Universidade Federal do Rio Grande do Sul
Instituto de Biociências
Curso de Biotecnologia - Ênfase em Bioinformática

Matheus Pereira Mai

Algoritmo de comparação de preditores para análises de doenças genéticas

Porto Alegre
2023

Matheus Pereira Mai

Algoritmo de comparação de preditores para análises de doenças genéticas

Trabalho de Conclusão de Curso apresentado como requisito parcial à obtenção do título de bacharel em Biotecnologia - Ênfase em Bioinformática do Instituto de Biociências da Universidade Federal do Rio Grande do Sul. Artigo formatado para o periódico *Genetics and Molecular Biology*
Orientador: Úrsula da Silveira Matte

Porto Alegre

2023

CIP - Catalogação na Publicação

Mai, Matheus Pereira
Algoritmo de comparação de preditores para análises de doenças genéticas / Matheus Pereira Mai. -- 2023.
60 f.
Orientadora: Úrsula da Silveira Matte.

Trabalho de conclusão de curso (Graduação) -- Universidade Federal do Rio Grande do Sul, Instituto de Biociências, Curso de Biotecnologia: Bioinformática, Porto Alegre, BR-RS, 2023.

1. Preditores de variantes. 2. Variantes genéticas. 3. Grupos gênicos. I. Matte, Úrsula da Silveira, orient. II. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da UFRGS com os dados fornecidos pelo(a) autor(a).

FOLHA DE APROVAÇÃO

Matheus Pereira Mai

Algoritmo de comparação de preditores para análises de doenças genéticas

Trabalho de Conclusão de Curso apresentado como requisito parcial à obtenção do título de bacharel em Biotecnologia - Ênfase em Bioinformática do Instituto de Biociências da Universidade Federal do Rio Grande do Sul. Artigo formatado para o periódico *Genetics and Molecular Biology*
Orientador: Úrsula da Silveira Matte

Porto Alegre, 12 de setembro de 2023.

BANCA EXAMINADORA:

_____
Dra. Úrsula da Silveira Matte
Orientadora
Universidade Federal do Rio Grande do Sul (UFRGS)

_____
Dra. Thayne Woycinck Kowalski
Universidade Federal do Rio Grande do Sul (UFRGS)

_____
Dr. Nureyev Ferreira Rodrigues
Universidade Federal do Rio Grande do Sul (UFRGS)

RESUMO

A análise molecular desempenha um papel fundamental no diagnóstico de doenças genéticas, pois caracteriza a relação entre variação de sequência e predisposição à ou surgimento da doença. Quando necessário, programas computacionais são utilizados para predizer o significado de determinada variação genética. Este trabalho apresenta uma análise sobre a relação entre a qualidade de assertividade de predição em diferentes grupos gênicos. Foi comparado o desempenho de 38 preditores de variantes em 39 genes, os quais foram agrupados em 8 categorias. Os dados foram obtidos do banco de dados *dbNSFP* e *ClinVar* para previsões e classificações e as classes de genes foram definidas no *Protein Class* do *PANTHER*. Nossos resultados mostraram que *ClinPred, BayesDel_addAF* e *Meta_RNN* superam consistentemente outros preditores, embora existam diferenças em classes genéticas específicas. Isso destaca a importância de avaliar qual preditor apresenta os melhores resultados caso a caso com um conjunto de dados selecionados.

Palavras chaves: Preditores de variantes. Variantes genéticas. Grupos gênicos.

ABSTRACT

Molecular analysis plays a key role in the diagnosis of genetic diseases, as it characterizes the relationship between sequence variation and predisposition to or onset of the disease. Computer programs are often used to predict the significance of a given genetic variation. This work presents an analysis of the relationship between the quality of prediction assertiveness in different gene groups. The performance of 38 predictors of variants in 39 genes was compared, which were grouped into 8 categories. Data were obtained from the dbNSFP and ClinVar database for predictions and classifications and the gene classes were defined in PANTHER's Protein Class. Our results showed that ClinPred, BayesDel_addAF and Meta_RNN consistently outperform other predictors despite differences in specific genetic classes. This highlights the importance of evaluating which predictor gives the best results on a case-by-case basis with a selected dataset.

Keywords: Variant Predictors. Genetic variants. Gene groups.

LISTA DE FIGURAS

LISTA DE ABREVIATURAS E SIGLAS

ACMG - *American College of Medical Genetics and Genomics*

AMP - *Association for Molecular Patology*

dbNSFP - Database for Non-Synonymous Functional Predictions

dbSNP - Database of Single Nucleotide Polymorphisms

EMMs - médias marginais estimadas

GM12878 - células linfoblastóides

HGVS - Human Genome Variation Society

H1 hESCs - células-tronco embrionárias humanas H1

mRNA – Ácido ribonucleico mensageiro

nsSNVs - variantes não-sinônimas de nucleotídeos únicos

PANTHER - Protein ANalysis Through Evolutionary Relationships

PC – Protein Class

SIFT - Sorting Intolerant From Tolerant

SSNVs - splicing-site *SNVs*

VCF - *Variant Call Format*

VEP - *Ensembl Variant Effect Pedictor*

VUS - *variant uncertain significance*

WGS – *whole genome sequencing*

SUMÁRIO

1. INTRODUÇÃO

1.1. ANÁLISE MOLECULAR DE DOENÇAS GENÉTICAS

A análise molecular desempenha um papel fundamental no diagnóstico de doenças genéticas, pois caracteriza a relação entre variação de sequência e predisposição à ou surgimento da doença, o que fornece uma ferramenta poderosa para identificar processos fundamentais para a patogênese e aplicar ou desenvolver novas estratégias para prevenção e tratamento (CLAUSSNITZER et al., 2020). O uso generalizado de *whole genome sequencing* (WGS) permite a detecção de uma ampla gama de variantes genéticas comuns e raras de diferentes tipos em quase todo o genoma, o que facilita a pesquisa de doenças raras e aplicações clínicas além de auxiliar na descoberta e anotação de variantes causais de doenças comuns (LAPPALAINEN et al., 2019). Nesse contexto, avanços tecnológicos tanto em métodos de bioinformática quanto em testes analíticos - que foram proporcionados por grandes projetos - (AUTON et al., 2015; LEK et al., 2016) possibilitaram a identificação de diversos genes causadores de doenças e também a elucidação de vias biológicas envolvidas no processo. Tais acontecimentos influenciam positivamente no desenvolvimento de terapias e diagnóstico de doenças o que contribui para a medicina de precisão e medicina genômica.

1.2. CLASSIFICAÇÃO DE VARIANTES GENÉTICAS

A análise molecular para diagnóstico de doenças envolve um conjunto de técnicas que objetivam encontrar marcadores biológicos em genomas e ou proteomas (CHOE et al., 2015). Uma das etapas do processo da análise e diagnóstico molecular é a classificação de variantes, considerada a pedra angular dos testes clínicos genéticos moleculares (NYKAMP et al., 2017). Sua importância se dá pelo fato de que a classificação, de maneira geral, consegue diferenciar uma variante quanto a sua importância em determinado diagnóstico, variando desde aquelas em que a variante é quase certamente patogênica para um distúrbio, até aquelas que são quase certamente benignas (RICHARDS et al., 2015).

