UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

AVALIAÇÃO DE VARIANTES GENÉTICAS NO GENE REGULADOR DA CONDUTÂNCIA TRANSMEMBRÂNICA DA FIBROSE CÍSTICA (*CFTR*) ATRAVÉS DE PREDITORES COMPUTACIONAIS (*IN SILICO*)

MARCUS SILVA MICHELS

Orientadora: Dra. Maria Teresa Vieira Sanseverino

Co-orientadora: Dra. Marina Siebert

Porto Alegre, maio de 2017

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

AVALIAÇÃO DE VARIANTES GENÉTICAS NO GENE REGULADOR DA CONDUTÂNCIA TRANSMEMBRÂNICA DA FIBROSE CÍSTICA (*CFTR*) ATRAVÉS DE PREDITORES COMPUTACIONAIS (*IN SILICO*)

MARCUS SILVA MICHELS

Dissertação submetida ao Programa de Pós-Graduação em Genética e Biologia Molecular da UFRGS como requisito parcial para a obtenção do grau de Mestre em Genética e Biologia Molecular.

Orientadora: Dra. Maria Teresa Vieira Sanseverino
Co-orientadora: Dra. Marina Siebert

Porto Alegre, maio de 2017

INSTITUIÇÃO E FONTES FINANCEIRAS

Este trabalho foi realizado no Serviço de Genética Médica do Hospital de Clínicas de Porto Alegre (HCPA), com colaboração da Unidade de Análises Moleculares e de Proteínas (UAMP) do HCPA e do Centro de Biotecnologia da Universidade Federal do Rio Grande do Sul (Cbiot-UFRGS), com apoio financeiro do Fundo de Incentivo à Pesquisa e Eventos (FIPE-HCPA).

O aluno recebeu bolsa de mestrado concedida pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) vinculada ao Programa de Pós-Graduação em Genética e Biologia Molecular da UFRGS (PPGBM-UFRGS).

DEDICATÓRIA

Dedico este estudo a todos os pacientes afetados pela fibrose cística, pois o diagnóstico dessa doença não é uma sentença e sim uma oportunidade de viver. Apesar da intensa rotina de cuidados e tratamento, o diagnóstico precoce é a melhor forma de dar às pessoas afetadas pela fibrose cística a possibilidade de sonhar e buscar seus objetivos, pois essa doença não pode limitá-los enquanto seres humanos.

AGRADECIMENTOS

A minha orientadora, Dra. Maria Teresa Vieira Sanseverino, pela resiliência. Sempre buscas o melhor para quem está ao teu redor, e comigo não é diferente.

A minha co-orientadora, Dra. Marina Siebert, que encontra sempre uma forma de fazer com que eu me sinta melhor. Sempre tem uma palavra de encorajamento e é um exemplo de profissional.

À Dra. Ursula Matte, pela contribuição significativa na concepção deste projeto de pesquisa.

À M.Sc. Aline Castello Branco Mancuso, por sua disponibilidade e empenho em facilitar o entendimento da estatística.

À minha família, que sempre me deu suporte incondicional e é o motivo da minha existência.

Aos meus amigos, Igor e Ethiane, por nunca terem deixado de estar comigo mesmo não estando perto. Tenho sorte de tê-los sempre comigo, mesmo que de longe.

Ao meu amor, pela sua incansável determinação e perseverança.

SUMÁRIO

LISTA DE ABREVIATURAS

ABC	“ATP-binding cassette”
ACMG	“American College of Medical Genetics and Genomics”
AMP	“Association for Molecular Pathology”
ATP	“Adenosine triphosphate” – Adenosina trifosfato
AUC	“Area under the curve”
Ca^{2+}	Íon cálcio
cAMP	“Cyclic adenosine monophosphate” - Adenosina monofosfato cíclico
CBAVD	“Congenital Bilateral Absence of Vas Deferens”
CF	“Cystic fibrosis”
CFF	“Cystic Fibrosis Foundation”
CFTR	Regulador da condutância transmembrânica da fibrose cística
CFTR	Gene do regulador da condutância transmembrânica da fibrose cística
CFTR1	“Cystic Fibrosis Mutation Database”
CFTR2	“The Clinical and Functional Translation of CFTR”
Cl^-	Íon cloreto
CNV	“Copy number variation”
del	Deleção
DNA	“Deoxyribonucleic acid” – Ácido desoxirribonucleico
DPTN	Diferença de Potencial Transepitelial Nasal
FAMED	Faculdade de Medicina
FC	Fibrose cística
GBEFC	Grupo Brasileiro de Estudos de Fibrose Cística
GPPG	Grupo de Pesquisa e Pós-Graduação
H_2O	Água
HCO_3^-	Íon bicarbonato
HCPA	Hospital de Clínicas de Porto Alegre
HGMD	“Human Gene Mutation Database”
ICM	“Intestinal current measurement”
IMC	Índice de Massa Corporal

IRT	“Immunoreactive trypsinogen”
K^+	Íon potássio
kb	quilobases
kDa	quiloDalton
LSVA	Líquido da Superfície da Via Aérea
MCI	Mensuração da Corrente Intestinal
MD	*Medicinae doctor*
miRNA	microRNA
mRNA	“messenger RNA” – RNA mensageiro
MSD1	“Membrane-spanning domain 1”
MSD2	“Membrane-spanning domain 2”
MVCC	“Mutation of varying clinical consequence”
NBD1	“Nucleotide-binding domain 1”
NBD2	“Nucleotide-binding domain 2”
NPD	“Nasal potential difference”
OMIM	“Online Mendelian Inheritance in Man”
P	Fosforilação
Pi	Fosfato Inorgânico
PUCRS	Pontifícia Universidade Católica do Rio Grande do Sul
RD	“Regulatory domain”
RNA	“Ribonucleic acid” – Ácido ribonucleico
ROC	“Receiver operating characteristic curve”
SNP	“Single nucleotide polymorphism”
Supp.	“Supplementary”
UFRGS	Universidade Federal do Rio Grande do Sul
VCCV	Variante de Consequência Clínica Variável
WHO	“World Health Organization”

LISTA DE FIGURAS

Figuras do Manuscrito

LISTA DE TABELAS

Tabelas do Manuscrito

RESUMO

A Fibrose Cística (FC) é uma doença monogênica com padrão de herança autossômico recessivo. Mutações no gene da Condutância Transmembrânica da Fibrose Cística (*CFTR*) causam alterações no canal de cloreto codificado por esse gene, o que aumenta a viscosidade do muco, levando às manifestações clínicas observadas nos pacientes com FC. Até o momento, mais de 2.000 variantes estão descritas no "Cystic Fibrosis Mutation Database" e quase 40% delas causam a substituição de um aminoácido por outro. Como a maioria dessas variantes "missense" não tem sua patogenicidade confirmada, a análise *in silico* dessas variantes poderia ajudar no diagnóstico molecular e no aconselhamento genético. Contudo, preditores *in silico* podem divergir durante a interpretação de variantes genéticas, gerando resultados que não contribuem para a interpretação da patogenicidade das mutações no *CFTR*. Neste estudo, a patogenicidade de 779 variantes "missense" no gene *CFTR* foi estabelecida pelo preditor consensual PredictSNP e comparada a dois bancos de anotação de variantes (CFTR2 e ClinVar). A concordância Kappa entre todos preditores foi fraca a moderada. O mesmo se aplica para os preditores em relação ao CFTR2 e ao ClinVar. A análise da acurácia esperada mostrou que o PredictSNP e o PhDSNP tiveram as melhores acurácias quando comparados aos outros preditores. As substituições do aminoácido isoleucina por valina foram associadas ao desfecho "neutro" enquanto que as substituições de leucina por prolina foram associadas ao desfecho "deletério". Variantes que causam a substituição de aminoácidos localizados nos domínios de ligação a nucleotídeo e em β-folhas estão associadas com concordância completa entre os preditores ao passo que variantes que geram troca de aminoácidos localizados no domínio transmembrânico 1 e em α-hélices estão associadas a pelo menos uma discordância entre os preditores, independentemente da posição. A análise de sensibilidade e especificidade relevou que o PhDSNP e o PredictSNP são os melhores preditores para analisar variantes "missense" no *CFTR* enquanto que o nsSNPAnalyzer não seria recomendado para analisar as mesmas variantes. Esses resultados corroboram dados da literatura ao mostrar que os preditores *in silico* divergem quando analisam a patogenicidade de variantes "missense" no *CFTR*. Também se observou que eles divergem dos bancos de anotação de variantes. Por isso, os resultados gerados pelos preditores

devem ser usados com cautela e os dados clínicos devem ser levados em considera
anotação de variantes no *CFTR*.

Palavras-chave: Fibrose cística; *CFTR*; variantes “missense”; predição *in silico*; PredictSNP; CFTR2; ClinVar

ABSTRACT

Cystic Fibrosis (CF) is a monogenic disease with autosomal recessive model of inheritance. Mutations in the Cystic Fibrosis Transmembrane Conductance Regulator gene (*CFTR*) cause an impairment in the chloride channel encoded by this gene, which increases the viscosity of the mucus, leading to the clinical manifestations observed in CF patients. To date, over 2000 variants have been described at the Cystic Fibrosis Mutation Database and almost 40% of them cause an amino acid substitution. Since most of these missense variants do not have their pathogenicity confirmed, *in silico* analysis of these variants could help on molecular diagnosis and genetic counseling. However, *in silico* predictors may diverge when interpreting genetic variants, generating results that do not contribute to the pathogenicity interpretation of *CFTR* mutations. In this study, the pathogenicity of 779 *CFTR* missense variants was established by consensus predictor PredictSNP and compared to two variant annotation databases (CFTR2 and ClinVar). Kappa agreement between all predictors was fair or moderate. The same applies to predictors and CFTR2/ClinVar. Analysis of expected accuracy showed that PredictSNP and PhDSNP had greater accuracies when compared to other predictors. Amino acid substitutions of isoleucine for valine were associated with neutral outcomes while the substitutions of leucine for proline were associated with deleterious outcomes. Variants that cause amino acid substitutions located in nucleotide-binding domains and β-strands are associated with full agreement between predictors while variants that cause amino acid changes located in the membrane-spanning domain 1 and in α-helices are associated with at least one disagreement between predictors, regardless of position. Sensitivity and specificity analysis showed that PhDSNP and PredictSNP are the best predictors to analyze *CFTR* missense variants whilst nsSNPAnalyzer would not be recommended to analyze the same variants. These results corroborate literature data showing that in silico predictors diverge when analyzing the pathogenicity of *CFTR* missense variants. Also, it has been observed that they diverge from variant annotation databases. Hence, their results should be treated carefully and clinical data should be taken into account when asserting variant annotation for *CFTR*.

Keywords: Cystic fibrosis; *CFTR*; missense variants; *in silico* prediction; PredictSNP; CFTR2; ClinVar.

1. INTRODUÇÃO

1.1. Fibrose Cística

A fibrose cística (FC - OMIM #219700) ou mucoviscidose é uma doença monogênica que afeta mais de 70 mil indivíduos em todo o mundo, a maioria crianças. No Brasil, mais de 75% dos indivíduos acometidos pela FC têm até 18 anos (GBEFC 2014; Cutting 2015). Ela acomete principalmente populações eurodescendentes, mas se manifesta em todas as etnias (WHO 2004; Cutting 2015). Na Europa, 1 a cada 2000-3000 recém-nascidos é diagnosticado como afetado pela FC. Nos Estados Unidos, a incidência de FC é de 1 para cada 3500 nascidos (WHO 2004; Saraiva-Pereira et al. 2011; Fanen et al. 2014; Brennan & Schrijver 2016). No Brasil, devido à grande heterogeneidade populacional, a incidência varia de 1 para cada 1600 nascidos em eurobrasileiros, a 1 para cada 14000 nascidos em afro-brasileiros (Silva Filho et al. 2016). Essa doença se caracteriza pelo excesso de viscosidade da secreção nas vias aéreas e nos ductos pancreáticos (Saraiva-Pereira et al. 2011; Cutting 2015; Brennan & Schrijver 2016). Isso gera obstrução dessas vias, culminando em inflamação, dano e destruição tecidual. Também são afetados outros órgãos que contém epitélio, como glândulas sudoríparas, ductos biliares do fígado, trato reprodutor masculino e intestino (Lucarelli et al. 2012; Bergougnoux et al. 2015; Cutting 2015).

A FC é causada por mutações no gene *CFTR* (“Cystic Fibrosis Transmembrane Conductance Regulator”), o qual codifica para o Regulador da Condutância Transmembrânica da Fibrose Cística. O padrão de herança é autossômico e recessivo, sendo necessário que haja mutações nos dois alelos do *CFTR* para que se tenha a doença (Saraiva-Pereira et al. 2011; Fanen et al. 2014; Cutting 2015; Brennan & Schrijver 2016). Dependendo da combinação de mutações presente em cada alelo, a proteína gerada tem sua função diminuída ou não é funcional, o que leva ao surgimento dos fenótipos observados na FC (Cutting 2015). Pelo menos uma mutação de perda de função em cada alelo do *CFTR* é necessária para que o paciente apresente um fenótipo mais grave, a FC clássica. Caso haja apenas mutações de perda parcial de função em ambos alelos e/ou mutação de perda de função em só uma cópia do *CFTR*, a FC se manifesta como um fenótipo brando

(FC não-clássica ou “atípica”). Geralmente o que ocorre na FC não-clássica é o comprometimento de parte dos tecidos e órgãos onde a expressão do gene *CFTR* é crítica (Groman et al. 2005; Simon 2010).

1.2. Sintomatologia

Os sinais e sintomas podem variar muito nos pacientes fibrocísticos, tanto temporalmente quanto fenotipicamente. Enquanto uns manifestam sintomas já ao nascimento, outros podem viver anos sem manifestar características clínicas de FC (Farrell et al. 2017; Sosnay et al. 2017). A **Figura 1** traz os principais órgãos comprometidos pela FC e os sintomas associados a eles.

Dentre os órgãos afetados pela doença, os pulmões e vias aéreas são os mais comprometidos, sendo que a doença obstrutiva pulmonar é atualmente a maior causa de morbidade e é responsável por aproximadamente 80% das mortes de pacientes com FC (CFF 2013; Cutting 2015). Consoante a isso, o Registro Brasileiro de Fibrose Cística de 2014 reporta que houve 46 óbitos dentre os 2.571 pacientes com FC acompanhados no mesmo ano pelos Centros de Referência em FC do Brasil. Dentre eles, 31 tiveram a insuficiência respiratória como causa do óbito (GBEFC 2014). Clinicamente, a perda da função pancreática exócrina resulta em desnutrição e atraso no crescimento, o que leva à morte na primeira década de vida em pacientes não tratados (CFF 2013; Cutting 2015).

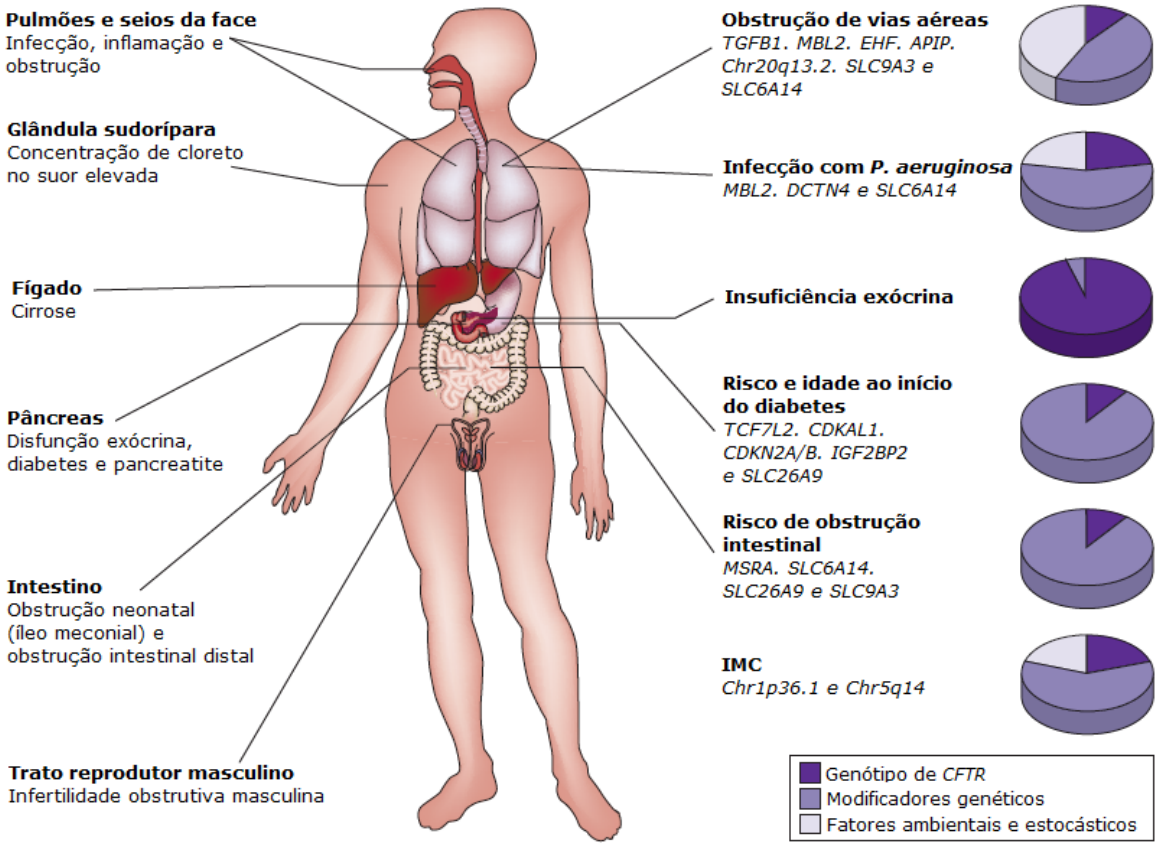


Figura 1. Características clínicas da fibrose císticas e contribuição relativa de modificadores genéticos para a variação em traços selecionados da fibrose cística. O diagnóstico de fibrose cística é baseado na presença dos achados clínicos mostrados à esquerda concomitante à concentração elevada de cloreto no suor (>60 mM). O grau de disfunção no sistema de órgãos varia consideravelmente entre indivíduos afetados. Ambos modificadores genéticos e fatores não-genéticos contribuem para a obstrução de vias aéreas e infecção por *Pseudomonas aeruginosa* — dois traços que definem a doença pulmonar na fibrose cística. O genótipo do gene regulador da condutância transmembrânica da fibrose cística (*CFTR*) é o primeiro determinante do grau de disfunção exócrina do pâncreas. A presença de variantes de *CFTR* associadas com disfunção exócrina pancreática grave é essencialmente um pré-requisito para o desenvolvimento de diabetes e obstrução intestinal. No contexto de disfunção endócrina grave, os modificadores genéticos determinam quando, e se, vai ocorrer diabetes e obstrução intestinal neonatal. A variação genética desempenha um papel predominante no estado nutricional como avaliado pelo índice de massa corporal (IMC). Adaptado de Cutting 2015.

1.3. Gene *CFTR*

O gene *CFTR* está localizado no braço longo do cromossomo 7 (*locus* 7q31.2) e apresenta 27 éxons distribuídos ao longo de aproximadamente 250 kb de DNA genômico. Após transcrição e processamento, resta um mRNA maduro de cerca de 6,5 kb, que é traduzido em uma proteína de 1480 aminoácidos com peso molecular de 168 kDa (Saraiva-Pereira et al. 2011).

Tal proteína (CFTR) é um canal de cloreto da superfamília ABC (“ATP binding cassette”), cuja regulação é feita por cAMP (adenosina monofosfato cíclico). Nesse canal, tanto a extremidade N-terminal quanto a C-terminal estão voltadas para o citoplasma. Além disso, ele possui 5 domínios proteicos: dois domínios transmembrânicos (MSD – “membrane-spanning domain”) de 6 subunidades cada; dois domínios citoplasmáticos de interação com ATP (NBD – “nucleotide binding domain”); e um domínio citoplasmático regulatório (RD), intrinsicamente desorganizado, que possui sítios de fosforilação para cinases – **Figura 2** (Saraiva-Pereira et al. 2011; Cutting 2015).

Esse transportador de cloreto se localiza na membrana apical de células epiteliais e auxilia no controle da quantidade de água no meio extracelular. Dependendo do tecido em que é expresso, o regulador da condutância transmembrânica da FC tem uma função diferente (Lucarelli et al. 2012; Wang et al. 2014; Cutting 2015). Por exemplo, no intestino e nas vias aéreas, o canal secreta cloreto para o lúmen (**Figura 3**). Já nas glândulas sudoríparas, ele atua reabsorvendo cloreto. Ainda, o CFTR tem uma função adicional nos pulmões: corregular o transporte de Na^+ ao interagir com o canal de sódio epitelial (ENaC) (Saraiva-Pereira et al. 2011; Lucarelli et al. 2012; Wang et al. 2014; Cutting 2015).

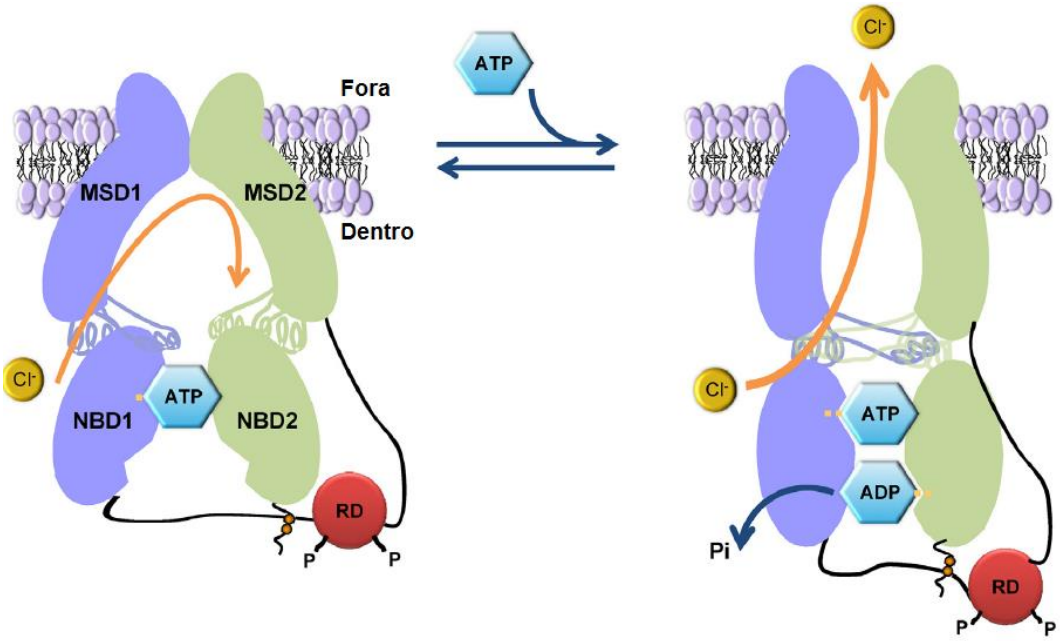


Figura 2. O ATP promove mudanças conformacionais que abrem o poro da proteína CFTR. Os modelos simplificados acima mostram o canal de cloreto CFTR nas configurações fechada (esquerda) e aberta (direita). MSD: "membrane-spanning domain"; NBD: "nucleotide-binding domain"; P: fosforilação do RD; Pi: fosfato inorgânico; RD: "regulatory domain". "Dentro" e "Fora" denotam os lados intra- e extracelular da membrana plasmática, respectivamente. Adaptado de Wang et al. 2014.

