SMITH, David P. Uncovering protein structure. **Essays in Biochemistry**, v. 64, n. 4, p. 649–680, 8 out. 2020.
- SUAREZKURTZ, G. Pharmacogenomics in admixed populations. **Trends in Pharmacological Sciences**, v. 26, n. 4, p. 196–201, abr. 2005.
- TORDAI, Hedvig *et al.* Analysis of AlphaMissense data in different protein groups and structural context. **Scientific Data**, v. 11, n. 1, p. 495, 14 maio 2024.
- VAN GOOR, Fredrick *et al.* Effect of ivacaftor on CFTR forms with missense mutations associated with defects in protein processing or function. **Journal of Cystic Fibrosis**, v. 13, n. 1, p. 29–36, jan. 2014.
- WALSH, Nicola *et al.* Variant reclassification and clinical implications. **Journal of Medical Genetics**, v. 61, n. 3, p. 207–211, mar. 2024.
- WANG, Juexin *et al.* Exploring Human Diseases and Biological Mechanisms by Protein Structure Prediction and Modeling. In: SHEN, Bairong; TANG, Haixu; JIANG, Xiaoqian (Orgs.). **Translational Biomedical Informatics**. Advances in Experimental Medicine and Biology. Singapore: Springer Singapore, 2016. v. 939 p. 39–61.
- WHIRL-CARRILLO, M. *et al.* Pharmacogenomics Knowledge for Personalized Medicine. **Clinical Pharmacology & Therapeutics**, v. 92, n. 4, p. 414–417, out. 2012.
- WU, Malinda *et al.* Racial inequities and rare CFTR variants: Impact on cystic fibrosis diagnosis and treatment. **Journal of Clinical & Translational Endocrinology**, v. 36, p. 100344, jun. 2024.
- WU, You; XIE, Lei. **AI-driven multi-omics integration for multi-scale predictive modeling of causal genotype-environment-phenotype relationships**. arXiv, , 8 jul. 2024. Disponível em: <<http://arxiv.org/abs/2407.06405>>. Acesso em: 15 jan. 2025
- YU, He *et al.* Paleolithic to Bronze Age Siberians Reveal Connections with First Americans and across Eurasia. **Cell**, v. 181, n. 6, p. 1232–1245.e20, jun. 2020.