Conforme as tecnologias de sequenciamento avançaram, as classificações de variantes também foram sendo aprimoradas com o intuito de viabilizar uma

classificação adequada utilizando padrões estipulados. Nesse contexto, são empregadas ferramentas e diretrizes específicas para auxiliar na classificação das variantes em contexto diagnóstico. Para seguir essas diretrizes, atualmente são utilizados dois principais protocolos, o do *American College of Medical Genetics and Genomics* (ACMG) juntamente com a *Association for Molecular Patology* (AMP) (RICHARDS et al., 2015) e o protocolo *Sherloc* que é uma especie de refinamento do protocolo supracitado (NYKAMP et al., 2017). A exemplo, o protocolo ACMG-AMP recomenda o uso de terminologia padrão específica – “patogênico”, “provavelmente patogênico”, “significado incerto”, “provavelmente benigno” e “benigno” – para descrever variantes (RICHARDS et al., 2015). Em contrapartida, ao verificar que algumas classificações se apresentam de maneira ambígua ou contraditória, o Sherloc refinou alguns passos, principalmente no sistema de métricas para computar a classificação (NYKAMP et al., 2017). Portanto, ambos protocolos apresentam certa divergência na classificação.

Essa colaboração entre especialistas tem permitido a disseminação do conhecimento e a melhoria contínua dos métodos de classificação, tornando possível uma abordagem mais robusta e precisa na interpretação clínica das variantes genéticas.

1.3. VARIAÇÕES *MISSENSE* E VARIANTES DE SIGNIFICADO INCERTO (VUS)

As variações que mais apresentam desafios para a classificação são as variantes *missense*. Variantes *missense* consistem em modificações pontuais na sequência nucleotídica, onde acontece a alteração de um único nucleotídeo que leva a mudança do aminoácido correspondente na proteína. Essa modificação não pode ser considerada diretamente patogênica (NYKAMP et al., 2017), diferentemente do que acontece com outras variantes, como por exemplo as *nonsense* pois a existência de um códon de parada prematuro geralmente resulta em proteínas truncadas e rapidamente degradadas (CASTIGLIA; ZAMBRUNO, 2010). Assim como as *nonsense*, as variantes de *frameshift* também apresentam um certo consenso sobre sua classificação devido ao impacto da variação na funcionalidade da proteína, entretanto, não são tratadas como perda de função quando ocorrem no último exon (LINDEBOOM; SUPEK; LEHNER, 2016).

Já as variações missense apresentam, muitas vezes, uma dificuldade na sua classificação, e então podem vir a ser elencadas como variantes de significado incerto ou VUS (do inglês *variant of uncertain significance).* Idealmente, o impacto funcional e consequentemente a classificação das *VUS* deveriam ser validadas por meio de estudos experimentais, porém, considerando a gigantesca gama de dados gerados a todo momento e que cada genoma pode gerar um arquivo *Variant Call Format (VCF)* de 125 MB, com 3 milhões de variantes cada, é impossível que a análise experimental abranja a descoberta e anotação de todas as novas variantes (WONG et al., 2019).

Nesses casos, uma importante ferramenta, fortemente utilizada, e indicada pelos protocolos ACMG - AMP e Sherloc, são as avaliações *in silico,* que, apesar de não apresentarem validação clínica, são utilizadas para formar um certo nível de evidência (NYKAMP et al., 2017; RICHARDS et al., 2015). Em certos casos, são utilizados até mesmo como única ferramenta para predizer a classificação de variantes genéticas (BORGES, 2021).

1.4. PREDITORES DE VARIANTES GENÉTICAS

Para obter uma classificação confiável e robusta das *VUS,* o uso de avaliações *in silico* ganhou e vem ganhando destaque. Essas ferramentas computacionais utilizam algoritmos e dados provenientes de bioinformática para predizer os possíveis efeitos funcionais das variações. Os preditores de variantes genéticas são, em sua maioria, construídos e baseados nos possíveis efeitos gerados por cada mutação, considerando fatores como conservação de aminoácidos e nucleotídeos, local e importância estrutural da alteração e fatores bioquímicos (TANG; THOMAS, 2016).

Ferramentas para a avaliação de variantes *missense* existem desde 1970 e a quantidade de programas que fazem a predição dessas variantes, a partir dos anos 2000, é crescente. Os primeiros preditores *Sorting Intolerant From Tolerant* (SIFT) (NG, 2003), PolyPhen (SUNYAEV, 2001) e PANTHER (THOMAS et al., 2003) tinham seus algoritmos baseados em alinhamentos e a divergência entre eles era especialmente em matrizes de pontuações internas e de probabilidades (BORGES, 2021). Depois desses iniciais, preditores baseados em informações estruturais também foram criados, por exemplo o MAPP (STONE; SIDOW, 2005), porém dependiam de estrutura tridimensional de proteínas, que era um dado escasso na época. Com o decorrer dos anos surgiram preditores com métodos que

compartilhavam informações de alinhamento e estrutural, e preditores que utilizavam técnicas de *machine learning* utilizando grupos de treinamento e considerando métodos de combinação (BORGES, 2021).

Atualmente existem diversos preditores de variantes genéticas que utilizam diferentes técnicas para classificar as *VUS,* o que levanta um tópico de importante discussão: “qual preditor eu devo utilizar?”. Ainda não existe um protocolo que indique quais programas são ideais para determinadas proteínas, o que gera uma escolha de usabilidade de maneira aleatória ou por notabilidade. Avaliando por citações na literatura, os preditores PolyPhen e SIFT apresentam maior número de usuários e, embora não se possa dizer que sejam os mais utilizados na prática clínica, de fato, frequentemente relatórios clínicos mencionam esses preditores (BORGES, 2021). Um aspecto importante que contribui para que alguns preditores sejam mais usados do que outros é a recomendação do ACMG-AMP (RICHARDS et al., 2015).

Além disso, as performances dos preditores variam amplamente de acordo com a sequência proteica avaliada (RICHARDS et al., 2015). Logo, a escolha do método utilizado varia de acordo com o problema a ser analisado (UÇAR et al., 2020). Em consequência da utilização de diferentes preditores genéticos, a divergência de resultados é frequente, e em casos de disparidade, deve-se utilizar alguma das estratégias disponíveis para ajustar os dados não balanceados (UÇAR et al., 2020).