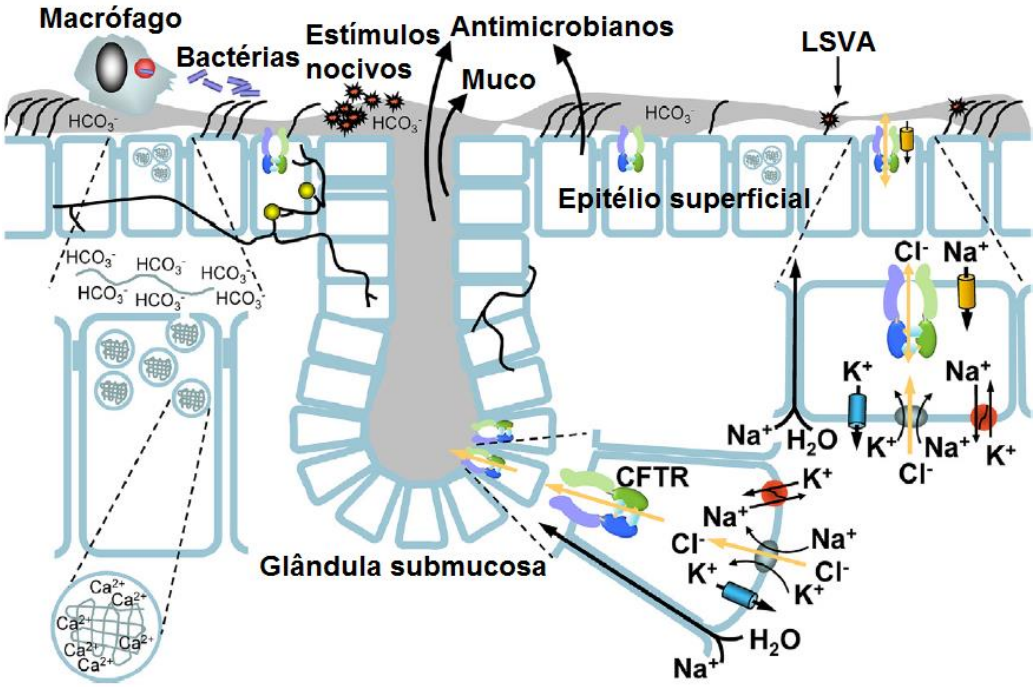


Figura 3. Papel de CFTR no transporte de íons transepitelial e defesa do hospedeiro nas vias aéreas respiratórias. A representação esquemática mostra uma fina camada de líquido da superfície da via aérea (LSVA) cobrindo o epitélio superficial e uma glândula submucosa. A atividade de CFTR e o canal de sódio epitelial (ENaC) na membrana apical controla a quantidade e a composição do LSVA. Adaptado de Wang et al. 2014.

1.3.1. Mutações no gene *CFTR*

De acordo com a Base de Dados de Mutações da Fibrose Cística ("Cystic Fibrosis Mutation Database" - http://www.genet.sickkids.on.ca/app), há atualmente 2.019 variantes listadas (CFTR1 1989). A mutação mais comum nesse gene é a deleção do códon 508 (CTT) que leva a não incorporação de uma fenilalanina durante a síntese proteica - c.1521_1523delCTT, p.Phe508del (F508del) (CFTR1 1989; Cutting 2015). Essa alteração de sequência, que ocorre no éxon 11 (domínio NBD1) do *CFTR*, está presente em 70% dos cromossomos de pacientes com FC em todo o mundo. Outras mutações no *CFTR* incluem: polimorfismos de nucleotídeo único (SNP), inserções, outras deleções, variação no número de cópias (CNV) de repetições em *tandem*. Somente outras quatro variantes representam mais de 1% dos casos de FC. São elas: p.Gly542X, p.Gly551Asp, p.Asn1303Lys, e p.Trp1282X (CFTR1 1989; Saraiva-Pereira et al. 2011; Wang et al. 2014; Cutting 2015).

Todas essas variantes têm consequências funcionais diversas e prevalências que variam entre as diferentes etnias (Cutting 2015). Estima-se que 40% delas causam a substituição de um aminoácido na proteína, 36% alteram o processamento do RNA (incluindo mutações sem sentido, de troca da fase de leitura e de erro no "splicing"), aproximadamente 3% envolvem grandes rearranjos no gene *CFTR*, 1% afeta a região promotora, 14% são variantes neutras e 6% têm efeito indeterminado até o momento. (CFTR1 1989; Cutting 2015). As mutações que causam FC podem afetar a quantidade e/ou a função do CFTR na membrana celular. Essas mutações estão divididas em seis classes (**Figura 4**) (Trujillano et al. 2013; Fanen et al. 2014; Wang et al. 2014; Martins et al. 2014; Cutting 2015). São elas:

- Classe I - Defeito na produção da proteína: grandes deleções, mutações "nonsense", "frameshift" e que causam erro no padrão de "splicing" geram códons de parada prematura, bloqueando a tradução e acelerando a degradação do mRNA. Exemplos: p.Gly542X; p.Trp128X; p.Arg553X; c.489+1G>T ("legacy name": 621+1G→T).
- Classe II - Defeito no processamento da proteína: mutações "missense" e deleções em fase podem gerar mal processamento da proteína CFTR, resultando na ausência de proteína funcional na membrana plasmática. Exemplos: p.Phe508del; p.Ile507del; p.Asn1303Lys.

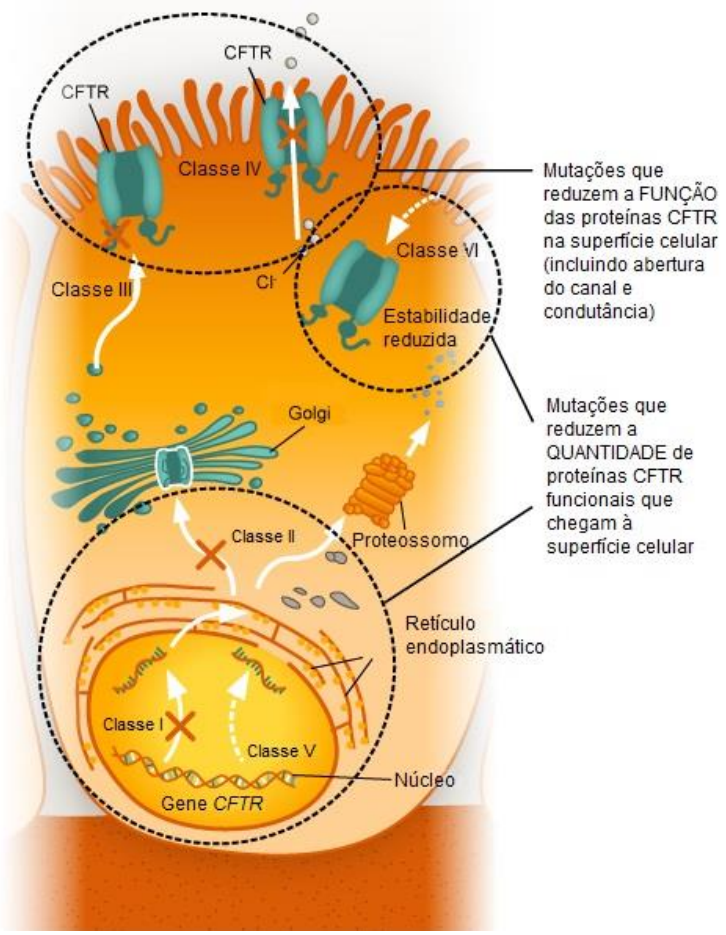


Figura 4. Classes de mutações no gene *CFTR* e o que elas afetam. É possível observar que as mutações de classe I, II, V e VI afetam a quantidade de CFTR funcional, ao passo que as mutações de classe III e IV afetam a função da proteína CFTR. Fonte: Adaptado de CFTR.info. Disponível em: <http://www.cftr.info/wp-content/uploads/m1_s1_03_03_step8 _just_right_ v23.jpg >. Acesso em 23/04/2017.

- Classe III - Defeito na regulação do canal: mutações dessa classe, geralmente "missense", não afetam a estrutura tampouco o processamento da proteína CFTR. Todavia, afetam o padrão de abertura e fechamento desse canal de cloreto. Exemplos: p.Gly551Asp; p.Gly178Arg; p.Gly551Ser; p.Ser549Asn.
- Classe IV - Defeito na condutância do canal: geralmente mutações "missense" que podem perturbar o fluxo de cloreto através do poro do canal. Exemplos: p.Arg117His; p.Arg347Pro; p.Arg117Cys; p.Arg334Trp.
- Classe V - Síntese proteica reduzida: mutações em sítios de "splicing" podem alterar a transcrição ou gerar "splicing" alternativo do *CFTR*, formando canais de cloreto funcionais, mas em menor quantidade que o fenótipo selvagem. Exemplos: c.3717+12191C>T ("legacy name": 3849+10kbC→T); c.2657+5G>A ("legacy name":

2789+5G→A); c.2988+1G>A (“legacy name”: 3120+1G→A); c.1210−12[5] (“legacy name”: 5T).

- Classe VI - Estabilidade reduzida da proteína: mutações “nonsense” e “frameshift” que aceleram o “turnover” do CFTR, diminuindo sua meia-vida. Exemplos: p.Cys1400X ou c.4196_4197delTC (“legacy name”: 4326delTC); p.Gln1412X; p.Ile1383Asn*fs*X3 ou c.4147_4148insA (“legacy name”: 4279insA).

1.3.1.1. O polimorfismo c.1210-12T[5_9] (trato poli-T)

Alelos complexos são aqueles que têm duas variantes em um cromossomo sendo que uma modula o efeito da outra (CFTR2 2011; Brennan & Schrijver 2016). Nesse contexto, o polimorfismo c.1210-12T[5_9], também conhecido como trato poli-T, está presente em toda cópia do gene *CFTR* em uma destas três formas: 5T (c.1210-12T[5]), 7T c.1210-12T[7] e 9T (c.1210-12T[9]). Ele está localizado na região consenso rica em pirimidinas próxima ao sítio aceptor de “splicing” no íntron 9 (junção íntron 9 - éxon 10). Dentre os três alelos possíveis para o trato poli-T, 5T afeta a eficiência de “splicing” do éxon 10 de maneira contexto-dependente. Essa variante está presente em pelo menos 5% dos alelos de *CFTR*, e está associada à penetrância parcial da FC (CFTR2 2011; Brennan & Schrijver 2016).

Quando 5T está disposta em *cis* com a variante p.Arg117His (c.350G>A), ela age como um modificador intragênico, ocorrendo o surgimento de um fenótipo de FC se houve uma mutação grave no outro alelo de *CFTR*. Ocasionalmente, as variantes 5T e p.Arg117His estão em alelos diferentes (*trans*) do gene *CFTR*. Nesse caso, 5T pode agir como uma mutação causadora de FC na presença de um outro polimorfismo, o trato TG (1210-34TG[9_13]). Este polimorfismo consiste em 9 a 13 repetições -TG- localizadas à montante (“upstream”) de 5T no íntron 9 de *CFTR*. O trato TG ocorre tipicamente em três formas: TG11, TG12 e TG13. Dependendo de qual forma está presente na mesma cópia de *CFTR* que 5T, diferentes desfechos podem acontecer. Quando há ocorrência de alelo TG12-T5 ou TG13-T5, essas combinações podem agir como mutação causadora de doença (contanto que haja uma mutação grave no outro alelo de *CFTR*), resultando em níveis elevados de cloreto no suor e sintomas clínicos de FC. Esses sintomas podem variar ou ser mais brandos que aqueles provocados por mutações graves, como p.Phe508del. Além

disso, há um risco aumentado para infertilidade masculina (CFTR2 2011; Brennan & Schrijver 2016).

Então, quando o alelo 5T ocorre no mesmo alelo que a mutação p.Arg117His ou 12-13 repetições -TG-, ele pode contribuir para o fenótipo clínico associado à FC e a transtornos relacionados ao *CFTR*, como a agenesia congênita de vasos deferentes (CFTR2 2011; Brennan & Schrijver 2016). Assim, a detecção exata do genótipo TG-poliT é criticamente importante como teste reflexo quando p.Arg117His é identificada em um painel de triagem de heterozigotos.

1.4. Diagnóstico de FC

Estudos envolvendo o gene *CFTR* representam uma das análises genéticas mais frequentes realizadas mundialmente, sejam para confirmar o diagnóstico clínico de FC e de doenças relacionadas ao *CFTR*, para a testagem de indivíduos heterozigotos, ou para o diagnóstico genético pré-natal e pré-implantacional (do inglês PGD – "Preimplantation Genetic Diagnosis") (Ong et al. 2001; Lucarelli et al. 2012; Bergougnoux et al. 2015; Brennan & Schrijver 2016). Por compreender sinais e sintomas heterogêneos, a confirmação diagnóstica da FC e das doenças relacionadas ao *CFTR* pode ser difícil, sobretudo na infância. Quanto mais cedo ocorre o diagnóstico, maior é a chance de que os pacientes tenham melhor qualidade de vida, pois se previne que indivíduos com insuficiência pancreática tenham problemas nutricionais, de crescimento e de diabetes associada à FC, bem como se possibilita um início precoce do tratamento preventivo de infecções pulmonares (Ong et al. 2001; Lucarelli et al. 2012; Bergougnoux et al. 2015; Brennan & Schrijver 2016).

O diagnóstico da FC é estabelecido em indivíduos que apresentam uma ou mais características fenotípicas de FC e evidência de anormalidade na função do CFTR. Tal diagnóstico se baseia em valores anormais de concentração de cloreto no suor (≥60 mmol/L) em teste de iontoforese com pilocarpina, ou na presença de duas mutações causadoras de doença no *CFTR* ou ainda na medição (característica de FC) em teste de diferença de potencial transepitelial nasal. Além disso, o diagnóstico da FC pode ser feito na ausência de características fenotípicas quando se trata de diagnóstico pré-natal (presença de duas variantes patogênicas dispostas em *trans* no *CFTR*) ou de programa de

triagem neonatal (valores de cloreto no suor ≥60 mmol/L e presença de duas mutações causadoras de FC, dispostas em *trans*) (Ong et al. 2001; Lucarelli et al. 2012; Bergougnoux et al. 2015; Brennan & Schrijver 2016; Farrell et al. 2017).

No Brasil, a inclusão da FC na triagem neonatal tem sido crucial para a identificação de casos antes do desenvolvimento de sintomas clínicos. O método usado para essa triagem é o teste do tripsinogênio imunorreativo (IRT – "immunoreactive trypsinogen"), realizado preferencialmente entre o terceiro e o quinto dia pós-natal, e por volta de três a cinco semanas de vida. De acordo com novas diretrizes da "Cystic Fibrosis Foundation" para o diagnóstico de FC, a triagem de casos com IRT positivo deve ser seguida por teste de cloreto no suor (**Figura 5**). Em caso de valores limítrofes (30-59 mmol/L), é realizado genotipagem do gene *CFTR* (Farrell et al. 2017). Segundo o relatório anual de 2014 do Registro Brasileiro de Fibrose Cística, 1.032 dos 3.511 pacientes registrados tiveram a triagem neonatal como uma das condições para o diagnóstico de FC (GBEFC 2014). Em parte dos recém-nascidos com IRT positivo, todavia, identifica-se uma ou nenhuma mutação (sendo estas ligadas a valores limítrofes no teste de cloreto no suor). Além dos casos falso-positivos, uma outra possível explicação para essa divergência entre diagnóstico laboratorial e molecular é o fato de que o painel recomendado pela ACMG ("American College of Medical Genetics and Genomics") contém apenas as 23 mutações mais frequentes na população mundial. Por mais que esses painéis sejam adaptados para diferentes populações e etnias, ainda assim há mutações que não são detectadas por estes painéis (Ong et al. 2001; Lucarelli et al. 2012; Bergougnoux et al. 2015; Brennan & Schrijver 2016).

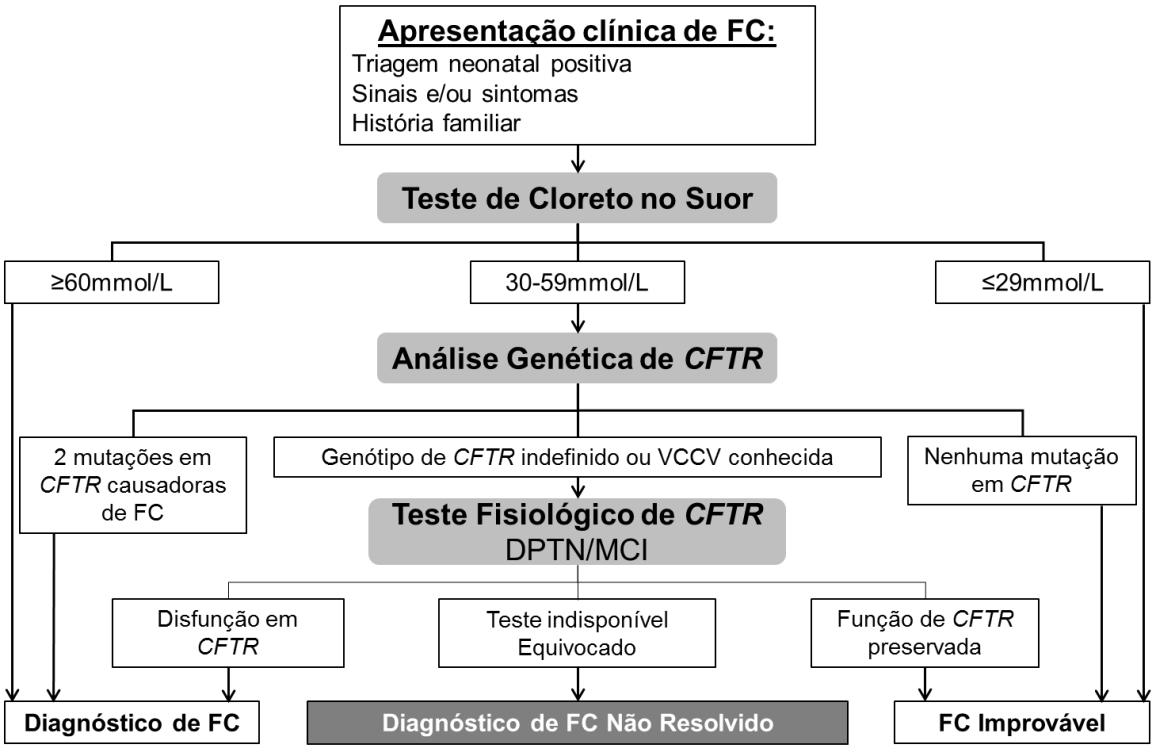


Figura 5. Algoritmo simplificado para o diagnóstico de FC. A FC é diagnosticada quando um indivíduo tem apresentação clínica da doença e evidência de disfunção no gene *CFTR*. Os testes de função do *CFTR* não são necessariamente feitos nessa ordem, mas hierarquicamente para estabelecer o diagnóstico de FC, o teste de cloreto no suor deveria ser considerado primeiro, depois a análise genética do *CFTR*, e então, os testes fisiológicos da proteína CFTR. Todos os indivíduos identificados com FC deveriam ter o teste de cloreto no suor e a análise genética de *CFTR* realizados. Raros indivíduos com cloreto no suor <30 mmol/L podem ser considerados fibrocísticos se alternativas foram excluídas e os outros testes confirmatórios (testes genéticos e fisiológicos) corroborarem o diagnóstico de FC. Se apenas uma variante de *CFTR* é identificada em análise limitada, mais testes no *CFTR* (testes mais extensivos) devem ser realizados. A FC é possível se ambos alelos possuem mutações causadoras de FC, indefinidas, ou variantes de consequência clínica variável; FC é improvável se somente mutações não causadoras de FC forem encontradas. Se o diagnóstico de FC não for resolvido, "Síndrome metabólica relacionada ao *CFTR*"/"Triagem para FC positiva, diagnóstico inconclusivo" (após a triagem neonatal) ou doença relacionada ao *CFTR* deve ser considerada. Raramente, a ausência de diagnóstico pode ser apropriada, mas é preciso seguimento desses pacientes. Nesses casos, o uso de "Portador de FC" ou do problema clínico específico deve ser usado como propósito de caracterização. DPTN: Diferença de potencial transepitelial nasal; MCI: Mensuração da corrente intestinal; VCCV: Variante de consequência clínica variável. Adaptado de Farrell et al. 2017.

Além da análise desses painéis de mutações frequentes, o diagnóstico molecular da FC ainda conta com a análise de grandes deleções no gene *CFTR*, por meio do ensaio de MLPA ("Multiplex Ligation-dependent Probe Amplification"), e com uma análise de sequência mais robusta, a qual pode envolver todos os éxons, junções éxon/íntron, regiões promotoras e regiões intrônicas específicas. Essa forma de sequenciamento pode detectar mais de 98% das mutações no *CFTR*. Em alguns casos, pode-se ainda sequenciar o gene completo. Entretanto, gera-se informação excessiva, que por vezes não se consegue interpretar, como variantes de significado clínico variável ou

incerto ainda não descritas e nem classificadas pela literatura (Ong et al. 2001; Lucarelli et al. 2012; Bergougnoux et al. 2015; Brennan & Schrijver 2016). Com isso, realizam-se outros testes que avaliam fisiologicamente a função da proteína CFTR, como a já citada medição da diferença de potencial transepitelial nasal. Todavia, ainda assim existem casos com IRT positivos que permanecem não resolvidos e por isso precisam ser acompanhados, além de casos em que se chega ao diagnóstico de síndrome metabólica relacionada ao CFTR (Farrell et al. 2017).

No cenário latino-americano, o acesso a ferramentas atualizadas de diagnóstico pode variar muito entre países e dentro do mesmo país. Por exemplo, o teste de cloreto no suor não está amplamente disponível, e em vários locais o teste de diferença de potencial transepitelial nasal (DPTN) é usado em seu lugar (Silva Filho et al. 2016).

Embora a análise molecular do *CFTR* seja reconhecidamente cada vez mais útil, ela não é absoluta e precisa ser usada como uma ferramenta adicional para o diagnóstico da FC, dada a baixa correlação genótipo-fenótipo observada na doença. Ademais, nem todos os centros de referência têm profissionais e instalações necessárias para gerar e interpretar dados de sequenciamento de forma robusta e precisa. Além disso, a heterogeneidade alélica da população latino-americana culmina em uma considerável variedade de variantes que ainda não estão descritas em bancos de anotação de variantes (Pérez et al. 2007; Dal'Maso et al. 2013; Silva Filho et al. 2016).