1.5. GRUPOS GÊNICOS E AGRUPAMENTO VIA *PANTHER*

Os grupos de genes, também conhecidos como famílias de genes, desempenham um papel crucial na compreensão da evolução e função dos elementos genéticos. São conjuntos de vários genes semelhantes, formados pela duplicação de um único gene original e geralmente com funções bioquímicas semelhantes. As proteínas podem ser classificadas ao longo de dois eixos primários: agrupamentos evolutivos (classe de proteína, família de proteína, subfamília) e agrupamentos funcionais (ontologia genética e vias) (MI et al., 2021). O agrupamento de genes através de homologia de sequências permite identificar genes com relação evolutiva, e, utilizando ferramentas de bioinformática e filogenia, é possível inferir suas funções biológicas.

O *PANTHER* (*Protein ANalysis Through Evolutionary Relationships*) é uma ferramenta computacional utilizada para classificar genes em grupos evolutivos e funcionalmente relacionados (THOMAS et al., 2003). A ferramenta oferece uma

plataforma online que utiliza uma grande quantidade de dados genômicos e filogenéticos para classificar genes em famílias e suas classes proteicas. A *Protein Class (PC)* foi projetada como uma classificação simples e de alto nível das funções de proteínas e famílias de proteínas. Além disso, as classificações de *PC* são mais fáceis de navegar e interpretar do que outras classificações no *PANTHER*, como *Gene Ontology* (MI et al., 2021).

Portanto, utilizando o *Panther Class* temos uma classificação para cada gene, que agrupa os semelhantes em uma única classe, o que é muito importante para uma possível validação e análise de preditores frente aos mesmos.

1.6. *dbNSFP*, *CLINVAR* E A AVALIAÇÃO DE PREDITORES

Atualmente existe um banco de dados chamado dbNSFP que foi desenvolvido para previsão funcional e anotação de todas as potenciais variantes não-sinônimas de nucleotídeos únicos (nsSNVs) no genoma humano. Sua versão atual é baseada na versão gencode 29 / Ensembl versão 94 e inclui um total de 84.013.490 nsSNVs e SSNVs (*splicing-site* SNVs). Ele compila pontuações de previsão de 38 algoritmos de predição além de pontuações de conservação e outras informações relacionadas como frequências alélicas (LIU et al., 2020). Para avaliar os preditores também é necessário dados referentes às amostras de pacientes, ou seja, dados curados. Nesse caso existe o banco de dados ClinVar que é um arquivo público livremente disponível de variantes genéticas humanas e interpretações de suas relações com doenças e outras condições, mantidos nos Institutos Nacionais de Saúde (NIH) (LANDRUM et al., 2020).

Sabendo que as escolhas de preditores geralmente ocorrem sem a utilização de critérios específicos, uma estratégia para fazer uma escolha mais objetiva e aumentar a confiabilidade da análise *in silico* é avaliar o desempenho de cada preditor para cada gene individualmente. Assim saberíamos se os critérios empregados na construção dos preditores são igualmente relevantes para todos os genes (BORGES, 2021). Nesse contexto, criamos uma ferramenta que aplica estatísticas para avaliar o desempenho de cada preditor frente a cada grupo de genes - utilizando o *Protein Class* do *PANTHER* - para verificar se existe uma real divergência dos resultados.

2. MATERIAL E MÉTODOS

2.1.1. *Download* dos dados do *dbNSFP*

O *dbNSFP* (Database for Non-Synonymous Functional Predictions) é um banco de dados que foi desenvolvido para previsão funcional e anotação de todas as potenciais variantes não-sinônimas de nucleotídeos únicos (*nsSNVs*) no genoma humano. Sua versão atual é baseada na versão *gencode* 29 / *Ensembl* versão 94 e inclui um total de 84.013.490 *nsSNVs* e *SSNVs* (splicing-site *SNVs*). Ele compila pontuações de previsão de mais de 38 algoritmos de predição além de pontuações de conservação e outras informações relacionadas como frequências alélicas (LIU et al., 2020). Essas informações são excelentes para fazer uma análise de preditores frente a variações genéticas, portanto, nesse trabalho utilizamos essa *database*, em sua versão *v4.2a*.

O processo de obtenção da versão *v4.2a* do *dbNSFP* foi realizado através do site https://sites.google.com/site/jpopgen/dbNSFP, fizemos o download da versão *v4.2a* de 06/04/2021 através do link https://drive.google.com/file/d/1OfAx1SrJVPPNYWfDpQd43S9Aqro3C6aA/view. A *database* foi baixada em formato .zip de aproximadamente 32 *gigabytes*. Após a descompactação, foram resgatados os arquivos de formato *.tsv* referentes a cada cromossomo do genoma humano, os quais foram posteriormente utilizados.

2.1.2. Seleção dos dados

A versão *v4.2a* do *dbNSFP* apresenta, em cada arquivo referente ao cromossomo, colunas com informações sobre a variante como: cromossomo, posição da variante no cromossomo, nucleotídeo de referência e o alterado, o aminoácido referência e o alterado, *rs_dbSNP* (que é um identificador atribuído a uma variante genética específica catalogada no *dbSNP* (Database of Single Nucleotide Polymorphisms) (SHERRY, 2001), o nome do gene em que está localizada a variante, anotações das variantes a nível de mRNA e a nível proteico, utilizando a nomenclatura do Human Genome Variation Society (HGVS) (GULLEY et al., 2007). Para essa anotação são utilizadas as geradas por dois programas, o *Ensembl Variant Effect Pedictor (VEP)* (MCLAREN et al., 2016) e o *ANNOVAR* (WANG; LI; HAKONARSON, 2010) (Figura 1).

chr	pos(1-based)	ref	alt	aaref	aaalt	rs_dbSNP	genename	HGVSc_ANNOVAR	HGVSp_ANNOVAR	HGVSp_VEP
1	93996138.0	G	A	R	X	rs998363634	ABCA4	c.3163C>T	p.R1055X	p.Arg1055Ter
1	93996161.0	C	A	S	I	rs6666652	ABCA4	c.3140G>T	p.S1047I	p.Ser1047Ile
1	93997877.0	T	C	Q	R	rs886044764	ABCA4	c.3089A>G	p.Q1030R	p.Gln1030Arg

Figura 1 - Exemplo de informações das variantes no dbNSFP. No exemplo, temos 3 variações para o gene *ABCA4.*

Além das informações referentes às localizações e anotações das variantes, o *dbNSFP* apresenta colunas referentes a classificação das mesmas, dada pelos preditores de variantes, como *SIFT*, *SIFT4G*, *PolyPhen2*, *MutationTaster*, etc. Também possui colunas referentes a frequências alélicas como as do banco de dados *gnomAD* (CHEN et al., [s.d.]) e do *1000 Genomes Project Phase 3 (1000Gp3)* (FAIRLEY et al., 2020). O banco também contém informações das variantes provenientes do *ClinVar* que é um arquivo público livremente disponível de variantes genéticas humanas e interpretações de suas relações com doenças e outras condições, mantidos nos Institutos Nacionais de Saúde (*NIH*) (LANDRUM et al., 2020).