1.4.1. Interpretação de Variantes Genéticas

Em 2015, a ACMG e a AMP ("Association for Molecular Pathology") publicaram em conjunto diretrizes para a interpretação de variantes genéticas de doenças monogênicas em laboratório clínico (Richards et al. 2015). Em tal publicação, há uma série de evidências a serem analisadas a fim de se estabelecer o grau de patogenicidade de cada variante. Dentre essas evidências, encontram-se os programas de predição computacional (*in silico*), que podem auxiliar na interpretação das variantes. Embora tais preditores possam usar diferentes algoritmos, eles podem ajudar a determinar o efeito da variante a nível de nucleotídeo e aminoácido, transcrito primário e transcritos alternativos, assim como o potencial impacto da variante sobre a proteína. Existem duas categorias principais de preditores: os que avaliam se uma mutação de troca de aminoácido (**Tabela 1**) é

prejudicial para estrutura e função da proteína resultante, e os que avaliam se existe efeito da mutação sobre o "splicing" (Richards et al. 2015). Esses preditores possibilitam o direcionamento de ensaios *in vitro* para a caracterização funcional das mutações (Cutting 2015; Richards et al. 2015).

Nesses preditores, o impacto de uma variante "missense" depende de critérios como a conservação evolutiva de um aminoácido ou nucleotídeo, a localização e o contexto da cadeia polipeptídica, e a consequência bioquímica da substituição de um aminoácido. A utilização de um ou a combinação desses critérios é usada por vários algoritmos para predizer o impacto da mutação de sentido trocado. Alguns estudos (Tavtigian et al. 2008; Hicks et al. 2011; Thusberg et al. 2011; Thompson et al. 2013) tentam comparar a performance dos preditores entre si e avaliar sua habilidade de predizer variantes que sabidamente causam doença. De modo geral, a acurácia dos preditores varia entre 65-80% quando analisam mutações causadoras de doença. Além disso, muitos preditores tendem a ter uma baixa especificidade, resultando em uma superpredição de mutações "missense" como sendo deletérias, e em predições questionáveis quando se trata de variantes de sentido trocado com efeito mais brando (Richards et al. 2015).

Em 2014, Bendl e colaboradores publicaram artigo que contém a metodologia utilizada para a criação de um preditor consenso para a avaliação de variantes "missense" (Bendl et al. 2014). Tal preditor, o PredictSNP, é o resultado da análise de oito ferramentas de predição bem estabelecidas: MAPP, nsSNPAnalyzer, PANTHER, PhD-SNP, PolyPhen-1, PolyPhen-2, SIFT e SNAP. Destas, as seis que obtiveram melhor desempenho (MAPP, PhD-SNP, PolyPhen-1, PolyPhen-2, SIFT e SNAP) na análise realizada pelos autores foram combinadas para desenvolver o PredictSNP, o qual está disponível online no seguinte endereço eletrônico: http://loschmidt.chemi.muni.cz/predictsnp, dos Laboratórios Loschmidt, na República Tcheca. Por ser um preditor consensual, o PredictSNP compila dados de outros preditores para criar sua própria predição, tal qual é descrito na **Tabela 1**. Como os preditores têm diferentes formas de exibir seus resultados, o PredictSNP faz uma uniformização deles. Assim, todos os resultados podem ser comparados entre si, o que facilita o entendimento do usuário. O PredictSNP representa cada predição como neutral ou deletéria associando a ela um percentual de acurácia esperada.

Tabela 1. Algoritmos de preditores *in silico* para variantes "missense"

Nome	Base do algoritmo	Observação
ConSurf	Conservação evolutiva	
FATHMM	Conservação evolutiva	
PANTHER	Conservação evolutiva	Alinhamentos de múltiplas sequências da biblioteca do PANTHER têm limitações
PhD-SNP	Conservação evolutiva	
SIFT	Conservação evolutiva	
SNPs&GO	Estrutura/função da proteína	
SNAP	Estrutura/função da proteína e conservação evolutiva	
Align GVGD	Estrutura/função da proteína e conservação evolutiva	
MAPP	Estrutura/função da proteína e conservação evolutiva	Existência de lacunas nos alinhamentos de múltiplas sequências
PolyPhen	Estrutura/função da proteína e conservação evolutiva	Regras empíricas para tomada de decisão
PolyPhen-2	Estrutura/função da proteína e conservação evolutiva	Método de aprendizagem de máquina
PROVEAN	Alinhamento e medida da similaridade entre a homologia da variante e da sequência da proteína	
nsSNPAnalyzer	Alinhamento múltiplo de sequência e na análise da estrutura da proteína	Requer proteína homóloga na base de dados ASTRAL
Condel	Combina SIFT, PolyPhen-2, e MutationAssessor	
CADD	Contrasta anotações derivadas de alelos fixados/quase fixados em humanos com variantes simuladas	
PredictSNP	Analisa ao mesmo tempo **MAPP, PhD-SNP, PolyPhen-1, PolyPhen-2, SIFT, SNAP**, nsSNPAnalyzer e PANTHER. Combina os destacados em negrito para criar seu próprio escore.	

Fonte: Adaptado de Richards et al. 2015.

Mais recentemente, em 2016, o mesmo grupo lançou o PredictSNP2 (Bendl et al. 2016). Essa nova versão compara preditores que avaliam variantes em regiões regulatórias, codificantes ("missense", sinônimas e "nonsense") e sítios de "splicing". O PredictSNP2 também leva em conta a avaliação realizada pelo PredictSNP quando se trata de variantes "missense". Na realidade, o mesmo estudo verificou que o PredicSNP apresenta acurácia superior aos outros preditores avaliados quando se trata de variantes "missense". Pelo fato de este estudo avaliar apenas variantes de sentido trocado, optou-se então por empregar o PredictSNP.

1.4.1.1. Modelagem molecular de CFTR

A identificação de resíduos importantes envolvidos na função e na estabilidade de interações peptídicas geralmente baseia-se em ensaios experimentais como mutagênese direcional. Todavia, esses experimentos são caros e demorados. De maneira geral, o conhecimento da estrutura tridimensional pode melhorar a entendimento e possibilitar a exploração dessa estrutura, focando no papel de resíduos de aminoácidos em posições que provavelmente afetariam a atividade da proteína (Thévenet et al. 2015).

Assim, num contexto em que o sequenciamento de nova geração gera dados de sequência de nucleotídeos e, por conseguinte, de aminoácidos, o estudo de estruturas tridimensionais envolvendo propriedades do canal de cloreto CFTR podem facilitar o entendimento do efeito de variantes no gene *CFTR* nessa proteína. Além disso, esses estudos podem inclusive gerar conhecimento útil para o desenvolvimento de ativadores, potenciadores, e corretores específicos para mutações que afetam a quantidade e/ou a atividade desse canal de membrana (Callebaut et al. 2017). De fato, algumas revisões da literatura exploram estudos que envolvem propriedades da estrutura tridimensional da proteína CFTR (Odolczyk & Zielenkiewicz 2014; Callebaut et al. 2017).

Uma das estruturas do CFTR disponíveis no “Protein Data Bank” (RCSB PDB) sob o identificador PDB ID: 5UAK revela a estrutura do CFTR humano defosforilado. Tal estrutura, desenvolvida por Liu e colaboradores (2017), revela algumas propriedades interessantes sobre o CFTR humano (Liu et al. 2017). Existe uma grande semelhança entre essa estrutura e o CFTR de *Danio rerio* (paulistinha, “zebrafish”) sob condições idênticas. Também, a estrutura de CFTR humano revela uma hélice previamente não resolvida pertencente ao domínio regulatório que impede a abertura do canal. Assim, essas informações podem auxiliar no entendimento de como as variantes presentes no gene *CFTR* podem afetar a atividade da proteína de mesmo nome, quais variantes têm maior potencial de causar dano e mais do que isso, fornecer informação útil para bancos de anotação de variantes genéticas.

1.4.1.2. Bancos de anotação de variantes genéticas

A informação gerada a respeito de variantes genéticas precisa ser compilada e interpretada de acordo com dados clínicos. Dessa forma, torna-se fundamental a existência de bancos de dados, sobretudo públicos, que disponibilizem informação a respeito de

variantes genéticas. Com relação ao gene *CFTR*, existem dois bancos de dados notoriamente mais completos e atualizados dentre os conhecidos pelos pesquisadores e clínicos que trabalham com Fibrose Cística e doenças relacionada ao CFTR. São eles: o "Cystic Fibrosis Mutation Database" (CFTR1 1989) e o CFTR2 ("The Clinical and Functional Translation of CFTR" - http://www.cftr2.org) (CFTR2 2011). Enquanto o primeiro compila informações moleculares a respeito das mutações no gene *CFTR*, o segundo agrupa dados populacionais, clínicos e funcionais a dados moleculares para interpretar a patogenicidade das variantes nesse gene. O CFTR2 contém atualmente apenas as 322 variantes mais frequentes na população de pacientes com Fibrose Cística.

Outro banco de dados de anotação da patogenicidade de variantes que tem grande importância é o ClinVar (http://www.ncbi.nlm.nih.gov/clinvar), do NCBI ("National Center for Biotechnology Information"). Esse banco de dados aceita submissões de diferentes centros e laboratórios. No caso do *CFTR*, o ClinVar tem o CFTR2 como uma de suas fontes de submissão da patogenicidade de variantes.

Portanto, este estudo pretende analisar as mutações "missense" no gene *CFTR* em diferentes preditores contidos no PredictSNP, comparando os resultados a dados clínicos e da literatura disponíveis no ClinVar e no CFTR2. Com isso, busca-se o estabelecimento da confiabilidade de cada um dos preditores em relação às mutações de sentido trocado, e assim poder-se-ia estabelecer um critério para predizer se variantes não patogênicas, de significado incerto ou de significado clínico variável precisam ou não ter sua anotação trocada, atualizando-as. Dessa forma, diminuir-se-ia a superpredição das variantes como sendo patogênicas e seria possível melhorar a acurácia dos preditores, contribuindo para a melhora nas decisões clínicas e no aconselhamento genético.

2. JUSTIFICATIVA

A fibrose cística é uma doença monogênica de alta prevalência, e as novas ferramentas de sequenciamento genético têm contribuído para o seu diagnóstico. Entretanto, identificam-se inúmeras variantes para as quais não se consegue estabelecer o significado clínico; os dados controversos da literatura se somam a preditores computacionais que podem divergir e não contribuir efetivamente para a interpretação das variantes genéticas. Essas disparidades prejudicam a interpretação correta das variantes, que têm grande impacto tanto no aconselhamento genético quanto nas decisões terapêuticas.

Assim, torna-se relevante identificar ferramentas mais sensíveis e específicas para melhor definição clínica dessas variantes.

3. OBJETIVOS

3.1. Objetivo Geral

Analisar variantes de sentido trocado no gene *CFTR* em diferentes preditores computacionais.

3.2. Objetivos Específicos

- Analisar concordância entre diferentes preditores computacionais para mutações "missense" no gene *CFTR*.
- Verificar concordância entre cada preditor com bancos de dados de anotação de variantes para o gene *CFTR* (CFTR2, ClinVar).
- Determinar sensibilidade e especificidade de cada preditor ao compará-lo com os bancos de dados de anotação de variantes no *CFTR*.

4. ARTIGO CIENTÍFICO

O artigo intitulado “Determining the Pathogenicity of *CFTR* Missense Variants: Multiple Comparison of *in silico* Predictors and Variant Annotation Databases” foi formatado conforme diretrizes e submetido para publicação junto ao periódico “Human Mutation”.

Determining the Pathogenicity of *CFTR* Missense Variants: Multiple Comparison of *in silico* Predictors and Variant Annotation Databases

Marcus Michels[1], Ursula Matte[1,2,3], Lucas Rosa Fraga[1], Aline Castello Branco Mancuso[3], Rodrigo Ligabue-Braun[4], Elias Figueroa Rodrigues Berneira[5], Marina Siebert[2], Maria Teresa Vieira Sanseverino[1,6,7,*]

[1] Programa de Pós-Graduação em Genética e Biologia Molecular, Departamento de Genética, Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, RS, Brazil

[2] Centro de Pesquisa Experimental, Hospital de Clínicas de Porto Alegre (HCPA), RS, Brazil

[3] Grupo de Pesquisa e Pós-Graduação (GPPG), Hospital de Clínicas de Porto Alegre, RS, Brazil

[4] Programa de Pós-Graduação em Biologia Celular e Molecular, Centro de Biotecnologia, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brazil

[5] Faculdade de Medicina (FAMED), Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brazil

[6] Serviço de Genética Médica, Hospital de Clínicas de Porto Alegre, Porto Alegre, RS, Brazil

[7] Escola de Medicina, Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS), Porto Alegre, RS, Brazil

*Corresponding author:

Dr. Maria Teresa Vieira Sanseverino, MD

Serviço de Genética Médica

Hospital de Clínicas de Porto Alegre

2350 Ramiro Barcelos Street, Porto Alegre - RS, Brazil, Postcode: 90035-903

Email: msanseverino@hcpa.edu.br

Abstract

Cystic Fibrosis (CF) is a monogenic autosomal recessive disease caused by mutations in the Cystic Fibrosis Transmembrane Conductance Regulator gene (*CFTR*). CF diagnosis relies on clinical, functional and molecular studies. To date, over 2,000 variants have been described on *CFTR* (~40% missense). Since few of them have their pathogenicity confirmed, *in silico* analysis could help genetic diagnosis and counseling. Here, the pathogenicity of 779 *CFTR* missense variants was established by consensus predictor PredictSNP and compared to CFTR2 and ClinVar. Kappa agreement between all predictors was fair or moderate. The same applies to predictors and CFTR2/ClinVar. Analysis of expected accuracy showed that PredictSNP and PhDSNP had higher accuracies than other predictors. Isoleucine-to-valine substitutions were significantly associated with the neutral outcome while Leucine-to-Proline changes were significantly associated with deleteriousness. Sensitivity and specificity analysis showed that PhDSNP and PredictSNP are more suitable to analyze *CFTR* missense variants. We could verify that the great difficulty of prediction tools is excluding the non-pathogenicity of missense variants. Our results should contribute to clarifying decision making when predicting the pathogenicity of *CFTR* missense variants.

Keywords: Cystic fibrosis; Mutation; Prediction; Bioinformatics; Systems biology.

Introduction

Cystic Fibrosis (CF; MIM# 219700) is a monogenic autosomal recessive inherited disease that affects over seventy thousand people worldwide. Even though CF is present in all ethnicities, most patients affected by this life-limiting disease have European ancestry . One in 2,000-3,000 newborns is affected with CF in Europe. In North America, the incidence of CF is 1 in 3500 newborns . In Brazil, where the population is very heterogeneous, the estimated incidence of CF in Euro-Brazilians is 1 in 1,600 newborns whereas in Afro-Brazilians it is 1 in 14,000 .

The disease has a broad spectrum of signal and symptoms. In the lungs and airways of CF patients, the viscosity of the mucus is increased, which causes bacterial infections that are difficult to treat and eventually lead to death. Other organs affected by CF are the liver, the intestines, the pancreas, the sweat glands and, in males, the reproductive tract. In the gastrointestinal tract, especially, CF patients may develop meconium ileus (newborn period), pancreatic insufficiency and cirrhosis. CF also causes obstructive azoospermia due to the congenital absence of vas deferens

CF is caused by mutations in the Cystic Fibrosis Transmembrane Conductance Regulator gene (*CFTR*; *ABCC7*; MIM# 602421) . This gene encodes for a transmembrane channel that regulates the flow of anions in the apical domain of epithelial cells. When this channel is impaired, CFTR protein malfunction causes damage in the tissues and organs where *CFTR* expression is critical. CFTR protein is a member of the ATP-binding cassette (ABC) superfamily , and it has five domains: two membrane spanning domains (MSD1 and MSD2), two nucleotide-binding domains (NBD1 and NBD2), and one intrinsically disordered region, the regulatory domain (RD) . The amount and/or the function of CFTR in the cell membrane can be affected by disease-causing variants .

To date, over 2,000 variants have been described in *CFTR* according to the Cystic Fibrosis Mutation Database (CFTR1 - http://www.genet.sickkids.on.ca/app) . Although p.Phe508del (c.1521_1523delCTT) is the most common mutation in CF patients, present in about 70% of CF chromosomes worldwide, the ones that cause amino acid substitution correspond to almost 40% of *CFTR* variants . There are two missense mutations that are carried by more than 1% of CF patients, which are p.Gly551Asp (c.1652G>A) and p.Asn1303Lys (c.3909C>G) . Even though most missense variants are rare, several may have clinical significance. Unfortunately, the minority of them has conclusive clinical data about their pathogenicity.

In 2015, the American College of Medical Genetics and Genomics (ACMG) and the Association for Molecular Pathology (AMP) published guidelines for the clinical laboratory interpretation of genetic variants regarding monogenic and mitochondrial diseases . The journal cites a plethora of evidence that should be taken into consideration when establishing the pathogenicity of a genetic variant. Amid them, computational (*in silico*) predictive programs can have an auxiliary role on variant interpretation . Among the main categories of *in silico* predictors is those which evaluate missense variants. The impact of these variants depends on criteria such as the functional consequence of an amino acid substitution, the location and the context within the protein structure and/or the evolutionary conservation of a nucleotide or amino acid. The algorithms used by those predictors consider one or more of the criteria above when assessing the impact of a missense variant .

Several studies have been published throughout the years in order to compare the performance of *in silico* predictors and to evaluate their ability to correctly predict disease-causing variants for different genes . Generally speaking, the accuracy of the predictors ranges from 65 to 80% when analyzing pathogenic variants. Furthermore, most predictors tend to have low specificity, which results in an overrepresentation of these missense variants

as deleterious. Also, these predictions may not be reliable when analyzing missense variants with mild effect . As an example of the applicability of these predictors in the clinic, in 2014 Manickam and colleagues identified different potential deleterious SNPs in the *GBA* gene that could be associated with Gaucher's disease . Specifically for CF, Dorfman and collaborators evaluated three popular predictors by comparing the predicted pathogenicity against the diagnosis of CF and its clinical manifestations in cohorts of subjects with CF-disease and CFTR-related disorders carrying these mutations . More recently, Masica and Karchin discussed the prediction of endophenotypes in the CF spectrum in the context of different *CFTR* missense variants .

Bearing in mind the different grades of manifestation and severity of CF, the aim of this study was to predict the effect of *CFTR* missense variants and to compare the results to public clinical data available in variant annotation databases - CFTR2 (http://www.cftr2.org) and ClinVar (http://www.ncbi.nlm.nih.gov/clinvar) – in order to establish the agreement between predictors and to compare the accuracy between predictors and variant annotation databases, determining which predictor is more suitable for analyzing *CFTR* missense variants. In addition, we aimed to find out if there are any particular modifications in the CFTR protein that could make predictors agree or disagree more. Finally, ROC curves were built in order to establish which predictors are more reliable (and which ones are not) when analyzing *CFTR* missense variants available on CFTR2 and ClinVar.

Materials and Methods

Data collection

A summary of this study's workflow is represented in Figure 1. Until the date this analysis started (October 2016), there were 2,009 variants described at the Cystic Fibrosis Mutation Database . In this study we evaluated only variants that cause amino acid substitution; thus

779 missense variants in the *CFTR* gene (NM_000492.3, LRG_663, ENSG00000001626) were collected from the Human Gene Mutation Database® (HGMD) Professional 2016.2 Trial Version for posterior analysis of pathogenicity. This study was approved by the Hospital de Clínicas de Porto Alegre (HCPA) Ethics Committee (CAAE 59458516.5.0000.5327; GPPG 16-0644).

Prediction

In order to predict the effect of *CFTR* missense variants in the protein, we employed the consensus classifier PredictSNP (http://loschmidt.chemi.muni.cz/predictsnp1/) . The canonical protein sequence for the analysis was retrieved from the UniProt database (UniProtKB, Isoform 1 - http://www.uniprot.org/uniprot/P13569.fasta) . PredictSNP comprises scores from different predictors (MAPP, PhDSNP, PolyPhen-1, PolyPhen-2, SIFT, SNAP, nsSNPAnalyzer, and PANTHER) and also uses the information of six of them (MAPP, PhDSNP, PolyPhen-1, PolyPhen-2, SIFT, SNAP) to create its own score. PredictSNP then classifies variants as “neutral” or “deleterious” and transforms the individual confidence scores of each predictor into one comparable scale ranging from 0–100%, which represents the percentage of expected accuracy, as described elsewhere . By doing it, this tool homogenizes the analysis. Variants were inputted from codons 1 to 1480 using their legacy names, e.g. S1251N (c.3752G>A; p.Ser1251Asn), and then submitted to analysis of pathogenicity.

Variant Annotation Databases

In order to compare the predicted pathogenicity of missense variants to what has been described in the literature, we used data from CFTR2 and ClinVar as a reference for determining if the predictors asserted the pathogenicity correctly.

The Clinical and Functional Translation of *CFTR* (CFTR2) is an online database for health professionals and patients, gathering clinical, molecular, and functional information. Also, it publishes every year a list of variants already detected in patients across the globe. CFTR2 classifies variants as "CF-causing", "Non CF-causing", "Varying clinical consequence" and "Unknown significance". Sometimes one variant may change from one class to another. For this study, we used the most up-to-date list available on CFTR2 at the time (https://www.cftr2.org/mutations_history; CFTR2_8August2016.xlsx) as a reference to compare the predicted and the clinical information of pathogenicity. Only the minority of variants that we analyzed on PredictSNP are recorded in the CFTR2 list (74 variants).

The other variant annotation database, ClinVar, is a platform of the National Center for Biotechnology Information (NCBI) that aggregates information about genomic variation and its relationship to human health. ClinVar follows the classification proposed by Richards and cols. (2015) to interpret variants, classifying them as "Pathogenic", "Likely pathogenic", "Unknown significance", "Likely benign" and "Benign". Only 146 of the 779 *CFTR* missense variants analyzed in this study have been described on ClinVar so far (until November,20th 2016).

CFTR topology, domains, and secondary structure

Information about CFTR structure was gathered from different tools. Data from CFTR topology were retrieved from the UniProt database (UniProtKB, http://www.uniprot.org/uniprot/P13569) , and divided into "cytoplasmic", "transmembrane" and "extracellular", according to the amino acid position. The information about CFTR domains (MSD1, NBD1, RD, MSD2 and NBD2) was collected both from Pfam (http://pfam.xfam.org/protein/P13569) and CFTR1. When data diverged between them, CFTR1 data were chosen since it is a specific database for the *CFTR* gene. Regarding the

secondary structure of CFTR, information was collected from the Protein Data Bank (RCSB PDB - http://www.rcsb.org) , using the PDB ID: 5UAK . Features represented in the secondary structure of CFTR were divided according to RCSB PDB into: “β-strand”, “turn”, “empty (no secondary structure assigned)”, “3/10-helix”, “β-bridge”, “bend” and “α-helix”.

Modeling of CFTR variants

The structural modeling of the CFTR protein (UniProtKB number: P13569) was performed using the I-TASSER package . Through sequential steps of identification of possible template structures, template fragmentation, incremental model construction and evaluation, the tool was able to construct a high-quality model for the protein (residues 1-1480). The visualization of the structures was performed with the software PyMOL (The PyMOL Molecular Graphics System, Version 1.8 Schrödinger, LLC). This model was created to verify the possible implications of four different *CFTR* missense variants that were chosen based on the agreement shown by all predictors for each one of them. The variants picked for the model are: p.Met1Val (c.1A>G); p.Arg117His (c.350G>A); p.Gly551Asp (c.1652G>A); p.Ile1027Thr (c.3080T>C).

Statistical analysis

In order to determine the agreement between individual predictors, we used Kappa coefficient . We also analyzed the agreement between predictors and variant annotation databases (CFTR2 and ClinVar). For this analysis, “Non CF-causing”, “Benign” and “Likely benign” variants were considered “neutral” while “CF-causing”, “Pathogenic” and “Likely pathogenic” variants were considered “deleterious”. Since PredictSNP not only determines if a variant is “neutral” or “deleterious” but also provides a percentage of expected accuracy for the predicted pathogenicity, we compared the accuracies between predictors and we also filtered the analysis for CFTR2 and ClinVar outcomes. To do so, variants reported in the

annotation databases as “CF-causing”, “Pathogenic” or “Likely pathogenic” were considered as having 100% accuracy. By doing it, we could compare the predicted accuracies to a standard. For this analysis, Friedman’s non-parametric test was applied.

We also wanted to analyze if there were specific amino acid changes that could make predictors assert a right or a wrong result in comparison to CFTR2 and ClinVar. For this analysis, only amino acid substitutions with absolute frequency of 10 or higher among the 779 CFTR missense variants were included. Pearson’s Chi-Squared or Fisher’s Exact Test was used when appropriate. In order to verify if there was any amino acid change, any particular region, domain, or any secondary feature of the CFTR protein that was associated with a higher or lower agreement between predictors, Chi-Squared or Fisher’s Exact Test was used when appropriate. We counted the agreement between predictors based on the predicted pathogenicity (“neutral” or “deleterious”) for a given variant, as follows: “0” (full agreement), “1” (1 disagreement), “2” (2 disagreements), and “3” (3 disagreements). Predictors that could not assign the pathogenicity of any of the 779 CFTR missense variant (considering them as “missing”) were excluded from the analysis.

Lastly, sensitivity and specificity analysis was performed and receiver operating characteristic (ROC) curves were constructed for pathogenic and non-pathogenic variants on CFTR2 and ClinVar in order to determine which of the used *in silico* tools predicts variant annotation best. For this analysis, accuracy values were used as the continuous variable. Variants predicted as neutral were analyzed as negative accuracies, differentiating them from those accuracies of variants predicted as deleterious. Moreover, Youden Index J was employed in order to measure the performance of diagnostic tests, since it takes sensitivity and specificity into account in its analysis . Results were analyzed using SPSS v18.0 and MedCalc Statistical Software version 17.4. Data were considered statistically significant

when $p<0.05$.

Results

Descriptive Analysis

Descriptive data of the predicted variants are shown in Table 1. All predictors had 779 variants validated except for MAPP (missing = 53) and PANTHER (missing=488). At least 25% of the predictions made by PredictSNP had an accuracy of 87%. Also, nsSNPAnalyzer had the lowest amplitude and its predictions had low accuracy in comparison to other predictors. At least half of the accuracies provided by nsSNPAnalyzer were 63%.

Agreement between predictors and variant annotation databases

Analysis of agreement is depicted in Table 2. PredictSNP had a moderate agreement with CFTR2 ($\kappa = 0.518$; $p<0.001$) but only fair agreement with ClinVar ($\kappa = 0.302$; $p<0.001$). PredictSNP and the other predictors also achieved fair to moderate agreement between each other. Since PredictSNP takes into account the scores from MAPP, PhDSNP, PolyPhen1, PolyPhen2, SIFT, and SNAP to provide its own score, the agreement between PredictSNP and these other predictors is dependent, which contradicts the premise that judges must be independent when calculating Kappa agreement. Likewise, CFTR2 is one of the submitters of variants to ClinVar, which means that the agreement between them is not independent as well. Hence, those kappa values were not included in our results. PANTHER did not have significant agreement to CFTR2 nor ClinVar due to small sample number. Therefore, PANTHER was excluded from further analyses. Overall, these results do not lead to a clear conclusion. Interestingly, PhDSNP had moderate agreement only with PolyPhen2, which is one of the predictors with higher sensitivity and specificity (shown in the following sections). Also, nsSNPAnalyzer showed fair agreement with some of the other predictors, even though

it has been the predictor with the poorest performance in further analyses.

Comparison of accuracy between predictors and variant annotation databases

In the comparison between accuracies, all differences described and depicted (Supp. Table S1) are statistically significant ($p<0.05$). When analyzing just predictors without any comparison to variant annotation databases, it was observed that PredictSNP, PhDSNP, and SIFT had greater median of accuracy than other predictors (Figure 2A). PredictSNP and PhDSNP did not differ from each other. In this analysis, SIFT had the highest median of accuracy, which was also significantly higher than PredictSNP and PhDSNP. In contrast, PolyPhen1, PolyPhen2, and nsSNPAnalyzer had lower median of accuracy than the others, with nsSNPAnalyzer as the predictor with the lowest median of accuracy (Figure 2A).

The comparison of accuracy between predictors for variants annotated on CFTR2 was also performed. For this analysis, when we tried to compare the accuracy of predictors for "non CF-causing" variants, Friedman's test did not show any significant differences, probably because the sample number was insufficient (n=8) to reveal any significance (data not shown). Nevertheless, when analyzing the accuracy of "CF-causing" variants, the same non-parametric test showed significant differences, which are also depicted in Supp. Table S1. Here, PredictSNP and PhDSNP were the predictors with higher median of accuracy, not showing significant difference between each other (Figure 2B). PredictSNP had also the highest median of accuracy. On the other hand, PolyPhen1 and nsSNPAnalyzer were the predictors with lower median of accuracy. Again, nsSNPAnalyzer had the lowest median of accuracy (Figure 2B).

Lastly, we compared the accuracy between predictors for variants already annotated on ClinVar. When we tried to compare the accuracy of predictors for "Benign" and "Likely benign" variants, once more Friedman's test did not show significant differences, probably

because the sample number was small (n=12) and no significance was found (data not shown). However, when we analyzed the accuracy of predictors for "Pathogenic" and "Likely pathogenic" variants, the same test has shown significant differences (Supp. Table S1). Like before, PredictSNP and PhDSNP showed higher median of accuracy. PhDSNP had the highest median of accuracy in this analysis (Figure 2C). On the other side, once again PolyPhen1 and nsSNPAnalyzer had lower median of accuracy, with nsSNPAnalyzer as the one with the lowest median of accuracy (Figure 2C). Altogether, this set of results indicates that PhDSNP and PredictSNP had the best performance when analyzing *CFTR* missense variants. On the contrary, PolyPhen1 and nsSNPAnalyzer seem to perform poorly when analyzing the same variants.

Amino acid substitutions

In order to know if predictors could associate any amino acid substitution with a specific outcome (neutral or deleterious) regardless of amino acid position, 272 amino acid changes with absolute frequency of 10 or higher were analyzed, and the ones significantly associated with predicted outcomes are shown in Table 3 and Supp. Table S2. Noteworthy, isoleucine-to-valine (Ile-Val) changes were significantly associated with a neutral outcome ($p<0.05$) by all predictors with the exception of MAPP (adjusted residue not significant; $p<0.001$) and nsSNPAnalyzer ($p=0.120$). Another substitution that is worthy of attention is the leucine-to-proline (Leu-Pro) change. Most Leu-Pro changes were significantly associated with a deleterious outcome ($p<0.05$) by all predictors (24/25 out of 26 Leu-Pro variants analyzed). Again, nsSNPAnalyzer was an exception ($p=0.120$). Overall, PolyPhen1 was the predictor that had the highest number of amino acid substitutions significantly associated with the outcomes. Since associations made by nsSNPAnalyzer were not statistically significant ($p=0.120$), this predictor was not represented in Table 3.

CFTR Structure Analysis

When we analyzed if any data of topology, domain, or feature in the secondary structure could be associated with a variant already annotated on CFTR2 or ClinVar, statistical analysis has not shown any significance. We believe this happened because the number of variants of CFTR2 (n=57) and ClinVar (n=109) is not enough for this type of comparison (data not shown).

Agreement between predictors according to CFTR structure and amino acid substitutions

We wanted to evaluate the agreement between predictors in relation to CFTR's topology (Figure 3A), domains (Figure 3B), features of secondary structure (Figure 3C), and amino acid substitutions. MAPP and PANTHER were excluded from this analysis because they were not capable of assigning the pathogenicity of a considerable number of *CFTR* missense variants. The results showed that predictors have the tendency to fully agree when amino acid changes are located in the cytoplasm, although it was not statistically significant (Figure 3A; $p=0.052$). To what concerns protein domains, CFTR amino acid changes located in NBD1 and NBD2 are significantly associated with full agreement between predictors ($p<0.001$). For those changes located in MSD1, on the contrary, predictors tend to not fully agree, being directly associated with two disagreements between predictors (Figure 3B; $p<0.001$). Taking the features of CFTR secondary structure into account, amino acid substitutions located in β-strands and bends are associated with full agreement between predictors whereas amino acid changes in α-helices are associated with at least one disagreement between predictors (Figure 3C; $p=0.001$). In the context of amino acid changes, the analysis did not show any significance when comparing these substitutions to the agreement between predictors ($p=0.379$).

Variants in the CFTR model

Variants that would be featured in the CFTR model (Figure 4) were chosen based on the agreement between predictors, as shown in the next paragraphs.

Predictors fully agreed that p.Met1Val (c.1A>G; legacy name M1V) is a neutral variant. However, our model (Figure 4A) shows that this variant would deviate the translation initiation to the second methionine codon in the mRNA molecule, causing the loss of the first 81 amino acids of the protein, which is corroborated by data available on CFTR2 and ClinVar.

One example of a variant that showed disagreement between predictors is p.Arg117His (c.350G>A; legacy name R117H). Considering that the residue of arginine (positively charged) is preceded by two glutamic acids (negatively charged), its substitution for histidine, which is basic, aromatic, and uncharged, could disturb the local neutralization. Besides, codon 117 is in the interface between transmembrane and extracellular segments of the protein, which could also cause a local disturbance (Figure 4A). Since p.Arg117His is well known for being pathogenic when in *cis* with the 5T (c.1210-12[5]) allele of the poly-T tract (c.1210-12[5-9]), we also elaborated a model contemplating both *CFTR* variants (not shown). The combination p.[Arg117His;5T] (c.[350G>A;1210−12[5]]) in the same CFTR model did not offer conclusive results of its pathogenicity, which corroborates the disagreement verified in the analysis of *in silico* predictors.

Another chosen variant was p.Gly551Asp (c.1652G>A; legacy name G551D), which had 100% agreement between predictors as being deleterious. In our model (Figure 4A), the p.Gly551Asp could change the local molecular environment, establishing new interactions by replacing an amino acid that lacks a side chain for a negatively charged one, which may be the cause of its pathogenicity.

The last variant picked from the list was p.Ile1027Thr (c.3080T>C; legacy name I1027T).

This variant was considered neutral by all predictors with the exception of Polyphen1, and also considered benign by CFTR2 and ClinVar. This variant changes isoleucine (hydrophobic) for threonine (polar and uncharged) in the middle of a transmembrane helix, where the side chain is exposed to the lipid portion of the cell membrane. Theoretically, this unfavorable contact could induce conformational alterations that would hide the side chain more to the interior of the channel, perhaps affecting CFTR's function.

Sensitivity and specificity analysis

Sensitivity and specificity analysis was performed for CFTR2 and ClinVar. For each predictor, various cut-off ranges produced different sensitivity and specificity values as is evident by the receiver operating characteristic (ROC) curves on Figures 5A and 5B. For this analysis, we analyzed 57 missense variants (49 CF-causing and 8 Non CF-causing variants) for CFTR2 and 109 variants (97 CF-causing and 12 Non CF-causing variants) for ClinVar. Results are depicted in Table 4.

Sensitivity and specificity analysis for CFTR2 has shown that the highest areas under de curve (AUC) were obtained by PredictSNP (0.902), PhDSNP (0.899) and PolyPhen2 (0.903). For PredictSNP, the optimal cut-off value of >55% maximized both sensitivity (88%) and specificity (87.5%) for asserting *CFTR* missense variants as deleterious. PhDSNP had an optimal cut-off value of >77% (same value as the median of PhDSNP's expected accuracy) where both sensitivity (80%) and specificity (100%) were maximized for predicting a missense variant as pathogenic. Concerning PolyPhen2, the best cut-off value was >54%, where a good sensitivity (71%) and a great specificity (100%) can be achieved. Moreover, Youden Index J measures the performance of diagnostic tests, taking sensitivity and specificity into account . From the three predictors, PhDSNP has shown the highest Index J (0.7959), indicating that this predictor had the best performance when evaluating the CF-

causing and non CF-causing variants already annotated on CFTR2. On the opposite direction, the lower AUCs were generated by nsSNPAnalyzer (0.782), SIFT (0.784), MAPP (0.792) and SNAP (0.815). These predictors also had the lowest Index J (nsSNPAnalyzer=0.5638; SIFT=0.5867; MAPP=0.5459; SNAP=0.4668). As we can see in Table 5, these predictors have either low sensitivity or specificity. We believe that the reason for the poor performance of nsSNPAnalyzer, SIFT, MAPP and SNAP is the proportion of false positives and/or false negatives.

The analysis of sensitivity and specificity for ClinVar has shown that PhDSNP had the greatest AUCs (0.822). At a cut-off of >82%, a sensitivity of 48% and a specificity of 100% was achieved for predicting missense variants as deleterious. PhDSNP also had the highest Index J (0.4845) in the analysis of missense variants reported on ClinVar. On the other hand, MAPP and nsSNPAnalyzer were the tools with the lowest AUCs, respectively, 0.663 and 0.673. As we can see in Table 5, these predictors have either low sensitivity or specificity. Besides, nsSNPAnalyzer had the lowest Index J (0.3454). Overall, all predictors had a worse performance for ClinVar than they had for CFTR2, but both analyses showed that PhDSNP and Predict SNP seem to be the most appropriate predictors to assess the pathogenicity of *CFTR* missense variants.

Discussion

In this study, we submitted 779 *CFTR* missense variants to prediction analysis in the consensus classifier PredictSNP and compared outcomes to annotations on CFTR2 and ClinVar in order to determine which predictors present the most reliable interpretation of these variants. We found that PhDSNP and PredictSNP were the predictors with best performances in the analyses of expected accuracy, sensitivity and specificity. On the other hand, nsSNPAnalyzer presented the poorest results in these analyses.

Our results showed that predictors diverge when interpreting the pathogenicity of *CFTR* missense variants. Kappa analysis revealed that all predictors have only fair to moderate agreement among them. Furthermore, predictors also diverge from the information available in well-known variant annotation databases like CFTR2 and ClinVar. Kappa agreement was fair or moderate between predictors and variant annotation databases (Table 2). This analysis did not shed light of which predictor may be the best one to use, since nsSNPAnalyzer, for instance, has a moderate agreement to CFTR2 but has been one of the predictors with the poorest performance in most analyses. Interestingly, PhDSNP has shown fair agreement to most of the other predictors except for PolyPhen2, which showed good performance in the analysis of sensitivity and specificity.

The medians of expected accuracy of PhDSNP and PredictSNP do not differ significantly from each other, but both medians are significantly higher than the medians of most predictors included in the analysis. In addition, PhDSNP and PredictSNP seem to have the best parameters of sensitivity and specificity in the analysis of ROC curves for missense variants reported on CFTR2 and ClinVar (Figure 5). In contrast, PANTHER and nsSNPAnalyzer are the least reliable predictors to estimate the pathogenicity of *CFTR* missense variants. PANTHER did not have enough cases to show any significance in the analysis of kappa agreement between predictors and variant annotation databases, being excluded from further analysis. This can be explained by the fact that most of the investigated variants could not be found at their positions in the pre-computed multiple sequence alignments of PANTHER library . Concerning nsSNPAnalyzer, this predictor not just had a relatively low median of expected accuracy but it also presented the worst results in the analysis of sensitivity and specificity for both CFTR2 and ClinVar.

One factor that may contribute to the disagreements between the outcomes attributed by predictors resides in the algorithms employed by each one of these tools . Four predictors included in the analysis generated by PredictSNP – PhDSNP , PolyPhen2 , SNAP , nsSNPAnalyzer – apply machine-learning methods to train their decision models . From the other predictors, SIFT and PANTHER use only evolutionary information while MAPP also considers differences in the physicochemical properties of wildtype and mutated amino acids in their prediction. PolyPhen1 use a set of empirical rules in order to classify missense variants. This diversity in the way predictors analyze missense variants indicates that these tools should not be used as the only way to assert the pathogenicity of this type of variant .

It was evaluated if predictors would associate amino acid substitutions with a specific outcome, not taking protein location into consideration. In this context, Ile-Val changes were significantly associated with a neutral outcome by prediction tools. Biochemically, changing an isoleucine for a valine results in the substitution of a non-polar and aliphatic residue for another with a similar side chain. However, there were few Ile-Val substitutions that were asserted as pathogenic (Table 3). Noteworthy, p.Ile1234Val (c.3700A>G) is the only Ile-Val change already annotated on CFTR2. According to the database, p.Ile1234Val is a CF-causing variant. Concerning ClinVar, p.Ile1234Val is annotated as likely pathogenic. Moreover, ClinVar describes p.Ile119Val (c.355A>G) as pathogenic while p.Ile506Val (c.1516A>G) and p.Ile507Val (c.1519A>G) are considered benign/likely benign. The data above point to the fact that PhDSNP was able to predict correctly the outcome of the only Ile-Val substitution already reported in both databases. Also, predictors asserted wrongly two benign/likely benign variants that are adjacent to the pathogenic p.Phe508del, located in the nucleotide-binding domain 1. On the other hand, Leu-Pro substitutions were significantly associated with a deleterious outcome. Here, changing a leucine for a proline promotes the substitution of a non-polar and aliphatic residue for a polar residue with a cyclic side chain,

which could indeed alter the local stability where these amino acid changes are inserted. In fact, proline is an amino acid with very peculiar characteristics, lacking an amide proton, which makes the amide nitrogen in the main chain incapable of forming hydrogen bonds . In addition, the pyrrolidine ring of proline is rigid, constraining the ϕ angle of the peptide chain to values close to -65°. Proline also restricts the conformation of the preceding residue in the protein sequence . Only two Leu-Pro substitutions were associated with the neutral outcome: p.Leu636Pro and Leu1480Pro. The first one is located in an α-helix of the regulatory domain and has been postulated as having mild effect, being associated with congenital bilateral absence of vas deferens – CBAVD . However, p.Leu636Pro is not found on CFTR2 nor ClinVar. In addition, SNAP also considered the missense variant p.Leu1480Pro (c.4439T>C) as neutral. Unfortunately, information about the annotation of p.Leu1480Pro is not available on CFTR2, ClinVar or in the literature.

Considering the structure of the CFTR protein, we had some interesting findings. Regarding the agreement between predictors in relation to neutral or deleterious outcomes, we observed that missense variants located in nucleotide-binding domains (NBD1 and NBD2), β-strands and bends are associated with full agreement between computational tools (Figure 3). One explanation for it resides in the fact that all β-strands present in the CFTR structure are located either in NBD1 or NBD2. Noteworthy, the CFTR protein a peculiar member of the ABC superfamily and it is fundamentally composed by two halves, each half composed of one MSD and one NBD (Figure 4B). On the other hand, missense variants located in the membrane-spanning domain 1 (MSD1) and in α-helices are associated with at least one disagreement between predictors (Figure 3). Herein, not just MSD1 and MSD2 are basically formed α-helices but this secondary structure is present in other domains. The structure 5UAK curated in the Protein Data Bank used as a reference to determine the secondary structure where is each *CFTR* missense variant corroborates the data above. Our results

suggest that predictors tend to agree more when they assert variants as deleterious, and the opposite is also true, since predictors tend to disagree more when asserting variants as neutral.

The molecular model elaborated with p.Met1Val, p.Arg117His, p.Gly551Asp and p.Ile1027Thr made it possible to better rationalize the effect of these variants in the CFTR protein, and some of the affirmations generated by our model were corroborated by CFTR2 and ClinVar data. In the case of p.Met1Val, translation initiation at the first methionine would be aborted, promoting the loss of the first 81 amino acids of the protein sequence (Figure 4B), which includes the Lasso domain . Our data is corroborated by CFTR2 and ClinVar, where p.Met1Val is classified, respectively, as “CF-causing” and “Pathogenic”. In fact, there are 26 patients in the CFTR2 database that carry this mutation .

A variant that showed disagreement between predictors is p.Arg117His (c.350G>A; legacy name R117H), which is described as having varying clinical consequence by CFTR2. In the same database, there are 1,817 patients that carry p.Arg117His . According to ClinVar, this variant is pathogenic, has conflicting interpretation of pathogenicity and is also a risk factor. In addition, one intragenic modifier, the poly-T tract (c.1210-12[5-9]), is present in every copy of *CFTR* in one of its three forms: 9T (c.1210-12[9]), 7T (c.1210-12[7]) and 5T (c.1210-12[5]). The 5T form, which is present in ~5% of the *CFTR* alleles – ~10% of individuals in general population – may cause the skipping of exon 10 during mRNA processing and it is associated with partial penetrance of CF. Therefore, when p.Arg117His is in *cis* with c.1210-12[5], this combination is considered as CF-causing according to CFTR2, being carried by 102 patients . CFTR2 highlights that some patients who carry p.Arg117His may also be carrying c.1210-12[5] in the same chromosome, even though these patients do not have their specific poly-T tract determined. It is important to emphasize that p.Arg117His

and c.1210-12[5] do not cause CF when they are alone or in *trans* . According to our model, the combination c.[350G>A;1210−12[5]] does not offer conclusive results of its pathogenicity, which reinforces the inconclusive predictions of *in silico* tools and contradicts curated data . Perhaps, the effect of p.Arg117His would be better predicted if prediction tools had editing fields to input relevant information that might result in a more efficient outcome prediction. Finally, the amino acid change generated by this mutation affects the function of the CFTR protein. When arginine is substituted by a histidine, the conductance of CFTR is affected, thus impairing the flow of chloride ions . Recently, ivacaftor (Kalydeco®; Vertex Pharmaceuticals Inc., Boston, MA), a drug already approved by the **U.S. Food and Drug Administration** for the treatment of *CFTR* missense variants that affect gating, was approved to treat patients carrying p.Arg117His as well .