No contexto desse trabalho, não foram utilizadas todas as colunas e informações provenientes do *dbNSFP*. Portanto, utilizamos as informações das variantes, a classificação de 38 preditores, SIFT_pred, SIFT4G_pred, Polyphen2_HDIV_pred, Polyphen2_HVAR_pred, LRT_pred, MutationTaster_pred, MutationAssessor_pred, FATHMM_pred, PROVEAN_pred, VEST4_score, MetaSVM_pred, MetaLR_pred, MetaRNN_pred, M-CAP_pred, REVEL_score, MutPred_score, MVP_score, MPC_score, PrimateAI_pred, DEOGEN2_pred, BayesDel_addAF_pred, BayesDel_noAF_pred, ClinPred_pred, LIST-S2_pred, CADD_phred, CADD_phred_hg19, DANN_score, fathmm-MKL_coding_pred, fathmm-XF_coding_pred, Eigen-phred_coding, Eigen-PC-phred_coding, GenoCanyon_score, integrated_fitCons_score, GM12878_fitCons_score, H1-hESC_fitCons_score, HUVEC_fitCons_score, LINSIGHT e as classificações do *ClinVar.* Devido ao grande tamanho de memória ocupado pelos arquivos *.tsv* do banco de dados, impossibilitando a utilização de determinadas ferramentas em computador pessoal, o resgate das informações foi feito via *bash Linux*.

2.1.3. Verificação dos genes que mais apresentam variantes no *CLINVAR* e seleção no *dbNSFP*

Ao fazer a análise do banco de dados verificou-se a impossibilidade de executar o algoritmo para todo o genoma, visto que muitos genes apresentam poucas variantes para serem analisadas e comparadas com um resultado padrão (no nosso caso o *ClinVar*) e isso geraria estatísticas sem a devida confiabilidade. Para isso, como o banco *dbNSFP* já inclui dados da relação das variantes com doenças genéticas e outras condições, além de classificações computadas do banco *ClinVar*, optou-se pela busca de genes com maior quantidade de variantes classificadas neste mesmo banco.

Para realizar isso, navegamos pelo site oficial do *ClinVar* (https://www.ncbi.nlm.nih.gov/clinvar/) e fizemos o *download* dos dados via *FTP*. Usamos o arquivo de formato *.vcf* e, utilizando a linguagem de programação *python* na versão 3.8, analisamos o arquivo e computamos todos os genes que apresentavam mais de 100 variantes computadas como "*Pathogenic", "Likely Pathogenic", "Benign"* e *"Likely Pathogenic"* somadas. Com isso, selecionamos 149 genes e então voltamos para os dados do *dbNSFP* para resgatar todas as variações e predições para esses genes em específico.

Para realizar essa seleção dos genes, novamente, devido ao grande tamanho de memória ocupado pelos arquivos *.tsv* do banco de dados, impossibilitando a utilização de determinadas ferramentas em computador pessoal, utilizamos o *bash Linux* para retornar arquivos individuais para cada um dos genes. Desse modo, posteriormente, podemos aplicar o algoritmo de avaliação com maior facilidade.

A partir dessa seleção de genes, realizamos mais uma verificação, como o banco *dbNSFP* pode não conter todas as classificações correntes do *ClinVar* que avaliamos devido às versões incompatíveis e/ou outras ocorrências, fizemos a verificação da quantidade de variantes desses genes mencionados, no *dbNSFP*. Ou seja, resgatamos todas as informações dos 149 genes correntes no *dbNSFP* e realizamos a contagem de suas variantes "*pathogenic*" e "*benign*". Com esse corte, tivemos uma queda de genes para serem avaliados de 149 para 86.

2.1.4. Limpeza e tratamento dos dados

Ao passo que resgatamos as informações de cada gene desejado, ainda precisamos avaliar todos os dados das tabelas, para remover dados indesejados,

corrigir valores faltantes e tratar valores inconsistentes para uma melhor confiabilidade dos dados. Uma das coisas importantes que foram tratadas era a quantidade de valores inexistentes nas tabelas que eram marcadas com um ponto final (‘ . ‘), o que atrapalhava na posterior aplicação dos cálculos desejados pois apresentava conflito de *typing*.

Como o *dbNSFP* apresenta a classificação dada pelos preditores compilando os valores de cada um dos programas, para uma avaliação efetiva, devemos padronizar os *scores*. Portanto, para as classificações, foram atribuídos valores binários, 1 e 0, sendo 1 = “*pathogenic”* e 0 = “*benign*”, seguindo o funcionamento de cada *score* dos programas. Exemplificando, para o preditor *SIFT* o valor “D” significa que é *pathogenic*, logo, receberá valor 1, porém, se o valor for “T” significa que é *benign*, logo, receberá valor 0. E assim foi feito para cada um dos preditores, utilizando os *scores* internos de cada programa, provenientes da própria documentação do *dbNSFP* e/ou da publicação realizada.

Além dos preditores, esse mesmo processo foi realizado para a classificação dada pelo *ClinVar*. No caso desse banco de dados, foi aplicado o seguinte conceito para pontuar em binários: se o valor for “*drug response*”, “*likely pathogenic*”, “*risk factor*”, ou esses valores somados, era considerado “*pathogenic*”, logo, recebe valor 1; se o valor for “*benign*”, “*likely benign”*, “*protective*”, ou esses valores somados, era considerado “*benign*”, logo, recebe valor 0; qualquer outro valor ou, se existisse uma soma de valores contraditórios para a mesma variante, por exemplo, “*pathogenic and likely benign*”, era considerado como incerto, logo, não recebia pontuação e era removido o dado, visto que não pode ser computado para as avaliações posteriores.

2.1.5. Criação e aplicação do algoritmo que mede sensibilidade, especificidade, acurácia e valor de *kappa*

Para fazer uma avaliação dos preditores frente a cada gene, foi desenvolvido um algoritmo em linguagem *python* na versão 3.8 que calcula métricas cruciais de avaliação, a sensibilidade, especificidade, acurácia e valor de kappa, proporcionando uma análise abrangente da eficácia dos preditores. Os cálculos de sensibilidade, especificidade e acurácia foram feitos utilizando uma lógica própria, já o teste de *kappa* foi realizado utilizando a biblioteca *scikit-learn*, com o método *cohen_kappa_score*.

Para trabalhar com os dados, foi utilizado a biblioteca *pandas* e para manipular os diretórios utilizou-se a biblioteca *os*. Para a realização dos cálculos, era necessário utilizar uma classificação como "padrão de referência", e, como o *dbNSFP* já apresenta a classificação do *ClinVar*, ele foi utilizado como esse tipo de modelo, visto que geralmente as classificações mantidas neste banco são, de certa forma mais curadas.

Todo o algoritmo foi construído no formato de funções, sem a utilização de programação orientada a criação de objetos.