The substitution of a glycine for an aspartate on codon 551 (p.Gly551Asp; c.1652G>A; legacy name G551D) has been reported as CF-causing by CFTR2 and pathogenic by ClinVar. This variant does not affect the amount of CFTR protein available in the cell membrane. Instead, its pathogenicity relies upon the functional activity of CFTR, impairing the gating of this chloride channel. The p.Gly551Asp variant was also chosen to be featured in our model (Figure 4A) because it is one of the two *CFTR* missense variants carried by more than 1% of CF patients , being a well-known therapeutic target of ivacaftor. In fact, there are 2,915 patients carrying this variant in the CFTR2 database. All these pieces of evidence corroborate the fact that predictors could predict correctly the deleterious effect of p.Gly551Asp.

The last variant picked from the list was p.Ile1027Thr (c.3080T>C; legacy name I1027T). Although this substitution could affect the permeability of the CFTR channel, it was considered neutral by all predictors except for PolyPhen1. Concerning CFTR2, p.Ile1027Thr is a non-CF causing variant. This annotation was based on clinical information of patients

carrying this variant, experimental data generated from this variant, and on groups of healthy individuals that carry p.Ile1027Thr. CFTR2 also reports that there are 36 patients carrying this variant in its database. Concerning ClinVar, p.Ile1027Thr is a benign or likely benign variant.

In our analysis of sensitivity and specificity, we could observe in both ROC curves that PhDSNP and PredictSNP had the best performances, which adds to the conclusion that these predictors seem to be the most reliable when evaluating *CFTR* missense variants. Even though PhDSNP does not present the highest sensitivity, it is important to highlight that a higher specificity would be preferred instead of a higher sensitivity, since molecular diagnosis is not used for screening but as a complementary form of CF diagnosis, where the clinical features and the sweat chloride test have already been performed in the patients . Hence, the goal would be avoiding false positive results.

One important limitation that we encountered was the lack of missense variants reported in the variant annotation databases. In the case of CFTR2, especially, only the 322 most common variants in the *CFTR* gene can be found and only 74 of them cause amino acid substitutions . Regarding ClinVar, this database receives information from different submitters (CFTR2 being one of them), and there are 146 *CFTR* missense variants curated in it. However, these numbers represent less than 20% of the 779 missense variants submitted to prediction analysis in this study.

We could verify that the great difficulty of prediction tools is excluding the non-pathogenicity of missense variants, corroborating previous studies that showed the low specificity of *in silico* predictors and the superprediction of missense variants as deleterious . Overall, i*n silico* analysis, as part of the molecular analysis of *CFTR*, should always be correlated with clinical – signals and symptoms – and physiological data in order to

determine CF diagnosis. Concomitantly, the further determination of pathogenicity and the reevaluation of missense variants curated in annotation databases like CFTR2 and ClinVar are fundamental, mainly because there are those cases of positive NBS, inconclusive diagnosis, and CFTR-related metabolic syndrome , where CF diagnosis is very difficult to achieve.

To our knowledge, this is the first study that employed a consensus predictor to evaluate a large number of *CFTR* missense variants and compare the predicted outcomes to publicly available variant annotation databases (CFTR2 and ClinVar). As shown by the results presented in previous sections, PhDSNP and PredictSNP seem to be the most appropriate tools to analyze missense variants in the *CFTR* gene.

Disclosure

None of the contributing authors have a conflict of interest.

Acknowledgements

Financial support: The authors would like to thank CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) and FIPE-HCPA (Fundo de Apoio à Pesquisa e Eventos – Hospital de Clínicas de Porto Alegre).

References

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. 2010. A method and server for predicting damaging missense mutations. Nat Methods 7:248–249.

Bajaj K, Madhusudhan MS, Adkar B V, Chakrabarti P, Ramakrishnan C, Sali A, Varadarajan R. 2007. Stereochemical Criteria for Prediction of the Effects of Proline Mutations on Protein Stability. PLoS Comput Biol 3:e241.

Bao L, Zhou M, Cui Y. 2005. nsSNPAnalyzer: Identifying disease-associated nonsynonymous single nucleotide polymorphisms. Nucleic Acids Res 33:480–482.

Bendl J, Stourac J, Salanda O, Pavelka A, Wieben ED, Zendulka J, Brezovsky J, Damborsky J. 2014. PredictSNP: Robust and Accurate Consensus Classifier for Prediction of Disease-Related Mutations. PLoS Comput Biol 10:e1003440.

Bergougnoux A, Taulan-cadars M, Claustres M, Raynal C. 2015. New Molecular Diagnosis Approaches — From the Identification of Mutations to their Characterization. In: Wat D, editor. Cystic Fibrosis in the Light of New Research, InTech, p 201–231.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. 2000. The protein data bank. Nucleic Acids Res 28:235–242.

Brennan ML, Schrijver I. 2016. Cystic Fibrosis: A Review of Associated Phenotypes, Use of Molecular Diagnostic Approaches, Genetic Characteristics, Progress, and Dilemmas. J Mol Diagnostics 18:3–14.

Bromberg Y, Rost B. 2007. SNAP: Predict effect of non-synonymous polymorphisms on function. Nucleic Acids Res 35:3823–3835.

Brunham LR, Singaraja RR, Pape TD, Kejariwal A, Thomas PD, Hayden MR. 2005. Accurate prediction of the functional significance of single nucleotide polymorphisms and mutations in the ABCA1 gene. PLoS Genet 1:e83.

Capriotti E, Calabrese R, Casadio R. 2006. Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. Bioinformatics 22:2729–2734.

CFTR1. 1989. Cystic Fibrosis Mutation Database (CFTR1). Cystic Fibrosis Centre, The Hospital for Sick Children, Toronto, Canada. Available at: http://www.genet.sickkids.on.ca/app. Accessed on: 2016/10/08.

CFTR2. 2011. The Clinical and Functional Translation of CFTR (CFTR2). Available at: http://cftr2.org. Accessed on: 2017/05/25.

Chandonia J, Hon G, Walker NS, Conte L Lo, Koehl P, Levitt M, Brenner SE. 2004. The ASTRAL Compendium in 2004. Nucleic Acids Res 32:D189–D192.

Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. 2012. Predicting the Functional Effect of Amino Acid Substitutions and Indels. PLoS One 7:e46688.

Cutting GR. 2015. Cystic fibrosis genetics: from molecular understanding to clinical application. Nat Rev Genet 16:45–56.

Dorfman R, Nalpathamkalam T, Taylor C, Gonska T, Keenan K, Xw Y, Corey M, L-c T, Zielenski J, Do DP. 2010. Do common in silico tools predict the clinical consequences of amino-acid substitutions in the CFTR gene ? Clin Genet 77:464–473.

Fanen P, Wohlhuter-Haddad A, Hinzpeter A. 2014. Genetics of cystic fibrosis: CFTR mutation classifications toward genotype-based CF therapies. Int J Biochem Cell Biol 52:94–102.

Farrell PM, White TB, Ren CL, Hempstead SE, Accurso F, Derichs N, Howenstine M, McColley SA, Rock M, Rosenfeld M, Sermet-Gaudelus I, Southern KW, et al. 2017. Diagnosis of Cystic Fibrosis: Consensus Guidelines from the Cystic Fibrosis Foundation. J Pediatr 181:S4–S15.e1.

Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M,

Qureshi M, Sangrador-Vegas A, Salazar GA, Tate J, et al. 2016. The Pfam protein families database: Towards a more sustainable future. Nucleic Acids Res 44:D279–D285.

Gadsby DC, Vergani P, Csanády L. 2006. The ABC protein turned chloride channel whose failure causes cystic fibrosis. Nature 440:477–483.

Hicks S, Wheeler DA, Plon SE, Kimmel M. 2011. Prediction of missense mutation functionality depends on both the algorithm and sequence alignment employed. Hum Mutat 32:661–668.

Holland IB, Cole SPC, Kuchler K, Higgins CF. 2003. ABC Proteins From Bacteria to Man. London: Academic Press.

Larrañaga P, Calvo B, Santana R, Bielza C, Galdiano J, Inza I, Lozano JA, Armañanzas R, Santafé G, Pérez A, Robles V. 2006. Machine learning in bioinformatics. Brief Bioinform 7:86–112.

Liu F, Zhang Z, Csanády L, Gadsby DC, Chen J. 2017. Molecular Structure of the Human CFTR Ion Channel. Cell 169:85–95.e8.

Lucarelli M, Pierandrei S, Bruno SM, Strom R. 2012. The Genetics of CFTR : Genotype – Phenotype Relationship , Diagnostic Challenge and Therapeutic Implications. In: Sriramulu D, editor. Cystic Fibrosis – Renewed Hopes Through Research, InTech, p 91–122.

Manickam M, Ravanan P, Singh P, Talwar P. 2014. In silico identification of genetic variants in glucocerebrosidase (GBA) gene involved in Gaucher's disease using multiple software tools. Front Genet 5:148.

Masica DL, Karchin R. 2016. Towards Increasing the Clinical Relevance of In Silico Methods to Predict Pathogenic Missense Variants. PLoS Comput Biol 12:e1004726.

Ng PC, Henikoff S. 2003. SIFT: Predicting amino acid changes that affect protein function.

Nucleic Acids Res 31:3812–3814.

OMIM. 2017. Online Mendelian Inheritance in Man, OMIM®. Johns Hopkins University, Baltimore, MD. MIM Number: {602421}: {12/08/2016}. World Wide Web URL: https://omim.org/.

Palomaki GE, Fitzsimmons SC, Haddow JE. 2004. Clinical sensitivity of prenatal screening for cystic fibrosis via CFTR carrier testing in a United States panethnic population. Genet Med 6:405–414.

Ramensky V, Bork P, Sunyaev S. 2002. Human non-synonymous SNPs: server and survey. Nucleic Acids Res 30:3894–3900.

Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hegde M, Lyon E, Spector E, Voelkerding K, Rehm HL. 2015. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med 17:405–423.

Riordan JR, Rommens JM, Kerem B, Alon N, Rozmahel R, Grzelczak Z, Zielenski J, Lok S, Plavsic N, Chou J, Drumm ML, Iannuzzi MC, et al. 1989. Identification the Cystic Fibrosis Gene : Cloning and Characterization of Complementary DNA. Science (80-) 245:1066–1073.

Rommens JM, Iannuzzi MC, Kerem B, Mitchell L, Melmer G, Dean M, Rozmahel R, Cole JL, Kennedy D, Hidaka N, Zsiga M, Buchwald M, et al. 1989. Identification of the Cystic Fibrosis Gene: Chromosome Walking and Jumping. Science (80-) 245:1059–1065.

Roy A, Kucukural A, Zhang Y. 2010. I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protoc 5:725–738.

Salvatore D, Dell'Edera D, Colangelo C, Smaldore G. 2015. p.Leu636Pro mutation is

associated with cystic fibrosis transmembrane conductance regulator-related disorders (congenital bilateral absence of vas deferens). Int J Urol 22:803–804.

Saraiva-Pereira ML, Fitarelli-Kiehl M, Sanseverino MTV. 2011. Genetics of Cystic Fibrosis. Rev HCPA 31:160–167.

Sheppard DN, Rich DP, Ostedgaard LS, Gregory RJ, Smith AE, Welsh MJ. 1993. Mutations in CFTR associated with mild-disease-form Cl^- channels with altered pore properties. Nature 362:160–164.

Silva Filho LVRF, Castaños C, Ruíz HH. 2016. Cystic fibrosis in Latin America—Improving the awareness. J Cyst Fibros 15:791–793.

Stone EA, Sidow A. 2005. Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. Genome Res 15:978–986.

Tavtigian S V., Greenblatt MS, Lesueur F, Byrnes GB. 2008. In silico analysis of missense substitutions using sequence-alignment based methods. Hum Mutat 29:1327–1336.

The UniProt Consortium. 2017. UniProt: the universal protein knowledgebase. Nucleic Acids Res 45:D158–D169.

Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, Diemer K, Muruganujan A, Narechania A. 2003. PANTHER: A library of protein families and subfamilies indexed by function. Genome Res 13:2129–2141.

Thompson BA, Greenblatt MS, Vallee MP, Herkert JC, Tessereau C, Young EL, Adzhubey IA, Li B, Bell R, Feng B, Mooney SD, Radivojac P, et al. 2013. Calibration of Multiple In Silico Tools for Predicting Pathogenicity of Mismatch Repair Gene Missense Substitutions. Hum Mutat 34:255–265.

Thusberg J, Olatubosun A, Vihinen M. 2011. Performance of mutation pathogenicity prediction methods on missense variants. Hum Mutat 32:358–368.

Vertex Pharmaceuticals Inc. 2014. U.S. Food and Drug Administration Approves KALYDECO® (ivacaftor) for Use in People with Cystic Fibrosis Ages 6 and Older Who Have the R117H Mutation. Available at: http://investors.vrtx.com/releasedetail.cfm?ReleaseID=889027. Accessed on: 2017/07/25.

Viera AJ, Garrett JM. 2005. Understanding Interobserver Agreement: The Kappa Statistic. Fam Med 37:360–363.

WHO Human Genetics Programme. 2004. The molecular genetic epidemiology of cystic fibrosis: report of a joint meeting of WHO/IECFTN/ICF(M)A/ECFS, Genoa, Italy, 19 June 2002. Geneva: World Health Organization, p 1–24.

Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y. 2015. The I-TASSER Suite: Protein structure and function prediction. Nat Methods 12:7–8.

Youden WJ. 1950. Index for rating diagnostic tests. Cancer 3:32–35.

Zhang Y. 2008. I-TASSER server for protein 3D structure prediction. BMC Bioinformatics 9:40.

Figure Legends

Figure 1. Workflow. Starting from 2,009 *CFTR* variants on CFTR1 (on November, 26th 2016), we used HGMD® Professional 2016.2 Trial Version to obtain the 779 missense variants analyzed in the study (insertions, deletions, splice site, nonsense, frameshift variants were excluded). Selected variants were then submitted to prediction analysis on consensus classifier PredictSNP. Variant annotation was retrieved from CFTR2 and ClinVar. Information about the location (topology, domain and secondary structure) of amino acid substitutions generated by *CFTR* missense variants were gathered from UniProt, Pfam, CFTR1 and PDB. Also, we chose four variants that represent agreement and/or disagreement of the outcome attributed by each predictor to build a model of the CFTR structure. Lastly, statistical analysis was performed and finally, we analyzed all data we collected. CFTR1: Cystic Fibrosis Mutation Database; HGMD: Human Gene Mutation Database; SNP: single nucleotide polymorphism; UniProt: Universal Protein Resource; PDB: Protein Data Bank; ROC: Receiver operating characteristic curve.

Figure 2. Comparison of accuracy among predictors. A) Comparison of accuracy for each predictor. Columns represent median and percentiles (25;75). Bars represent minimum and maximum; ○ represent outliers; n=726; $p<0.05$. **B)** Comparison of accuracies for CF-causing variants according to CFTR2 annotation. Columns represent median and percentiles (25,75). Bars represent minimum and maximum; * and ○ represent outliers; n=49; $p<0.05$. **C)** Comparison of accuracies according to ClinVar annotation as "Pathogenic" or "Likely pathogenic". Columns represent median and percentiles (25,75). Bars represent minimum and maximum; * and ○ represent outliers; n=97; $p<0.05$. PANTHER was not included in this analysis.

Figure 3. Agreement between predictors according to CFTR protein structure. A) Agreement between predictors in relation to CFTR topology. Even though there was a tendency of association between full agreement and cytoplasm, it was not statistically significant (p=0.052) **B)** Agreement between predictors in relation to CFTR domains. NBD1 and NBD2 are significantly associated with full agreement between predictors. MSD1 is significantly not associated with full agreement and significantly associated with two disagreements between predictors (p<0.001). **C)** Agreement between predictors in relation to the secondary structure of CFTR. Beta-strands and bends are significantly associated with full agreement between predictors whilst alpha-helices are significantly not associated with full agreement between predictors. Overall, predictors tend to agree more when they predict missense variants as deleterious (black columns). As the disagreements increase, the neutral outcome gets more frequent (white columns). *p<0.05; †: p>0.05. MAPP and PANTHER were not included in this analysis.

Figure 4. Modeled structure of CFTR. A) Modeled structure for CFTR in its proposed location when inserted in biomembranes (dotted lines), evidencing the absence of extracellular domains. Proposed variants are highlighted, while two segments that may be lost (at least partially) as a function of these mutations are highlighted in orange (associated with p.Met1Val) and in yellow (associated with p.Arg117His). Details regarding these deletions are given in the results session. **B)** Main structural features of human CFTR. The semi-symmetric structure is composed of two equivalent halves, comprising two domains (MSD and NBD) each. The N-terminus half of the protein has exclusive topological regions, namely the Lasso domain and the R domain (or R insertion). MSD: Membrane-spanning domain; NBD: Nucleotide-binding domain; R domain: Regulatory domain.

Figure 5. ROC curve for variants on CFTR2 and ClinVar. A) ROC curve for variants on CFTR2. Receiver operating characteristic (ROC) curves elaborated for missense variants according to CFTR2. n=57 (CF-causing = 49; non CF-causing = 8). **B)** ROC curve for variants on ClinVar. Receiver operating characteristic (ROC) curves elaborated for missense variants according to ClinVar. n=109 (Pathogenic/Likely pathogenic = 97; Benign/Likely benign = 12). PANTHER was not included in this analysis.

Table 1. Descriptive analysis of each predictor for *CFTR* missense variants (n=779[a])

Predictor		PredictSNP	MAPP	PhDSNP	PolyPhen1	PolyPhen2	SIFT	SNAP	nsSNPAnalyzer	PANTHER
Sample number	**Valid**	779	726	779	779	779	779	779	779	291
	Missing	0	53	0	0	0	0	0	0	488
Pathogenicity	**Neutral**	286 (36.7%)	339 (46.7%)	218 (28.0%)	400 (51.3%)	287 (36.8%)	211 (27.1%)	264 (33.9%)	292 (37.5%)	192 (66.0%)
	Deleterious	493 (63.3%)	387 (53.3%)	561 (72.0%)	379 (48.7%)	492 (63.2%)	568 (72.9%)	515 (66.1%)	487 (62.5%)	99 (34.0%)
Mean of Expected Accuracy[b] (SD)		73 (11.8)	69 (12.3)	73 (12.6)	67 (5.2)	66 (13.9)	71 (13.0)	70 (12.0)	64 (1.0)	61 (7.9)
Minimum		51	41	45	59	40	43	50	63	47
Maximum		87	92	98	74	87	90	89	65	74
Percentiles	**25**	63	62	61	67	55	65	61	63	56
	50 (Median)	74	72	77	67	68	79	72	63	63
	75	87	77	86	74	81	79	81	65	68

[a]Missense variants were retrieved from HGMD® Professional 2016.2 Trial Version on 09/29/2016.

[b]Expected accuracy is a comparable scale ranging from 0–100% which represents the transformed confidence scores of individual predictors. Descriptive data of each predictor are depicted. Missing values represent that predictors were not able to assign the pathogenicity of a variant. SD: Standard deviation.

Table 2. Kappa agreement between predictors and variant annotation databases (CFTR2 and ClinVar) for *CFTR* missense variants

	PredictSNP	MAPP	PhDSNP	PolyPhen1	PolyPhen2	SIFT	SNAP	nsSNPAnalyzer	PANTHER
PredictSNP	-	-	-	-	-	-	-	0.538	0.492
MAPP	-	-	0.322	0.496	0.449	0.362	0.325	0.368	0.304
PhDSNP	-	-	-	0.335	0.416	0.366	0.324	0.331	0.230
PolyPhen1	-	-	-	-	0.488	0.379	0.353	0.439	0.482
PolyPhen2	-	-	-	-	-	0.545	0.335	0.508	0.433
SIFT	-	-	-	-	-	-	0.352	0.504	0.375
SNAP	-	-	-	-	-	-	-	0.285	0.251
nsSNPAnalyzer	-	-	-	-	-	-	-	-	0.378
PANTHER	-	-	-	-	-	-	-	-	-
CFTR2	0.518	0.332	0.324	0.482	0.481	0.363	0.398	0.561	0.459†
ClinVar	0.302	0.039†	0.407	0.208	0.306	0.297	0.242	0.268	0.215†

Kappa agreement between predictors and variant annotation databases (CFTR2 and ClinVar) for *CFTR* missense variants is shown.
Categorical outcome ("neutral" or "deleterious") was employed in this analysis.
Interpretation of Kappa (Viera & Garrett, 2005): <0.000 - Less than chance agreement; 0.010-0.200 - Slight agreement; 0.210-0.400 - Fair agreement; 0.410-0.600 - Moderate agreement; 0.610-0.800 - Substantial agreement; 0.810-0.990 - Almost perfect agreement; 1.000 – Perfect agreement.
All values displayed in this table are statistically significant ($p<0.05$) except the labeled ones (†: $p>0.05$).
The valid sample number of each predictor (shown in Table 1) was used for calculating Kappa:
- For most predictors, n=779; for MAPP, n=726; for PANTHER, n=291.
- For CFTR2, 57 missense variants listed on CFTR2 were included (CF-causing = 49; Non CF-causing = 8).
- For ClinVar, 109 missense variants reported on ClinVar were included (Pathogenic/Likely pathogenic = 97; Benign/Likely benign = 12).

Table 3. Amino acid substitutions associated with "neutral" or "deleterious" outcomes by each predictor regardless of protein location

Amino acid change	PredictSNP*	MAPP*	PhDSNP*	PolyPhen1*	PolyPhen2*	SIFT*	SNAP*
Asp-Asn			n (5/11) d (6/11)	n (9/11)			
Asp-Tyr					n (10/10)		
Glu-Lys		n (10/14) d (4/14)					
Gly-Arg		d (16/20)	d (21/21)	d (17/21)			
Gly-Val				d (11/13)			
Ile-Thr			n (6/13) d (7/13)				
Ile-Val	n (13/14)		n (13/14)	n (14/14)	n (9/14) d (5/14)	n (11/14)	n (12/14)
Lys-Glu				n (9/10)	n (7/10)		
Leu-Phe				n (9/11)			
Leu-Pro	d (25/26)	d (25/25)	d (26/26)	d (25/26)	d (25/26)	d (25/26)	d (24/26)
Leu-Ser							d (13/14)
Met-His				n (11/12)			
Met-Ile				n (12/13)			
Ser-Phe				d (9/10)			
Thr-Ile			n (6/12) d (6/12)				

The first amino acid represents the one in the canonical protein sequence and the second one represents the mutated amino acid.
"n" = neutral; "d" = deleterious; the parentheses represent, respectively, the number of substitutions associated with the outcome and the absolute frequency of this substitution in the wild-type CFTR protein.
Pearson's Chi-squared or Fisher's Exact Test was used when appropriate. *: $p<0.05$.
Associations depicted in this table represent amino acid substitutions that are significantly different from the rest of the substitutions analyzed by each predictor, whether polarizing the substitution towards one outcome or not.
nsSNPAnalyzer did not associate any outcome to the substitutions analyzed ($p>0.05$). PANTHER was not included in this analysis
Leu-Pro substitutions that were not associated with the deleterious outcome by each preditor are: p.Leu636Pro (c.1907T>C; PredictSNP, PolyPhen1, Polyphen2, SIFT, SNAP) and p.Leu1480Pro (c.4439T>C; SNAP).