2.1.6. Classificação dos genes resgatados via *PANTHER Protein Class*

Como o objetivo do presente trabalho é avaliar os preditores frente a grupos gênicos, foi necessário buscar uma forma de agrupar cada gene em classes coesas. Para isso, optou-se por agrupá-los através de classes proteicas, fazendo então a utilização do *Protein ANalysis Through Evolutionary Relationships* (PANTHER). O PANTHER que é uma plataforma online que utiliza uma grande quantidade de dados genômicos e filogenéticos para classificar genes em famílias e suas classes proteicas, apresenta uma classificação de genes por classes proteicas, o *Protein Class* (*PC*) que é mais fácil de navegar e interpretar do que, por exemplo, agrupar por *Gene Ontology*.

Portanto, agrupamos os genes escolhidos em classes seguindo a classificação dada pelo *PC*. Por exemplo, o gene *FBN1* (fibrilina) apresenta um *PC* de *extracellular matrix structural protein*, então, todos os outros genes selecionados que apresentaram o mesmo *PC,* serão agrupados juntos. Dessa forma, podemos realizar a análise dos preditores frente à grupos gênicos.

Após agruparmos esses genes, verificamos a quantidade de classes que poderiam ser analisadas. Para isso, adicionamos um *cutoff* de no mínimo 3 genes presentes por classes. Com esse *cutoff* teríamos poucos dados a serem avaliados, tornando irrelevante a aplicação estatística. Portanto, optamos por fazer agrupamentos menores em classes relacionadas.

2.1.7. Aplicação de estatísticas via linguagem *R*

Para analisar as relações entre preditores de variantes genéticas e grupos de genes com base na *PC*, usando o índice de *Youden* como variável de resposta. A

análise foi realizada utilizando a linguagem de programação R. Na aplicação das estatísticas agrupamos as informações em uma tabela com as seguintes colunas: *programs* contendo o nome de cada preditor avaliado, no nosso caso, 38 preditores; *PANTHER_CLASS* contendo o valor *PC* sendo as classes dos genes; *Youden*, contendo o valor do cálculo de *Youden* para cada preditor (*(sensibilidade + especificidade) - 1*); *genename* contendo o nome de cada gene que foi avaliado.

Inicialmente, ajustamos um modelo de regressão linear, utilizando a função *lm()*, no qual consideramos a variável resposta *Youden* em relação às variáveis preditoras *PANTHER_CLASS* e *programs*. Isso nos permitiu modelar a relação entre os preditores e a métrica *Youden*. Em seguida, o pacote *emmeans* foi utilizado para calcular as médias marginais estimadas (EMMs). Isso envolvia calcular as médias do índice de Youden em diferentes combinações de *PANTHER_CLASS* e *programas*. Esses EMMs representam o efeito médio de cada grupo na variável de resposta, onde o objetivo é realizar uma análise das médias dos efeitos estimados, com foco na interação entre as variáveis *programs* e *PANTHER_CLASS*. Essa etapa proporcionou uma visão mais aprofundada das diferenças nas médias entre os grupos gênicos e os preditores de variantes genéticas. Para efetuar comparações múltiplas entre as médias, aplicamos os procedimentos de ajuste de p-valor usando o método de *Benjamini-Hochberg* (adjust = "BH"). Isso nos permitiu controlar o erro global ao realizar várias comparações, ajudando a identificar associações estatisticamente significativas. Para simplificar a interpretação das diferenças entre pares os resultados foram visualizados e interpretados através da função *cld()*, onde foram gerados agrupamentos baseados em letras. Esta função atribui letras a grupos que não são significativamente diferentes uns dos outros, ajudando a identificar quais preditores e grupos gênicos apresentam diferenças estatisticamente significativas entre si, ou seja, o resultado ajuda a identificar grupos com efeitos semelhantes no índice de *Youden*.

Esse resultado foi atribuído então em uma tabela com as seguintes colunas: *programs*, contendo os nomes dos preditores de variantes; *PANTHER_CLASS*, contendo a classificação *PC*; *emmeans*, sendo os valores de médias marginais estimadas; *lower.CL* e *upper.CL*, sendo os intervalos de confiança; *.group*, sendo as letras que vão identificar os grupos e preditores com diferenças estatísticas entre si. Dessa maneira, foi possível fazer uma avaliação minuciosa do resultado.

3. RESULTADOS

3.1.1. Genes utilizados

Conforme mencionado, foram realizados três *cutoffs* de genes visando a melhor otimização das estatísticas. Primeiramente utilizamos o *ClinVar* e obtivemos 149 genes com mais de 100 variantes "*pathogenic*" e "*benign*" somadas. A partir desses genes, foi verificado quantas variantes a coluna *ClinVar* presente no *dbNSFP* apresentava. Nesse contexto, 86 genes retornaram uma soma de variantes maior que 100, portanto, nossa amostra reduziu. Como o objetivo do presente trabalho era avaliar os preditores frente a grupos gênicos, foi realizado a divisão dos genes em classes Panther, e, nesse contexto, optamos por utilizar apenas classes que apresentam 3 ou mais genes, portanto, nossa amostra reduziu novamente, de 86 genes, passamos para 39 (*FBN1, BRCA1, DMD, COL4A5, SCN1A, USH2A, COL3A1, ABCA4, COL1A1, COL1A2, SCN2A, MECP2, SCN5A, CDKL5, KCNH2, ABCC6, NIPBL, DNAH5, BMPR2, COL2A1, COL7A1, DNAH11, SCN8A, CACNA1A, COL4A3, ACVRL1, DYNC2H1, SPAST, COL6A3, BRAF, GRIN2B, VWF, COL4A4, CHD7, ABCB1, CACNA1H, GRIN2A, VHL, KCNT1, MBD5 e FBN2*). Desses genes selecionados o de maior quantidade de variantes foi *FBN1* com 1298 variantes, dessas, 1283 foram consideradas "*pathogenic*" e 15 "*benign*". Já o gene com menos variantes foi *FBN2* com 101 variantes, sendo, 51 consideradas "*pathogenic*" e 50 "*benign*" (Figura 2).

Genename	Qnt_patogenics	Qnt_benign	Total_variantes
FNB1	1283	15	1298
FBN2	51	50	101

Figura 2 - Genes com mais e menos variantes catalogadas no dbNSFP usados na nossa análise

3.1.2. Agrupamentos

Como já mencionado, foram feitos agrupamentos por classes gênicas utilizando *PC* do *Panther* e, posteriormente, utilizamos agrupamentos em classes menores relacionadas entre si. Nesse contexto, as modificadas foram: a classe *extracellular matrix structural protein* virou *extracellular matrix protein*; *non-receptor serine/threonine protein kinase* e *serine/threonine protein kinase receptor* para *serine/threonine protein kinase receptor proteins*; *microtubule binding motor protein* e

non-motor microtubule binding protein para *microtubule binding protein*; *voltage-gated ion channel* para *ion channel* (Figura 3).