Ile-Val substitutions that were not associated with the neutral outcome by each predictor are: p.Ile119Val (c.355A>G; SNAP), p.Ile497Val (c.1489A>G; PolyPhen2), p.Ile506Val (c.1516A>G; PredictSNP, PolyPhen2, SIFT, SNAP), p.Ile507Val (c.1519A>G; SIFT), p.Ile556Val (c.1666A>G; PolyPhen2), p.Ile1051Val (c.3151A>G; PolyPhen2), p.Ile1131Val (c.3391A>G; PolyPhen2, SIFT) and p.Ile1234Val (c.3700A>G; PhDSNP).

Table 4. Sensitivity and Specificity of Each Predictor according to Variant Annotation Databases

	CFTR2 (n=57)					ClinVar (n=112)				
Predictor	**AUC**	**Cut-off (% of expected accuracy)**	**Sensitivity (%)**	**Specificity (%)**	**Youden Index J**	**AUC**	**Cut-off (% of expected accuracy)**	**Sensitivity (%)**	**Specificity (%)**	**Youden Index J**
PredictSNP	0.902	>55	88	87.5	0.7526	0.763	>-63	86	58	0.4390
MAPP	0.792	>-70	80	75	0.5459	0.663	>57	44	100	0.4433
PhDSNP	0.899	>77	80	100	0.7959	0.822	>82	48	100	0.4845
PolyPhen1	0.874	>-67	82	87.5	0.6913	0.697	>-67	71	67	0.3780
PolyPhen2	0.903	>54	71	100	0.7143	0.783	>-61	79	67	0.4605
SIFT	0.784	>45	84	75	0.5867	0.720	>45	79	67	0.4605
SNAP	0.815	>72	59	87.5	0.4668	0.735	>56	77	58	0.3565
nsSNPAnalyzer	0.782	>-65	94	62.5	0.5638	0.673	>-65	84.5	50	0.3454

Negative cut-off values represent cut-offs for the neutral outcome.
AUC = area under the curve.
Youden Index J (Youden, 1950) is a statistical test to rate diagnostic tests, ranging from 0 to 1 (values close to 1 represent predictors with better performances).

All data presented in this table are statistically significant ($p<0.05$). Youden Index is calculated according to the equation: $J = \frac{ad-bc}{(a+b)(c+d)}$

PANTHER was not included in this analysis.

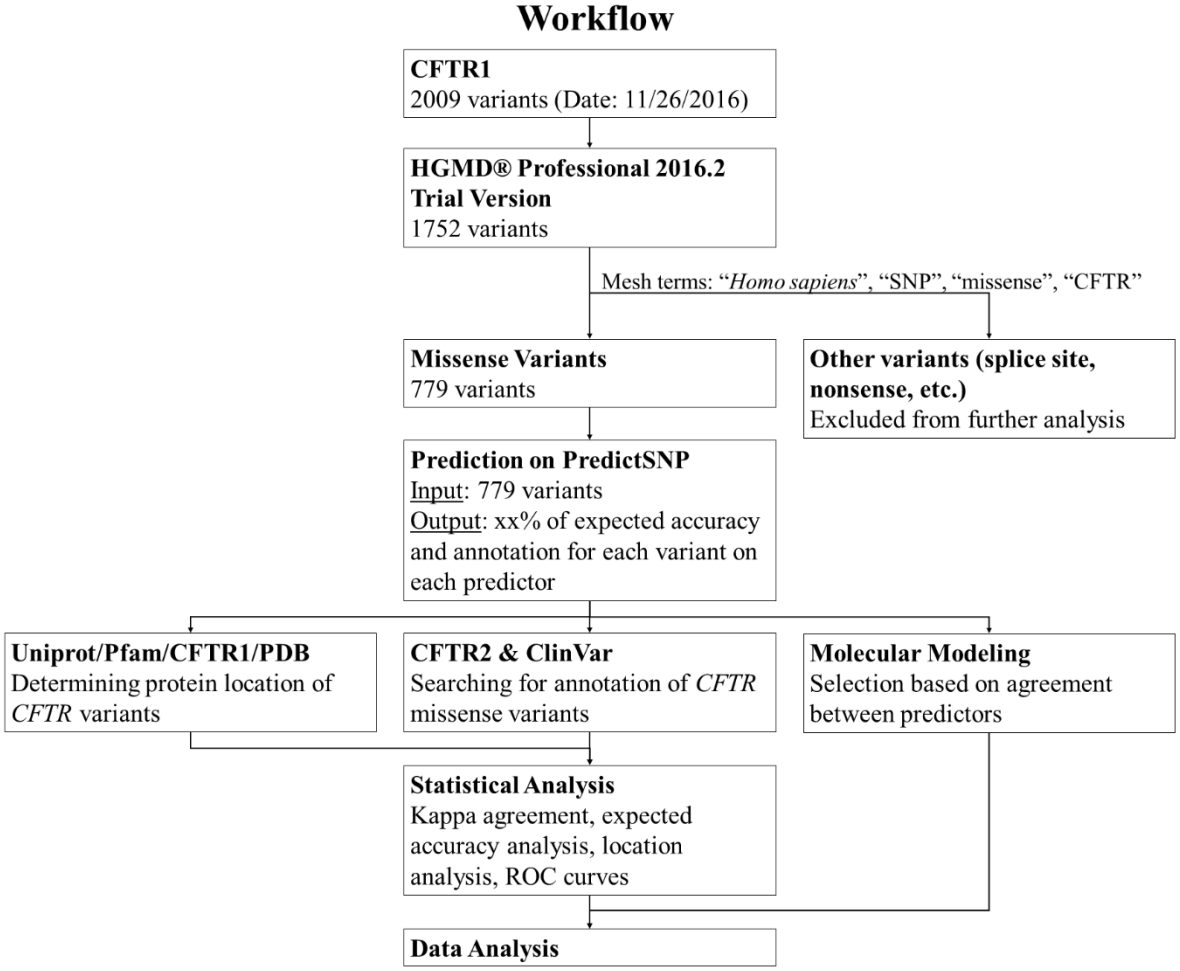


Figure 1. Workflow. Starting from 2,009 *CFTR* variants on CFTR1 (on November, 26th 2016), we used HGMD® Professional 2016.2 Trial Version to obtain the 779 missense variants analyzed in the study (insertions, deletions, splice site, nonsense, frameshift variants were excluded). Selected variants were then submitted to prediction analysis on consensus classifier PredictSNP. Variant annotation was retrieved from CFTR2 and ClinVar. Information about the location (topology, domain and secondary structure) of amino acid substitutions generated by *CFTR* missense variants were gathered from UniProt, Pfam, CFTR1 and PDB. Also, we chose four variants that represent agreement and/or disagreement of the outcome attributed by each predictor to build a model of the CFTR structure. Lastly, statistical analysis was performed and finally, we analyzed all data we collected. CFTR1: Cystic Fibrosis Mutation Database; HGMD: Human Gene Mutation Database; SNP: single nucleotide polymorphism; UniProt: Universal Protein Resource; PDB: Protein Data Bank; ROC: Receiver operating characteristic curve.

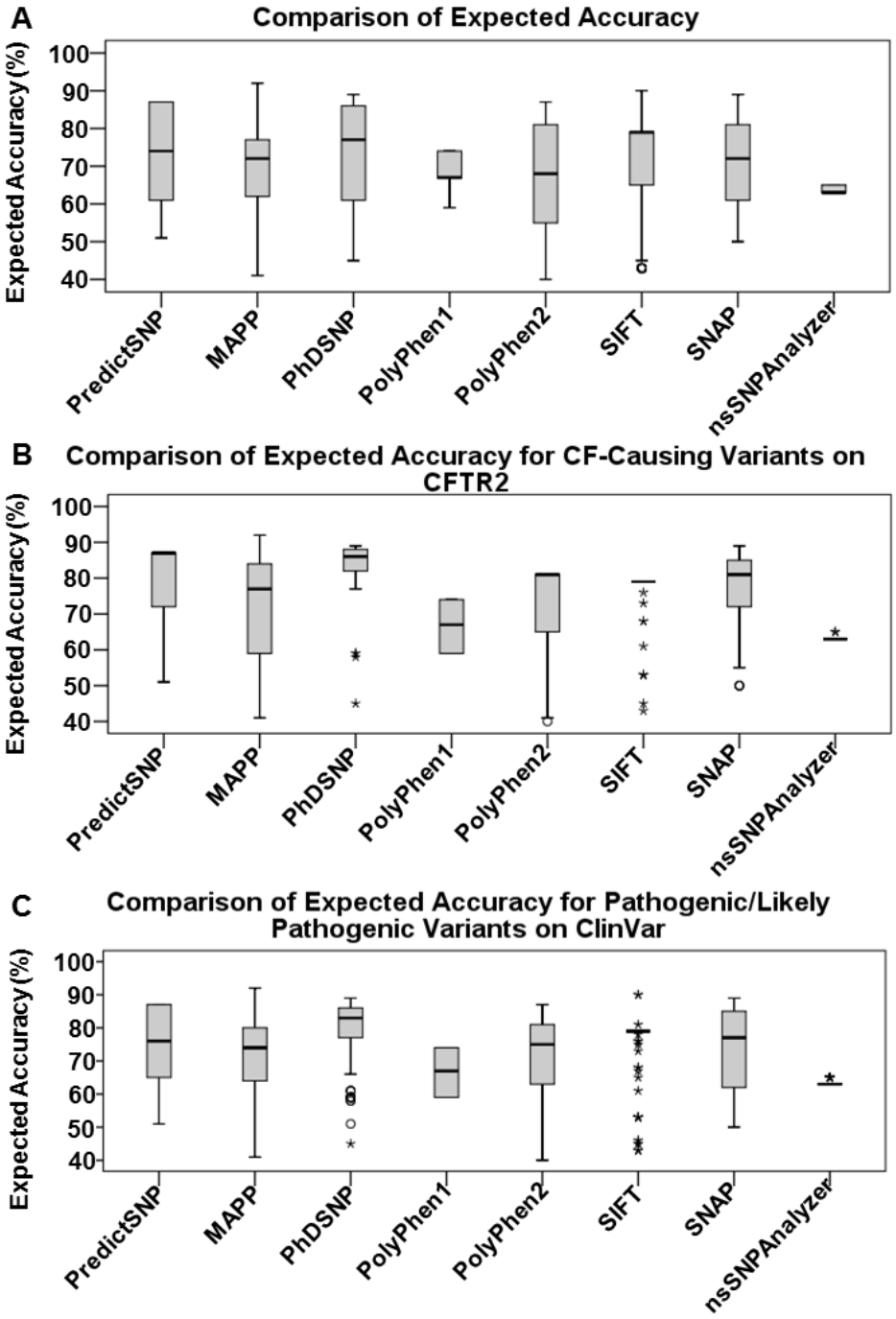


Figure 2. Comparison of accuracy among predictors. A) Comparison of accuracy for each predictor. Columns represent median and percentiles (25;75). Bars represent minimum and maximum; ○ represent outliers; n=726; $p<0.05$. **B)** Comparison of accuracies for CF-causing variants according to CFTR2 annotation. Columns represent median and percentiles (25,75). Bars represent minimum and maximum; * and ○ represent outliers; n=49; $p<0.05$. **C)** Comparison of accuracies according to ClinVar annotation as "Pathogenic" or "Likely pathogenic". Columns represent median and percentiles (25,75). Bars represent minimum and maximum; * and ○ represent outliers; n=97; $p<0.05$. PANTHER was not included in this analysis.

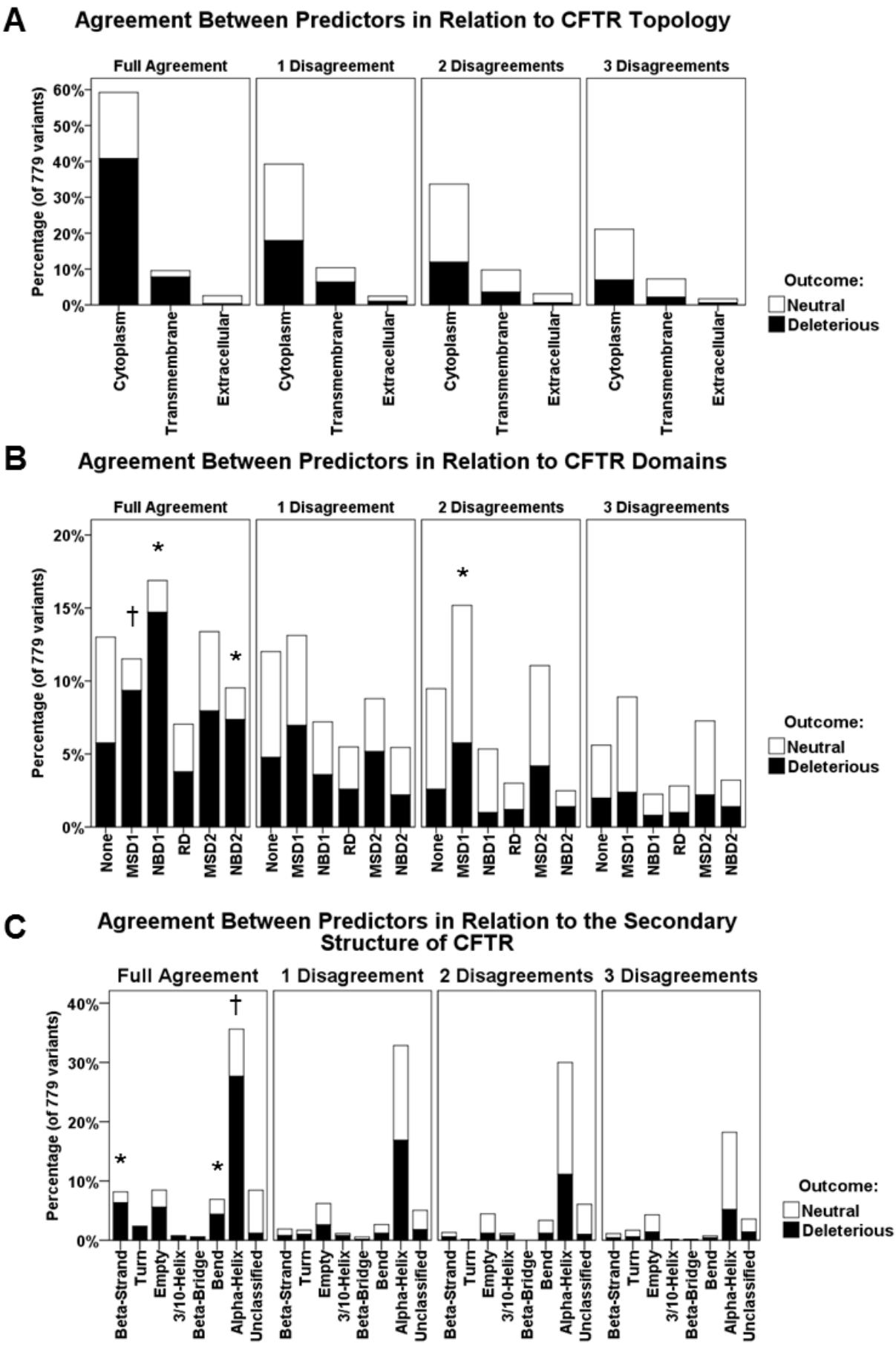

A
Agreement Between Predictors in Relation to CFTR Topology
Full Agreement
1 Disagreement
2 Disagreements
3 Disagreements
Percentage (of 779 variants)
60%
50%
40%
30%
20%
10%
0%
Cytoplasm
Transmembrane
Extracellular
Outcome:
Neutral
Deleterious
B
Agreement Between Predictors in Relation to CFTR Domains
Full Agreement
1 Disagreement
2 Disagreements
3 Disagreements
Percentage (of 779 variants)
20%
15%
10%
5%
0%
None
MSD1
NBD1
RD
MSD2
NBD2
*
†
Outcome:
Neutral
Deleterious
C
Agreement Between Predictors in Relation to the Secondary Structure of CFTR
Full Agreement
1 Disagreement
2 Disagreements
3 Disagreements
Percentage (of 779 variants)
40%
30%
20%
10%
0%
Beta-Strand
Turn
Empty
3/10-Helix
Beta-Bridge
Bend
Alpha-Helix
Unclassified
*
†
Outcome:
Neutral
Deleterious

Figure 3. Agreement between predictors according to CFTR protein structure. A) Agreement between predictors in relation to CFTR topology. Even though there was a tendency of association between full agreement and cytoplasm, it was not statistically significant (p=0.052) **B)** Agreement between predictors in relation to CFTR domains. NBD1 and NBD2 are significantly associated with full agreement between predictors. MSD1 is significantly not associated with full agreement and significantly associated with two disagreements between predictors (p<0.001). **C)** Agreement between predictors in relation to the secondary structure of CFTR. Beta-strands and bends are significantly associated with full agreement between predictors whilst alpha-helices are significantly not associated with full agreement between predictors. Overall, predictors tend to agree more when they predict missense variants as deleterious (black columns). As the disagreements increase, the neutral outcome gets more frequent (white columns). *p<0.05; †: p>0.05. MAPP and PANTHER were not included in this analysis.

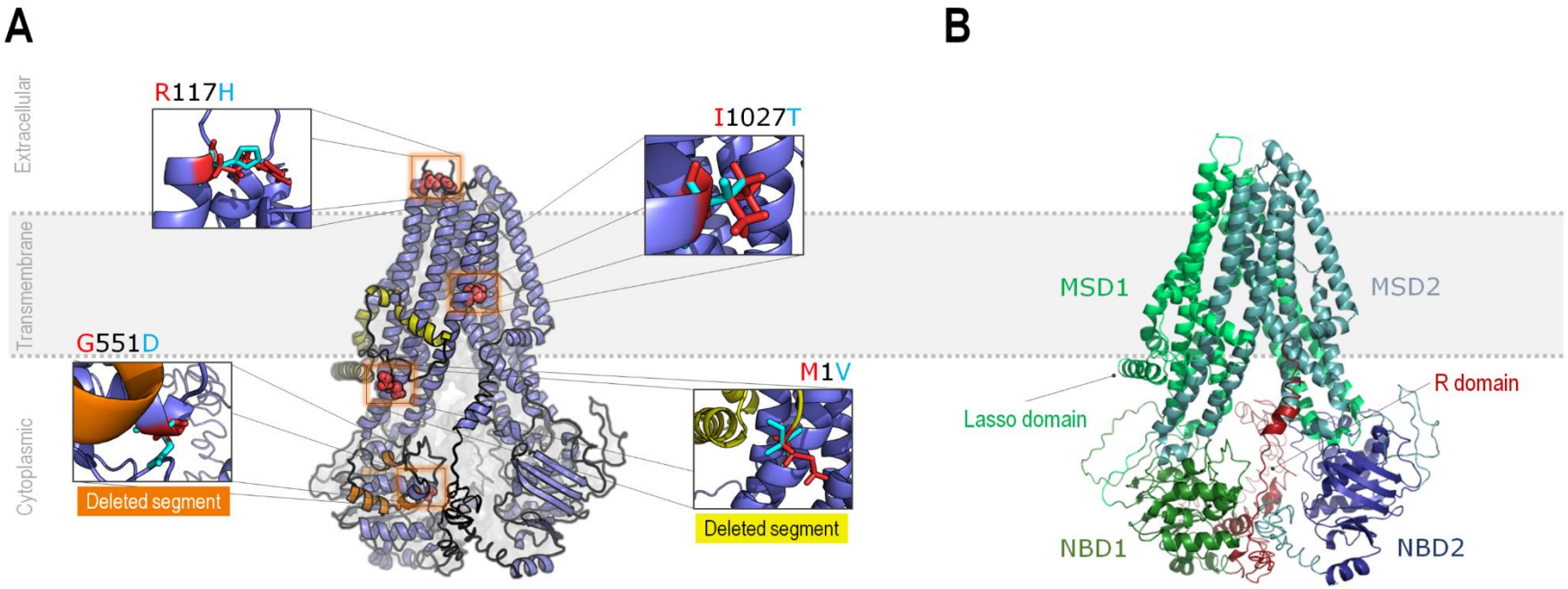


Figure 4. Modeled structure of CFTR. A) Modeled structure for CFTR in its proposed location when inserted in biomembranes (dotted lines), evidencing the absence of extracellular domains. Proposed variants are highlighted, while two segments that may be lost (at least partially) as a function of these mutations are highlighted in orange (associated with p.Met1Val) and in yellow (associated with p.Arg117His). Details regarding these deletions are given in the results session. **B)** Main structural features of human CFTR. The semi-symmetric structure is composed of two equivalent halves, comprising two domains (MSD and NBD) each. The N-terminus half of the protein has exclusive topological regions, namely the Lasso domain and the R domain (or R insertion). MSD: Membrane-spanning domain; NBD: Nucleotide-binding domain; R domain: Regulatory domain.

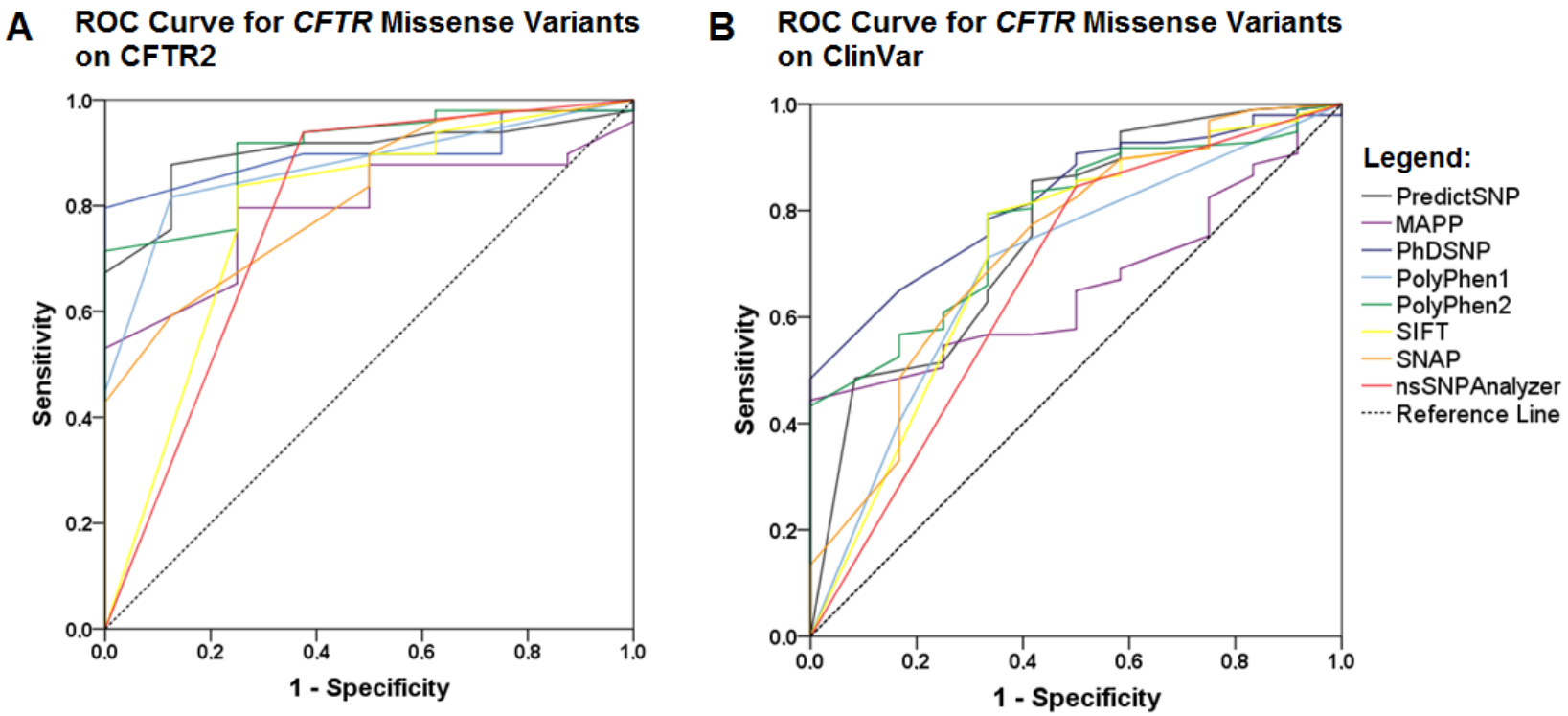


Figure 5. ROC curve for variants on CFTR2 and ClinVar. A) ROC curve for variants on CFTR2. Receiver operating characteristic (ROC) curves elaborated for missense variants according to CFTR2. n=57 (CF-causing = 49; non CF-causing = 8). **B)** ROC curve for variants on ClinVar. Receiver operating characteristic (ROC) curves elaborated for missense variants according to ClinVar. n=109 (Pathogenic/Likely pathogenic = 97; Benign/Likely benign = 12). PANTHER was not included in this analysis.

Table S1. Comparison of accuracy between predictors and variant annotation databases for *CFTR* missense variants

Comparison between predictors for *CFTR* missense variants (n=726)		
Predictor	**Median of Expected Accuracy is significantly higher than**	**Median of Expected Accuracy is significantly lower than**
PredictSNP	MAPP, PolyPhen1, PolyPhen2, SNAP, nsSNPAnalyzer	SIFT
MAPP	PolyPhen1, PolyPhen2, nsSNPAnalyzer	PredictSNP, PhDSNP
PhDSNP	MAPP, PolyPhen1, PolyPhen2, SNAP, nsSNPAnalyzer	SIFT
PolyPhen1	nsSNPAnalyzer	PredictSNP, MAPP, PhDSNP, SIFT, SNAP
PolyPhen2	nsSNPAnalyzer	PredictSNP, MAPP, PhDSNP, SIFT, SNAP
SIFT	PredictSNP, PhDSNP, PolyPhen1, PolyPhen2, nsSNPAnalyzer	
SNAP	PolyPhen1, PolyPhen2, nsSNPAnalyzer	PredictSNP, PhDSNP
nsSNPAnalyzer		PredictSNP, MAPP, PhDSNP, PolyPhen1, PolyPhen2, SIFT, SNAP
Comparison between predictors and CFTR2 for CF-causing missense variants (n=49)		
Predictor	**Median of Expected Accuracy is significantly higher than**	**Median of Expected Accuracy is significantly lower than**
PredictSNP	PolyPhen1, PolyPhen2, SIFT, nsSNPAnalyzer	
MAPP	PolyPhen1, nsSNPAnalyzer	PhDSNP
PhDSNP	MAPP, PolyPhen1, PolyPhen2, SIFT, nsSNPAnalyzer	
PolyPhen1		PredictSNP, MAPP, PhDSNP, SNAP
PolyPhen2		PredictSNP, PhDSNP
SIFT	nsSNPAnalyzer	PredictSNP, PhDSNP
SNAP	PolyPhen1, nsSNPAnalyzer	
nsSNPAnalyzer		PredictSNP, MAPP, PhDSNP, SIFT, SNAP
Comparison between predictors and ClinVar for Pathogenic/Likely pathogenic missense variants (n=97)		
Predictor	**Median of Expected Accuracy is significantly higher than**	**Median of Expected Accuracy is significantly lower than**
PredictSNP	PolyPhen1, PolyPhen2, nsSNPAnalyzer	
MAPP	PolyPhen1, nsSNPAnalyzer	PhDSNP
PhDSNP	MAPP, PolyPhen1, PolyPhen2, SIFT, SNAP, nsSNPAnalyzer	
PolyPhen1		PredictSNP, MAPP, PhDSNP, SIFT, SNAP
PolyPhen2	nsSNPAnalyzer	PredictSNP, PhDSNP
SIFT	PolyPhen1, nsSNPAnalyzer	PhDSNP
SNAP	PolyPhen1, nsSNPAnalyzer	PhDSNP
nsSNPAnalyzer		PredictSNP, MAPP, PhDSNP, PolyPhen2, SIFT, SNAP

The top of the table represents a comparison between predictors for missense variants in the *CFTR* gene(n=726). In the middle, CF-causing variants were employed in order to compare predictors to CFTR2's variant annotation (n=49). In the bottom, Pathogenic/Likely pathogenic variants were used to compare predictors to ClinVar's variant annotation (n=97). Friedman's nonparametric test was used to compare the median of accuracies. Blank spaces represent a predictor that did not statistically differ from the others. All data presented in this table are statistically significant ($p<0.05$). PANTHER was not included in this analysis.

Table S2. Predictors that significantly associate each amino acid substitution with "neutral" or "deleterious" outcome regardless of protein location

Amino acid change	PredictSNP		MAPP		PhDSNP		PolyPhen1		PolyPhen2		SIFT		SNAP		nsSNPAnalyzer		p value
	n	d	n	d	n	d	n	d	n	d	n	d	n	d	n	d	
Arg-Gly	3	12	5	9	2	13	6	9	2	13	3	12	3	12	6	9	0.4411
Asp-Asn	6	5	8	3	5	6	9	2	5	6	3	8	5	6	5	6	0.2231
Asp-Gly	5	10	9*	5	2	13	9*	6	4	11	1	14*	8	7	3	12	0.0023*
Asp-Tyr	1	9	5*	4	0	10	3	7	0	10	0	10	5*	5	1	9	0.0009*
Gln-Arg	6	4	5	5	4	6	7	3	6	4	3	7	5	5	4	6	0.7554
Glu-Lys	6	9	10	4	4	11	10	5	5	10	6	9	4	11	6	9	0.1104
Gly-Arg	3	18	4	16	0	21	4	17	7	14	4	17	3	18	6	15	0.136
Gly-Val	3	10	5	7	1	12	2	11	5	8	4	9	2	11	7	6	0.1513
Ile-Thr	6	7	7	6	6	7	5	8	6	7	6	7	5	8	5	8	0.9985
Ile-Val	13	1	7	6*	13	1	14*	0	9	5	11	3	12	2	10	4	0.0284*
Lys-Glu	6	4	3	4	2	8*	9*	1	7	3	5	5	2	8*	5	5	0.0236*
Leu-Phe	5	6	7	4	3	8	9	2	5	6	3	8	3	8	4	7	0.1072
Leu-Pro	1	25	0	25	0	26	1	25	1	25	1	25	2	24	4	22	0.309
Leu-Ser	2	12	6	8	2	12	3	11	3	11	1	13	1	13	3	11	0.3949
Met-His	6	6	7	5	4	8	11	1	5	7	5	7	5	7	6	6	0.1193
Met-Ile	6	7	7	6	2	11*	12*	1	4	9	2	11*	6	7	6	7	0.0021*
Ser-Phe	1	9	2	6	0	10	1	9	3	7	1	9	2	8	2	8	0.7009
Ser-Pro	2	9	2	9	0	11	3	8	5	6	3	8	2	9	3	8	0.4026
Thr-Ile	5	7	7	3	6	6	5	7	5	7	3	9	6	6	3	9	0.4713
Tyr-Cys	3	13	11*	5	3	13	4	12	3	13	1	15*	4	12	5	11	0.0118*

In the rows, the first amino acid represents the one in the wild-type protein sequence and the second one represents the mutated amino acid.
Columns represent the number of substitutions associated with each outcome ("neutral" or "deleterious") by each predictor "n" = neutral; "d" = deleterious.
Pearson's Chi-squared or Fisher's Exact Test was used when appropriate. *: $p<0.05$.
Associations depicted in this table represent predictors that deviate significantly from the rest of the predictors analyzed, whether polarizing towards one outcome or not polarizing its prediction when other predictors do so. PANTHER was not included in this analysis.

5. DISCUSSÃO GERAL

A fibrose cística é uma doença clinicamente heterogênea, o que pode tornar seu diagnóstico muito complicado para os clínicos. Em um pequeno número de casos onde os pacientes apresentam doença pulmonar, insuficiência pancreática, elevados níveis de cloreto no suor, infertilidade masculina, o diagnóstico de FC é evidente. Por outro lado, há aqueles pacientes que apresentam apenas um ou alguns sinais e sintomas de FC e níveis de eletrólitos no suor normais ou intermediários – usualmente chamada de FC "atípica/não-clássica". Cabe reforçar que de acordo com as diretrizes consensuais recém-publicadas pela "Cystic Fibrosis Foundation", os termos "FC clássica/típica" e "FC não clássica/atípica" devem ser evitados pelos médicos, pois não têm uma definição harmônica e podem confundir pacientes e familiares (Brennan & Schrijver 2016; Farrell et al. 2017).

Com o advento da triagem neonatal para FC, neonatos podem ser diagnosticados antes do desenvolvimento dos sintomas da doença. No entanto, a trajetória entre um resultado positivo na triagem neonatal e um diagnóstico conclusivo de FC pode não ser simples como em outras doenças monogênicas. Existem casos em que os pacientes apresentam tripsinogênio imunorreativo anormal e teste de cloreto no suor ≥60 mmol/L e que ao mesmo tempo se consegue identificar apenas uma ou mesmo nenhuma mutação causadora de FC na análise molecular do *CFTR*. Em contrapartida, há casos em que pacientes com triagem neonatal positiva para FC apresentam valores normais (≤29 mmol/L) ou limítrofes (30-59 mmol/L) de cloreto no suor, mas que de fato possuem duas mutações patogênicas identificáveis nos alelos do *CFTR*. (Farrell et al. 2017; Ratkiewicz et al. 2017). Esses exemplos representam a baixa correlação genótipo-fenótipo vista na FC, o que denota a importância de um diagnóstico que inclua dados clínicos, funcionais e moleculares para se ter um quadro completo de cada caso.

Neste contexto, pesquisas envolvendo o gene *CFTR* tem mostrado que sua análise molecular é muito complexa. Além de ser um gene muito grande (250kb), há mais de 2.000 variantes identificadas na sua estrutura (CFTR1 1989). Ademais, o significado biológico, funcional e clínico da maioria dessas variantes é de difícil estimativa. Para determinar se uma variante em particular é causadora de FC, o CFTR2 usa três critérios diferentes (CFTR2 2011):

a. **Características clínicas** de indivíduos que têm uma cópia da variante em questão e uma cópia de outra variante que sabidamente causa FC. Os níveis médios de cloreto no suor nesses pacientes devem ser acima do ponto de corte para diagnóstico de FC (≥60 mmol/L) a fim de que a variante atenda aos critérios clínicos do CFTR2 para ser causadora de FC.
b. **Análise funcional** do grau de disfunção que a variante em questão pode causar. Algumas variantes são conhecidas por terem consequências funcionais graves e não são testadas. Outras variantes requerem investigação em laboratório. O CFTR2 tem revisado pesquisas previamente publicadas, trabalhado em colaboração com a "Vertex Pharmaceuticals Incorporated" (Cambridge, MA, EUA), com pesquisadores de outras instituições, e conduzido experimentos para determinar o quanto de disfunção a variante em questão pode causar. As variantes que resultam em <10% da função normal da proteína CFTR (avaliada por enovelamento/processamento, condutância de cloreto, ou "splicing") atendem aos critérios funcionais do CFTR2 para serem causadoras de FC.
c. **Análise populacional e de penetrância** da variante em questão. Se uma dada variante atende aos critérios clínicos e funcionais para ser causadora de FC, o CFTR2 também se assegura de que: 1) a variante é menos frequente numa população saudável comparado à população de pacientes com FC, ou 2) não há evidência de que indivíduos com a variante em questão e uma variante sabidamente causadora de FC são saudáveis (se a variante é causadora de FC, espera-se que esses indivíduos tenham FC).

Assim, os pesquisadores precisam estar sempre buscando maneiras razoáveis de interpretar as variantes genéticas do *CFTR*. Um mecanismo com crescente uso na tentativa de determinar a patogenicidade de variantes no *CFTR* são os preditores computacionais (*in silico*). Muitos desses preditores tentam estimar a patogenicidade de variantes que causam a substituição de um aminoácido por outro, levando em conta a conservação evolutiva do aminoácido selvagem, a localização e o contexto da cadeia polipeptídica, e a consequência bioquímica da substituição de um aminoácido (Tavtigian et al. 2008; Hicks et al. 2011; Thusberg et al. 2011; Thompson et al. 2013; Richards et al. 2015).

Com vistas a isso, neste estudo pudemos observar que os preditores analisados apresentam concordância fraca ou moderada entre si. Além disso, eles apresentam concordância fraca a moderada com os bancos de anotação de variantes usados neste

estudo (CFTR2 e ClinVar). Coletivamente, nossos resultados apontam para o PhDSNP e o PredictSNP como as melhores ferramentas computacionais para prever a patogenicidade de variantes "missense" do gene *CFTR*. As medianas de acurácia esperada do PhDSNP e do PredictSNP não diferem entre si significativamente, e estas são relativamente elevadas em relação às medianas dos demais preditores analisados. Além disso, PhDSNP e PredictSNP parecem ter os melhores parâmetros de sensibilidade e especificidade em análise de curva ROC ("Receiver operating characteristic curves") para as variantes "missense" listadas no CFTR2 e no ClinVar. Em contrapartida, os preditores PANTHER e nsSNPAnalyzer parecem ser os menos fidedignos em estimar a patogenicidade das variantes "missense" no gene *CFTR*. O nsSNPAnalyzer não só apresenta uma mediana de acurácia esperada relativamente menor que a dos demais preditores *in silico* como também apresentou piores resultados em análise de sensibilidade e especificidade para ambos os bancos de anotação de variantes genéticas.

Um fator que pode ter contribuído para as discordâncias entre os desfechos gerados pelos preditores reside nos algoritmos usados por cada um deles (Bendl et al. 2014). Quatro dos preditores inclusos na análise do PredictSNP – PhDSNP (Capriotti et al. 2006), PolyPhen2 (Adzhubei et al. 2010), SNAP (Bromberg & Rost 2007), nsSNPAnalyzer (Chandonia et al. 2004; Bao et al. 2005) – usam um modelo de tomada de decisão baseado em vários métodos de aprendizagem de máquina (Larrañaga et al. 2006). Dos demais preditores, SIFT (Ng and Henikoff 2003) e PANTHER (Thomas et al. 2003; Brunham et al. 2005) usam somente informação evolutiva, enquanto o MAPP (Stone and Sidow 2005) também considera as diferenças nas propriedades físico-químicas entre o aminoácido selvagem e o mutado. O PolyPhen1 (Ramensky et al. 2002) utiliza um grupo de regras empíricas para a classificação das variantes "missense". Toda essa diversidade na forma com que os preditores analisam as variantes de sentido trocado indicam que eles não dever ser usados como a única forma de determinar a patogenicidade dessas mutações (Richards et al. 2015). Assim, a análise *in silico*, como parte da análise molecular do *CFTR*, deve sempre ser correlacionada a achados clínicos – sinais e sintomas – e clínico-laboratoriais para fechar o diagnóstico de FC.

Do ponto de vista estrutural, algumas informações relevantes podem ser destacadas. Com relação à concordância entre os preditores quanto aos desfechos "neutro" ou "deletério", foi possível observar que variantes "missense" localizadas nos domínios

NBD1 e NBD2, e em β-folhas estão associadas à concordância completa entre os preditores. Também, verificou-se que variantes "missense" que estão presentes no domínio MSD1 e em α-hélices estão associadas a discordância de pelo menos um preditor. O modelo 5UAK do "Protein Data Bank" – RCSB PDB (Berman et al. 2000; Liu et al. 2017) –, que foi usado como referência para determinar a estrutura secundária onde se localizam as variantes "missense" do CFTR, corrobora os dados supracitados. Além disso, a modelagem das mutações p.Met1Val, p.Arg117His, p.Gly551Asp e p.Ile1027Thr possibilitou a racionalização do efeito dessas mutações na estrutura da proteína CFTR.

Além disso, as substituições de isoleucina (I, Ile) por valina (V, Val) foram de forma geral associadas ao desfecho benigno enquanto que as substituições de leucina (L, Leu) por prolina (P, Pro) foram associadas ao desfecho patogênico. Do ponto de vista bioquímico, a troca de isoleucina por valina resulta na substituição de um resíduo não polar e alifático (cadeia lateral aberta) por outro, ambos com grupos laterais semelhantes. Já a substituição de leucina por prolina acaba provocando a troca de um resíduo não polar e alifático por um resíduo polar de cadeia lateral cíclica. De fato, a prolina é um aminoácido com características peculiares, já que um dos prótons do grupo amida é ausente, o que impossibilita o nitrogênio deste grupo amida de realizar ligações de hidrogênio (Bajaj et al. 2007). Ademais, o anel pirrolidina da prolina é rígido, constringindo o ângulo ϕ da cadeia peptídica a valores próximos de -65º. Essa alteração poderia alterar a estabilização do local onde essa mutação está inserida.

Uma grande limitação encontrada foi a falta de variantes "missense" reportadas nas bases de dados de anotação de variante; no caso do CFTR2, especialmente, apenas as 322 variantes mais comuns no *CFTR* estão presentes, sendo que apenas 74 dessas são variantes que causam a substituição de um aminoácido por outro (CFTR2 2011). Já o ClinVar recebe anotações de diversas fontes (incluindo o CFTR2), havendo 146 variantes "missense" do gene *CFTR* anotadas. Porém, isso representa menos de 20% das 779 variantes "missense" submetidas à análise por preditores computacionais.

Assim, ressalta-se a importância dos bancos de anotação de variantes, que são uma forma de disponibilização pública de dados de variantes genéticas, a fim de que mais variantes possam ser validadas. Consoante a isso, a reavaliação da interpretação de patogenicidade é fundamental. Nas listas de variantes divulgadas pelo CFTR2 anualmente existem muitos exemplos de variantes que trocaram de anotação com o tempo, dentre elas

as variantes “missense”. Além disso, o CFTR é um canal de cloreto transmembrânico, o que torna sua análise funcional muito mais difícil, denotando a importância dos bancos de anotação de variantes, da reavaliação destas ao longo do tempo e da modelagem molecular para estudos *in silico*.

Considerando o espectro variável de sinais e sintomas da fibrose cística, a relativa imprecisão do teste de cloreto no suor de certos pacientes e a dificuldade em se definir o quadro de uma parte dos pacientes, torna-se importante que o diagnóstico da FC compreenda uma análise de características clínicas, funcionais e moleculares. Com isso, criar-se-ia um cenário completo da condição de cada paciente. Isso é o que preconiza a “Cystic Fibrosis Foundation” nas recentes diretrizes publicadas (Farrell et al. 2017). Ainda, a dificuldade verificada na correlação do teste de cloreto no suor com a análise molecular do *CFTR* em certos pacientes abre um precedente para que se considere a modelagem molecular da proteína CFTR. Isso possibilitaria uma melhor estimativa do que pode estar acontecendo com o canal de cloreto desses pacientes, baseado nas mutações em seus alelos. Outrossim, faz-se necessária a busca de outros marcadores funcionais que validem o impacto fisiológico das mutações, como o teste da diferença de potencial transepitelial nasal e outros testes eletrofisiológicos, que possam contribuir para a definição do diagnóstico de FC (Nick & Nichols 2016).

Então, a aplicação em larga escala da triagem neonatal, o surgimento de novas testes fisiológicos e moleculares e a possibilidade de novas opções terapêuticas, tornam o diagnóstico precoce e definitivo da FC cada vez mais importante. E concomitante a isso, o diagnóstico clínico e o diagnóstico molecular, através da reavaliação da patogenicidade das variantes genéticas, precisam ser reavaliados periodicamente, sobretudo em casos duvidosos (Farrell et al. 2017). Por fim, a anotação das variantes genéticas do *CFTR* e sua reavaliação faz toda a diferença do ponto de vista do aconselhamento genético, pois isso é determinante para as famílias dos pacientes em termos de risco de recorrência, em termos de decisões reprodutivas e em termos de custo para a opção do diagnóstico pré-implantacional.

6. CONSIDERAÇÕES FINAIS

Este estudo possibilitou a verificação da concordância entre os diferentes preditores que fazem parte do PredictSNP no contexto de variantes “missense” no gene *CFTR*.

Observou-se que os preditores apresentam concordância fraca a moderada entre si. Além disso, os preditores apresentam também concordância fraca ou moderada com os bancos de anotação de variantes usados como parâmetro para variantes que já têm sua patogenicidade determinada.

A análise de sensibilidade e especificidade de variantes “missense” depositadas no CFTR2 e no ClinVar revelou que o PhDSNP e PredictSNP parecem ser os preditores que conseguem avaliar melhor a patogenicidade de variantes “missense” no gene *CFTR*.

Em contrapartida, o nsSNPAnalyzer não se mostrou recomendável para a avaliação dessas variantes.

Assim, o presente estudo cumpriu todos os objetivos propostos e também permitiu estabelecer outras relações importantes entre a análise dos preditores e a estrutura do canal de cloreto CFTR.

Considerando as limitações dos preditores de variantes atualmente disponíveis, a modelagem molecular foi fundamental para a racionalização de possíveis alterações estruturais causadas pelas variantes incluídas no modelo.

Os resultados deste trabalho reforçam a importância da correlação dos achados clínicos e funcionais na interpretação e validação de resultados da investigação molecular na Fibrose Cística.

REFERÊNCIAS BIBLIOGRÁFIAS

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS e Sunyaev SR (2010) A method and server for predicting damaging missense mutations. Nat Methods 7:248–249.

Bajaj K, Madhusudhan MS, Adkar B V, Chakrabarti P, Ramakrishnan C, Sali A e Varadarajan R (2007) Stereochemical Criteria for Prediction of the Effects of Proline Mutations on Protein Stability. PLoS Comput Biol 3:e241.