Classe PANTHER	**Agrupamento**
Extracellular matrix structural protein	Extracellular matrix protein
Non-receptor serine/threonine protein kinase	Serine/Threonine protein kinase receptor proteins
Serine/Threonine protein kinase receptor	Serine/Threonine protein kinase receptor proteins
Microtubule binding motor protein	Microtubule binding protein
Non-motor microtubule binding protein	Microtubule binding protein
Voltage-gated ion channel	Ion channe

Figura 3 - Agrupamentos menores em classes do PANTHER relacionadas (classe Panther = classificação do PANTHER; Agrupamento menor = classificação criada para agrupar classes relacionas)

Neste trabalho foram analisadas 8 classes gênicas: *extracellular matrix protein*, *ion channel*, *chromatin/chromatin-binding, or -regulatory protein*, *microtubule binding protein*, *serine/threonine protein kinase receptor proteins*, *transmembrane signal receptor*, *ATP-binding cassette (ABC) transporter* e *ubiquitin-protein ligase*. Dos quais, o que apresentou mais genes avaliados, foi *extracellular matrix protein* com 11 genes, e 4 classes apresentaram apenas 3 genes avaliados cada (Figura 4).

PANTHER_CLASS	**Qnt_genes**
Extracelular matrix protein	11
Ion channel	8
Chromatin/chromatin-binding, or -regulatory protein	4
Microtubule binding protein	4

Serine/threonine protein kinase receptor proteins	3
Transmembrane signal receptor	3
ATP-binding cassete (ABC) transporter	3
Ubiquitin-protein ligase	3

Figura 4 - Quantidade de genes avaliados por classificação (qnt_genes = quantidade de genes; PANTHER_CLASS = classificação PANTHER)

Também foram medidas as quantidades de variantes e qual foi a média de variantes *pathogenic* e *benign* por grupo avaliado. Os grupos que obtiveram maior e menor média de variantes *pathogenic* foram, respectivamente, *ubiquitin-protein ligase* com 493 e *chromatin/chromatin-binding or -regulatory protein* com 101,5. Já os grupos que obtiveram maior e menor média de variantes *benign* foram, respectivamente, *ubiquitin-protein ligase* com 106 e *ATP-binding cassete (ABC) transporter* com 12,67 (Figura 5).

PANTHER_CLASS	**Média Pathogenic**	**Média Benign**
ATP-binding cassete (ABC) transporter	214,00	12,67
Chromatin/chromatin-binding, or -regulatory protein	101,50	60,00
Extracelular matrix protein	316,27	30,82
Ion channel	195,75	43,50
Microtubule binding protein	125,50	43,00
Serine/threonine protein kinase receptor proteins	152,00	20,33
Transmembrane signal receptor	159,33	60,67

Ubiquitin-protein ligase	493,00	106,00

Figura 5 - Médias de variantes por grupo gênico (PANTHER_CLASS = classificação do PANTHER)

3.1.3. Classes dos genes e avaliação dos preditores

Para fazer a avaliação dos preditores optou-se pela utilização de algumas estatísticas, os cálculos de sensibilidade, especificidade, acurácia e valor de *kappa*. Sensibilidade e especificidade foram utilizadas para inferir o valor de *Youden* que posteriormente foi utilizado para avaliar os preditores frente aos grupos gênicos.

Como já mencionado, a avaliação final que objetiva o presente trabalho utilizou as médias marginais de *Youden* para fazer comparações múltiplas par a par com intuito de identificar associações entre os preditores e os grupos gênicos estatisticamente significativos. Através disso, podemos observar que para a classe *ATP-binding cassette (ABC) transporter* os preditores *BayesDel_addAF* (preditor *BayesDel* que utiliza frequência alélica), *ClinPred* e *MetaRNN* que apresentaram valores de *emmeans*, respectivamente de, 0,75956 (0,46238 - 1,05675), 0,75661 (0,45942 - 1,05379) e 0,72163 (0,42445 - 1,01882), demonstraram estarem no mesmo grupo (.group = a ou .group = ab) onde as diferenças estatísticas não existem ou não são grandes o suficientes para separá-los em grupos diferentes dentro da determinada *PANTHER_CLASS*. Para a classe *chromatin/chromatin-binding or - regulatory protein* os preditores *MetaRNN* e *REVEL* não obtiveram diferenças estatísticas o suficiente para separá-los em grupos diferentes (*emmeans* de 0,88729 (0,62992 - 1,14466) e 0,84761 (0,59024 - 1,10498), respectivamente). É importante frisar que os preditores *BayesDel_addAF e ClinPred* apresentaram diferença estatística mas ainda sim ficaram bem próximos (.group = abc), salientando também, os bons valores de *emmeans*, respectivamente, 0,75818 (0,50081 - 1,01555) e 0,74431 (0,48694 - 1,00168).

programs	PANTHER_CLASS	emmean	SE	df	lower.CL	upper.CL	.group
BayesDel_addAF_pred	ATP-binding cassette (ABC) transporter	0,759565999	0,09227095335	1100	0,4623798926	1,056752105	a
ClinPred_pred	ATP-binding cassette (ABC) transporter	0,756607616	0,09227095335	1100	0,4594215096	1,053793722	a
MetaRNN_pred	ATP-binding cassette (ABC) transporter	0,7216317173	0,09227095335	1100	0,4244456109	1,018817824	ab
PROVEAN_pred	ATP-binding cassette (ABC) transporter	0,709071589	0,1130083769	1100	0,3450944293	1,073048749	abc
REVEL_score	ATP-binding cassette (ABC) transporter	0,6701221027	0,09227095335	1100	0,3729359962	0,9673082091	abc