Bao L, Zhou M e Cui Y (2005) nsSNPAnalyzer: Identifying disease-associated nonsynonymous single nucleotide polymorphisms. Nucleic Acids Res 33:480–482.

Bendl J, Musil M, Štourač J, Zendulka J, Damborský J e Brezovský J (2016) PredictSNP2: A Unified Platform for Accurately Evaluating SNP Effects by Exploiting the Different Characteristics of Variants in Distinct Genomic Regions. PLOS Comput Biol 12:e1004962.

Bendl J, Stourac J, Salanda O, Pavelka A, Wieben ED, Zendulka J, Brezovsky J e Damborsky J (2014) PredictSNP: Robust and Accurate Consensus Classifier for Prediction of Disease-Related Mutations. PLoS Comput Biol 10:e1003440.

Bergougnoux A, Taulan-cadars M, Claustres M e Raynal C (2015) New Molecular Diagnosis Approaches — From the Identification of Mutations to their Characterization. In: Wat D (ed) Cystic Fibrosis in the Light of New Research. InTech, pp 201–231

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN e Bourne PE (2000) The protein data bank. Nucleic Acids Res 28:235–242.

Brennan ML & Schrijver I (2016) Cystic Fibrosis: A Review of Associated Phenotypes, Use of Molecular Diagnostic Approaches, Genetic Characteristics, Progress, and Dilemmas. J Mol Diagnostics 18:3–14.

Bromberg Y & Rost B (2007) SNAP: Predict effect of non-synonymous polymorphisms on function. Nucleic Acids Res 35:3823–3835.

Brunham LR, Singaraja RR, Pape TD, Kejariwal A, Thomas PD e Hayden MR (2005) Accurate prediction of the functional significance of single nucleotide polymorphisms and mutations in the ABCA1 gene. PLoS Genet 1:e83.

Callebaut I, Hoffmann B, Lehn P e Mornon JP (2017) Molecular modelling and molecular dynamics of CFTR. Cell Mol Life Sci 74:3–22.

Capriotti E, Calabrese R e Casadio R (2006) Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. Bioinformatics 22:2729–2734.

CFF (2013) Cystic Fibrosis Foundation Patient Registry. 2012 Annual Data Report. Bethesda, Maryland

CFTR1 (1989) Cystic Fibrosis Mutation Database (CFTR1), Disponível em: <http://www.genet.sickkids.on.ca/app>. Acesso em: 08/10/2016.

CFTR2 (2011) The Clinical and Functional Translation of CFTR (CFTR2). Disponível em: <http://cftr2.org>. Acesso em: 25/05/2017.

Chandonia J, Hon G, Walker NS, Lo Conte L, Koehl P, Levitt M e Brenner SE (2004) The ASTRAL Compendium in 2004. Nucleic Acids Res 32:D189–D192.

Cutting GR (2015) Cystic fibrosis genetics: from molecular understanding to clinical application. Nat Rev Genet 16:45–56.

Dal'Maso VB, Mallmann L, Siebert M, Simon L, Saraiva-Pereira ML e Dalcin P de TR (2013) Diagnostic contribution of molecular analysis of the cystic fibrosis transmembrane conductance regulator gene in patients suspected of having mild or atypical cystic fibrosis. J Bras Pneumol 39:181–189.

Fanen P, Wohlhuter-Haddad A e Hinzpeter A (2014) Genetics of cystic fibrosis: CFTR mutation classifications toward genotype-based CF therapies. Int J Biochem Cell Biol 52:94–102.

Farrell PM, White TB, Ren CL, Hempstead SE, Accurso F, Derichs N, Howenstine M, McColley SA, Rock M, Rosenfeld M et al. (2017) Diagnosis of Cystic Fibrosis: Consensus Guidelines from the Cystic Fibrosis Foundation. J Pediatr 181:S4–S15.e1.

GBEFC (2014) Grupo Brasileiro de Estudos de Fibrose Cística. Registro Brasileiro de Fibrose Cística 2014.

Groman JD, Karczeski B, Sheridan M, Robinson TE, Fallin MD e Cutting GR (2005) Phenotypic and genetic characterization of patients with features of "nonclassic" forms of cystic fibrosis. J Pediatr 146:675–680.

Hicks S, Wheeler DA, Plon SE e Kimmel M (2011) Prediction of missense mutation functionality depends on both the algorithm and sequence alignment employed. Hum Mutat 32:661–668.

Larrañaga P, Calvo B, Santana R, Bielza C, Galdiano J, Inza I, Lozano JA, Armañanzas R, Santafé G, Pérez A et al. (2006) Machine learning in bioinformatics. Brief Bioinform 7:86–112.

Liu F, Zhang Z, Csanády L, Gadsby DC e Chen J (2017) Molecular Structure of the Human CFTR Ion Channel. Cell 169:85–95.e8.

Lucarelli M, Pierandrei S, Bruno SM e Strom R (2012) The Genetics of CFTR: Genotype – Phenotype Relationship, Diagnostic Challenge and Therapeutic Implications. In: Sriramulu D (ed) Cystic Fibrosis – Renewed Hopes Through Research. InTech, pp 91–122

Martins R da S, Fonseca AC, Acosta FE, Folescu T, Higa LY, Sad I, de Miranda Chaves CR, Cabello P e Cabello GM (2014) Severe phenotype in an apparent homozygosity caused by a large deletion in the CFTR gene: a case report. BMC Res Notes 7:583.

Ng PC & Henikoff S (2003) SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res 31:3812–3814.

Nick JA & Nichols DP (2016) Diagnosis of Adult Patients with Cystic Fibrosis. Clin Chest Med 37:47–57.

Odolczyk N & Zielenkiewicz P (2014) Molecular modelling approaches for cystic fibrosis transmembrane conductance regulator studies. Int J Biochem Cell Biol 52:39–46.

Ong T, Marshall SG, Karczeski BA, Sternen DL, Cheng E e Cutting GR (2001) Cystic Fibrosis and Congenital Absence of the Vas Deferens. University of Washington, Seattle; 1993-2017, Disponível em: <https://www.ncbi.nlm.nih.gov/books/NBK1250/>. Acesso em: 25/05/2017, Seattle, WA

Pérez MM, Luna MC, Pivetta OH e Keyeux G (2007) CFTR gene analysis in Latin American CF patients: Heterogeneous origin and distribution of mutations across the continent. J Cyst Fibros 6:194–208.

Ramensky V, Bork P e Sunyaev S (2002) Human non-synonymous SNPs: server and survey. Nucleic Acids Res 30:3894–3900.

Ratkiewicz M, Pastore M, McCoy KS, Thompson R, Hayes D e Sheikh SI (2017) Role of CFTR mutation analysis in the diagnostic algorithm for cystic fibrosis. World J Pediatr 13:129–135.

Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hegde M, Lyon E, Spector E et al. (2015) Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med 17:405–423.

Saraiva-Pereira ML, Fitarelli-Kiehl M e Sanseverino MTV (2011) Genetics of Cystic Fibrosis. Rev HCPA 31:160–167.

Silva Filho LVRF, Castaños C e Ruíz HH (2016) Cystic fibrosis in Latin America—Improving the awareness. J Cyst Fibros 15:791–793.

Simon L (2010) Análise da frequência do polimorfismo (TG)m-Tn no gene CFTR em pacientes com Fibrose Cística clássica e atípica. Universidade Federal do Rio Grande do Sul.

Sosnay PR, White TB, Farrell PM, Ren CL, Derichs N, Howenstine MS, Nick JA e De Boeck K (2017) Diagnosis of Cystic Fibrosis in Nonscreened Populations. J Pediatr 181:S52–S57.e2.

Stone EA & Sidow A (2005) Physicochemical constraint violation by missense

substitutions mediates impairment of protein function and disease severity. Genome Res 15:978–986.

Tavtigian S V., Greenblatt MS, Lesueur F e Byrnes GB (2008) In silico analysis of missense substitutions using sequence-alignment based methods. Hum Mutat 29:1327–1336.

Thévenet P, Rey J, Moroy G e Tuffery P (2015) Molecular Modeling of Peptides. In: Zhou P and Huang J (eds) Computational Peptidology. Springer, New York, p pp1-14

Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, Diemer K, Muruganujan A e Narechania A (2003) PANTHER: A library of protein families and subfamilies indexed by function. Genome Res 13:2129–2141.

Thompson BA, Greenblatt MS, Vallee MP, Herkert JC, Tessereau C, Young EL, Adzhubey IA, Li B, Bell R, Feng B et al. (2013) Calibration of Multiple In Silico Tools for Predicting Pathogenicity of Mismatch Repair Gene Missense Substitutions. Hum Mutat 34:255–265. doi: 10.1002/humu.22214

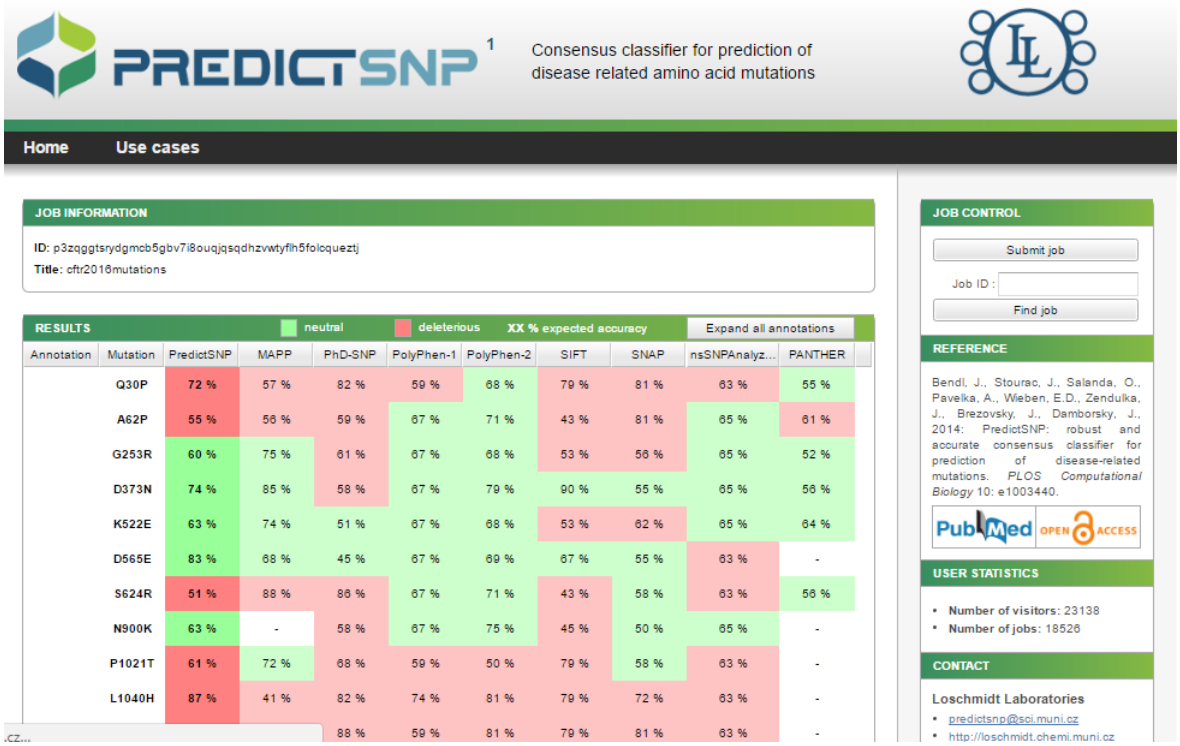
Thusberg J, Olatubosun A e Vihinen M (2011) Performance of mutation pathogenicity prediction methods on missense variants. Hum Mutat 32:358–368

Trujillano D, Ramos MD, González J, Tornador C, Sotillo F, Escaramis G, Ossowski S, Armengol L, Casals T, Estivill X et al. (2013) Next generation diagnostics of cystic fibrosis and CFTR-related disorders by targeted multiplex high-coverage resequencing of CFTR. 50:455–462.

Wang Y, Wrennall JA, Cai Z, Li H e Sheppard DN (2014) Understanding how cystic fibrosis mutations disrupt CFTR function: From single molecules to animal models. Int J Biochem Cell Biol 52:47–57.

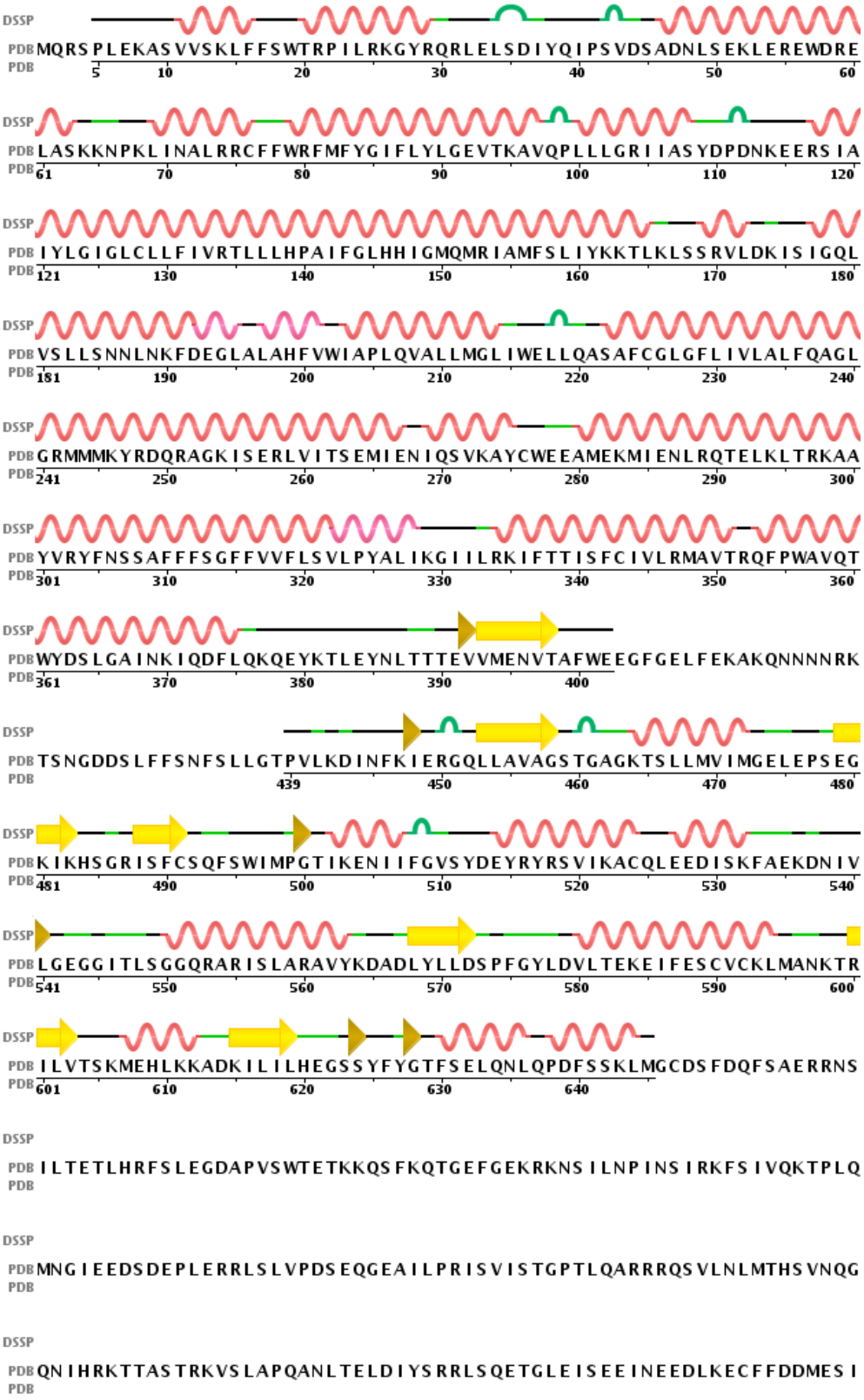
WHO Human Genetics Programme (2004) The molecular genetic epidemiology of cystic fibrosis: report of a joint meeting of WHO/IECFTN/ICF(M)A/ECFS, Genova, Itália, 19 June 2002. World Health Organization, Genebra, pp 1–24

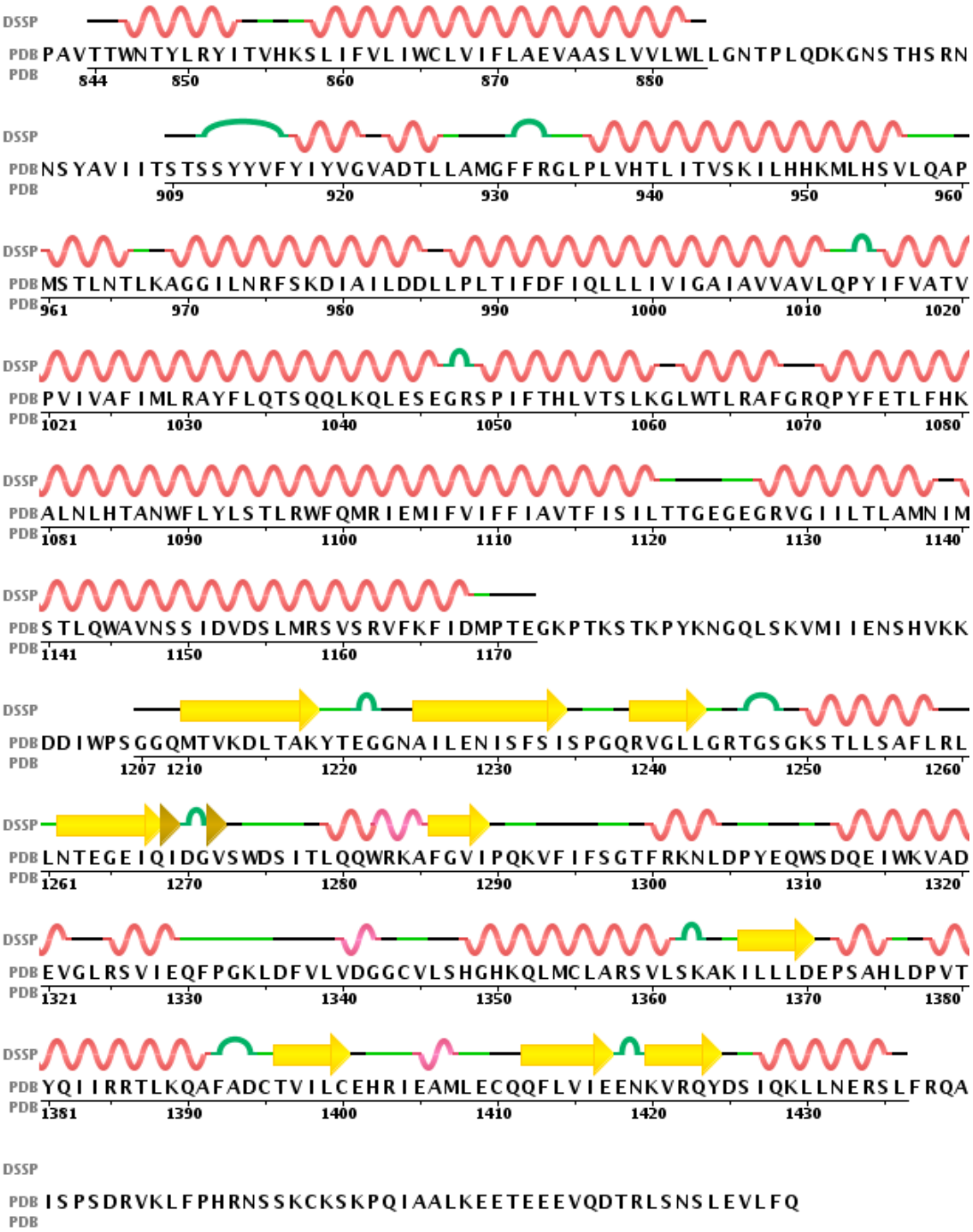
ANEXO A – INTERFACE DO PREDICTSNP E SEU *OUTPUT*



Interface do PredictSNP mostrando o resultado de análise de mutações “missense” no gene *CFTR*. Nas colunas encontram-se o PredictSNP e demais preditores. Células em verde indicam predição de mutação considerada “neutra” e células em vermelho indicam predição de mutação considerada “deletéria”. Células em branco contendo um “-“ indicam que o preditor naquela coluna não foi capaz de avaliar aquela mutação por razões metodológicas. Os valores percentuais indicam a acurácia esperada por cada preditor para o desfecho predito.

ANEXO B – ESTRUTURA SECUNDÁRIA DA PROTEÍNA CFTR





DSSP Legend

E: beta strand

T: turn

empty: no secondary structure assigned

G: 3/10-helix

B: beta bridge

S: bend

H: alpha helix

Estrutura secundária da proteína CFTR obtida a partir do modelo tridimensional gerado por Liu et al. 2017. Estrutura e modelo estão disponíveis online no “Protein Data Bank” (RCSB PDB – PDB ID: 5UAK), em: <http://www.rcsb.org/pdb/explore/remediatedSequence.do?structureId=5UAK>. Acesso em 26/03/2017.

Referência:

Liu F, Zhang Z, Csanády L, Gadsby DC, Chen J. 2017. Molecular Structure of the Human CFTR Ion Channel. Cell. 169:85-95.e8.

ANEXO C – CERTIFICADO DE APRESENTAÇÃO ORAL NO VI CONGRESSO BRASILEIRO DE FIBROSE CÍSTICA

VI CONGRESSO BRASILEIRO DE FIBROSE CÍSTICA

05 A 08 DE ABRIL DE 2017

ESTAÇÃO EVENTOS | CURITIBA | PR

CERTIFICADO

Certificamos que o trabalho

AVALIAÇÃO DE VARIANTES MISSENSE NO GENE CFTR ATRAVÉS DE PREDITORES IN SILICO

foi apresentado na modalidade Apresentação Oral, por Marcus Silva Michels, no evento VI Congresso Brasileiro de Fibrose Cística ocorrido de 05 a 08 de abril de 2017 no Estação Eventos em Curitiba/PR.
Curitiba, 08 de abril de 2017

PAULO KUSSEK
PRESIDENTE DO VI CONGRESSO BRASILEIRO DE FIBROSE CÍSTICA

JOCEMARA GURMINI
VICE-PRESIDENTE DO VI CONGRESSO BRASILEIRO DE FIBROSE CÍSTICA

CARLOS ANTONIO RIEDI
PRESIDENTE DA COMISSÃO CIENTÍFICA E TEMAS LIVRES DO VI CONGRESSO BRASILEIRO DE FIBROSE CÍSTICA

ANGELITA WISNIESKI DA SILVA
PRESIDENTE DA COMISSÃO CIENTÍFICA MULTIDISCIPLINAR DO VI CONGRESSO BRASILEIRO DE FIBROSE CÍSTICA

Realização

GBEFC GRUPO BRASILEIRO DE ESTUDOS DE FIBROSE CÍSTICA