BayesDel_noAF_pred	ATP-binding cassette (ABC) transporter	0,6633011993	0,09227095335	1100	0,3661150929	0,9604873058	abc
VEST4_score	ATP-binding cassette (ABC) transporter	0,6499017687	0,09227095335	1100	0,3527156622	0,9470878751	abc
MetaSVM_pred	ATP-binding cassette (ABC) transporter	0,6073654477	0,09227095335	1100	0,3101793412	0,9045515541	abcd
DEOGEN2_pred	ATP-binding cassette (ABC) transporter	0,5710447945	0,1130083769	1100	0,2070676348	0,9350219542	abcde
Polyphen2_HDIV_pred	ATP-binding cassette (ABC) transporter	0,5220517087	0,09227095335	1100	0,2248656022	0,8192378151	abcdef
SIFT_pred	ATP-binding cassette (ABC) transporter	0,5190463835	0,1130083769	1100	0,1550692238	0,8830235432	abcdef
SIFT4G_pred	ATP-binding cassette (ABC) transporter	0,508308004	0,09227095335	1100	0,2111218976	0,8054941104	abcdef
MutationTaster_pred	ATP-binding cassette (ABC) transporter	0,5080042103	0,09227095335	1100	0,2108181039	0,8051903168	abcdef
Polyphen2_HVAR_pred	ATP-binding cassette (ABC) transporter	0,4986636203	0,09227095335	1100	0,2014775139	0,7958497268	abcdef
MetaLR_pred	ATP-binding cassette (ABC) transporter	0,4836255157	0,09227095335	1100	0,1864394092	0,7808116221	abcdef
LIST-S2_pred	ATP-binding cassette (ABC) transporter	0,459283821	0,09227095335	1100	0,1620977146	0,7564699274	bcdefg
MutationAssessor_pred	ATP-binding cassette (ABC) transporter	0,4255549413	0,09227095335	1100	0,1283688349	0,7227410478	cdefg
fathmm-XF_coding_pred	ATP-binding cassette (ABC) transporter	0,3492943343	0,09227095335	1100	0,05210822789	0,6464804408	defgh
fathmm-MKL_coding_pred	ATP-binding cassette (ABC) transporter	0,3398294993	0,09227095335	1100	0,04264339289	0,6370156058	defghi
GenoCanyon_score	ATP-binding cassette (ABC) transporter	0,268091125	0,09227095335	1100	-0,02909498144	0,5652772314	efghij
Eigen-phred_coding	ATP-binding cassette (ABC) transporter	0,2553880977	0,09227095335	1100	-0,04179800878	0,5525742041	efghij
LRT_pred	ATP-binding cassette (ABC) transporter	0,240158456	0,09227095335	1100	-0,05702765044	0,5373445624	fghij
Eigen-PC-phred_coding	ATP-binding cassette (ABC) transporter	0,1708836137	0,09227095335	1100	-0,1263024928	0,4680697201	ghijk
MVP_score	ATP-binding cassette (ABC) transporter	0,1070652127	0,09227095335	1100	-0,1901208938	0,4042513191	hijkl
DANN_score	ATP-binding cassette (ABC) transporter	0,073380172	0,09227095335	1100	-0,2238059344	0,3705662784	hijkl
PrimateAI_pred	ATP-binding cassette (ABC) transporter	0,055222876	0,09227095335	1100	-0,2419632304	0,3524089824	ijkl
M-CAP_pred	ATP-binding cassette (ABC) transporter	0,0182057895	0,1130083769	1100	-0,3457713702	0,3821829492	ijkl
LINSIGHT	ATP-binding cassette (ABC) transporter	0	0,09227095335	1100	-0,2971861064	0,2971861064	jkl
CADD_phred	ATP-binding cassette (ABC) transporter	0	0,09227095335	1100	-0,2971861064	0,2971861064	jkl
CADD_phred_hg19	ATP-binding cassette (ABC) transporter	0	0,09227095335	1100	-0,2971861064	0,2971861064	jkl
MPC_score	ATP-binding cassette (ABC) transporter	0	0,09227095335	1100	-0,2971861064	0,2971861064	jkl
GERP++_RS	ATP-binding cassette (ABC) transporter	0	0,09227095335	1100	-0,2971861064	0,2971861064	jkl
FATHMM_pred	ATP-binding cassette (ABC) transporter	-0,030543406	0,1130083769	1100	-0,3945205657	0,3334337537	jkl
integrated_fitCons_score	ATP-binding cassette (ABC) transporter	-0,07367415967	0,09227095335	1100	-0,3708602661	0,2235119468	kl
HUVEC_fitCons_score	ATP-binding cassette (ABC) transporter	-0,07616603367	0,09227095335	1100	-0,3733521401	0,2210200728	kl
MutPred_score	ATP-binding cassette (ABC) transporter	-0,1102183087	0,09227095335	1100	-0,4074044151	0,1869677978	kl
H1-hESC_fitCons_score	ATP-binding cassette (ABC) transporter	-0,1205077947	0,09227095335	1100	-0,4176939011	0,1766783118	kl
GM12878_fitCons_score	ATP-binding cassette (ABC) transporter	-0,122208656	0,09227095335	1100	-0,4193947624	0,1749774504	l

Figura 6 - Parte da tabela de comparações múltiplas par a par referentes à classe *ATP-binding cassete (ABC) transporter*

programs	PANTHER_CLASS	emmean	SE	df	lower.CL	upper.CL	.group
MetaRNN_pred	chromatin/chromatin-binding. or -regulatory protein	0,8872857473	0,07990898963	1100	0,6299150294	1,144656465	a
REVEL_score	chromatin/chromatin-binding. or -regulatory protein	0,8476106988	0,07990898963	1100	0,5902399809	1,104981417	ab
BayesDel_addAF_pred	chromatin/chromatin-binding. or -regulatory protein	0,7581796165	0,07990898963	1100	0,5008088987	1,015550334	abc
ClinPred_pred	chromatin/chromatin-binding. or -regulatory protein	0,7443128475	0,07990898963	1100	0,4869421297	1,001683565	abc
BayesDel_noAF_pred	chromatin/chromatin-binding. or -regulatory protein	0,648992528	0,07990898963	1100	0,3916218102	0,9063632458	abcd
MutationAssessor_pred	chromatin/chromatin-binding. or -regulatory protein	0,6390109197	0,09227095335	1100	0,3418248132	0,9361970261	abcde
PrimateAI_pred	chromatin/chromatin-binding. or -regulatory protein	0,6262468783	0,09227095335	1100	0,3290607719	0,9234329848	abcde
Polyphen2_HVAR_pred	chromatin/chromatin-binding. or -regulatory protein	0,6209569808	0,07990898963	1100	0,3635862629	0,8783276986	bcde
VEST4_score	chromatin/chromatin-binding. or -regulatory protein	0,598573374	0,07990898963	1100	0,3412026562	0,8559440918	cdef
Polyphen2_HDIV_pred	chromatin/chromatin-binding. or -regulatory protein	0,567693391	0,07990898963	1100	0,3103226732	0,8250641088	cdefg
SIFT4G_pred	chromatin/chromatin-binding. or -regulatory protein	0,5446083943	0,07990898963	1100	0,2872376764	0,8019791121	cdefgh
DEOGEN2_pred	chromatin/chromatin-binding. or -regulatory protein	0,536623705	0,07990898963	1100	0,2792529872	0,7939944228	cdefgh
MutPred_score	chromatin/chromatin-binding. or -regulatory protein	0,4840438123	0,07990898963	1100	0,2266730944	0,7414145301	defghi
PROVEAN_pred	chromatin/chromatin-binding. or -regulatory protein	0,4409001083	0,07990898963	1100	0,1835293904	0,6982708261	defghi
SIFT_pred	chromatin/chromatin-binding. or -regulatory protein	0,3884507533	0,07990898963	1100	0,1310800354	0,6458214711	efghij
MetaSVM_pred	chromatin/chromatin-binding. or -regulatory protein	0,3597436318	0,07990898963	1100	0,1023729139	0,6171143496	fghij
LIST-S2_pred	chromatin/chromatin-binding. or -regulatory protein	0,3462191775	0,07990898963	1100	0,08884845967	0,6035898953	ghij
M-CAP_pred	chromatin/chromatin-binding. or -regulatory protein	0,3034822605	0,07990898963	1100	0,04611154267	0,5608529783	hijk
MetaLR_pred	chromatin/chromatin-binding. or -regulatory protein	0,2911840858	0,07990898963	1100	0,03381336792	0,5485548036	ijk
MVP_score	chromatin/chromatin-binding. or -regulatory protein	0,2704545455	0,07990898963	1100	0,01308382767	0,5278252633	ijkl
LRT_pred	chromatin/chromatin-binding. or -regulatory protein	0,194668334	0,07990898963	1100	-0,06270238383	0,4520390518	jklm

MutationTaster_pred	chromatin/chromatin-binding. or -regulatory protein	0,1909395595	0,07990898963	1100	-0,06643115833	0,4483102773	jklm
fathmm-MKL_coding_pred	chromatin/chromatin-binding. or -regulatory protein	0,06185803175	0,07990898963	1100	-0,1955126861	0,3192287496	klmn
DANN_score	chromatin/chromatin-binding. or -regulatory protein	0,03889433125	0,07990898963	1100	-0,2184763866	0,2962650491	lmn
Eigen-phred_coding	chromatin/chromatin-binding. or -regulatory protein	0,01875	0,07990898963	1100	-0,2386207178	0,2761207178	mn
FATHMM_pred	chromatin/chromatin-binding. or -regulatory protein	0,01220530725	0,07990898963	1100	-0,2451654106	0,2695760251	mn
Eigen-PC-phred_coding	chromatin/chromatin-binding. or -regulatory protein	0,009375	0,07990898963	1100	-0,2479957178	0,2667457178	mn
CADD_phred_hg19	chromatin/chromatin-binding. or -regulatory protein	0	0,07990898963	1100	-0,2573707178	0,2573707178	mno
MPC_score	chromatin/chromatin-binding. or -regulatory protein	0	0,07990898963	1100	-0,2573707178	0,2573707178	mno
CADD_phred	chromatin/chromatin-binding. or -regulatory protein	0	0,07990898963	1100	-0,2573707178	0,2573707178	mno
GERP++_RS	chromatin/chromatin-binding. or -regulatory protein	0	0,07990898963	1100	-0,2573707178	0,2573707178	mno
LINSIGHT	chromatin/chromatin-binding. or -regulatory protein	-0,001968504	0,07990898963	1100	-0,2593392218	0,2554022138	mno
GenoCanyon_score	chromatin/chromatin-binding. or -regulatory protein	-0,01322629475	0,07990898963	1100	-0,2705970126	0,2441444231	mno
HUVEC_fitCons_score	chromatin/chromatin-binding. or -regulatory protein	-0,1120652925	0,07990898963	1100	-0,3694360103	0,1453054253	no
integrated_fitCons_score	chromatin/chromatin-binding. or -regulatory protein	-0,1249430397	0,07990898963	1100	-0,3823137576	0,1324276781	no
H1-hESC_fitCons_score	chromatin/chromatin-binding. or -regulatory protein	-0,1249430398	0,07990898963	1100	-0,3823137576	0,1324276781	no
GM12878_fitCons_score	chromatin/chromatin-binding. or -regulatory protein	-0,1276902925	0,07990898963	1100	-0,3850610103	0,1296804253	no
fathmm-XF_coding_pred	chromatin/chromatin-binding. or -regulatory protein	-0,2566356347	0,09227095335	1100	-0,5538217411	0,04055047178	o

Figura 7 - Parte da tabela de comparações múltiplas par a par referentes à classe *chromatin/chromatin-binding. or -regulatory protein*.

A próxima classe avaliada *extracellular matrix protein* os preditores *ClinPred* e *MetaRNN* que apresentaram valores de *emmeans*, respectivamente de 0,91120 (0,75600 - 1,0664) e 0,89346 (0,73826 - 1,04866) demonstraram estarem no mesmo grupo para essa determinada *PANTHER_CLASS*, destaque, para, novamente o *BayesDel_addAF* estar com diferença estatística, mesmo que não elencado no mesmo grupo, apresentou bom valor de *emmeans* 0,87059 (0,71539 - 1,02579). Já,

na classe *ion channel* os preditores *MetaRNN* e *ClinPred* que apresentaram valores de *emmeans*, respectivamente de 0,6991 (0,51711 - 0,88108) e 0,68013 (0,49814 - 0,86212) demonstraram estarem no mesmo grupo para essa classe, e, novamente, destaque para o *BayesDel_addAF* que está na mesma circunstância da classe anterior com um bom valor de *emmeans* 0,65299 (0,47101 - 0,83499).

programs	PANTHER_CLASS	emmean	SE	df	lower.CL	upper.CL	.group
ClinPred_pred	extracellular matrix protein	0,9112012202	0,04818693382	1100	0,7560008378	1,066401603	a
MetaRNN_pred	extracellular matrix protein	0,8934608767	0,04818693382	1100	0,7382604944	1,048661259	ab
BayesDel_addAF_pred	extracellular matrix protein	0,8705910313	0,04818693382	1100	0,7153906489	1,025791414	abc
VEST4_score	extracellular matrix protein	0,7548247252	0,04818693382	1100	0,5996243428	0,9100251076	bcd
REVEL_score	extracellular matrix protein	0,7488199681	0,04818693382	1100	0,5936195857	0,9040203505	cd
BayesDel_noAF_pred	extracellular matrix protein	0,7450616902	0,04818693382	1100	0,5898613078	0,9002620726	cd
SIFT4G_pred	extracellular matrix protein	0,7126934899	0,05053888255	1100	0,5499179556	0,8754690242	de
SIFT_pred	extracellular matrix protein	0,7070003492	0,04818693382	1100	0,5517999668	0,8622007316	de
MutationAssessor_pred	extracellular matrix protein	0,7054784649	0,05650418845	1100	0,523489885	0,8874670447	de
DEOGEN2_pred	extracellular matrix protein	0,6656213123	0,06040551831	1100	0,4710673368	0,8601752878	defg
MetaSVM_pred	extracellular matrix protein	0,6585939268	0,04818693382	1100	0,5033935444	0,8137943092	def
Polyphen2_HVAR_pred	extracellular matrix protein	0,6356791484	0,05650418845	1100	0,4536905685	0,8176677282	defg
PROVEAN_pred	extracellular matrix protein	0,6292698119	0,04818693382	1100	0,4740694295	0,7844701943	defg
Polyphen2_HDIV_pred	extracellular matrix protein	0,5652712491	0,05650418845	1100	0,3832826693	0,747259829	efgh
LIST-S2_pred	extracellular matrix protein	0,5466499617	0,05053888255	1100	0,3838744274	0,709425496	fgh
MutationTaster_pred	extracellular matrix protein	0,5207040595	0,04818693382	1100	0,3655036771	0,6759044418	fgh
MetaLR_pred	extracellular matrix protein	0,5065045377	0,04818693382	1100	0,3513041554	0,6617049201	ghi
PrimateAI_pred	extracellular matrix protein	0,4490699894	0,04818693382	1100	0,293869607	0,6042703717	hij
LRT_pred	extracellular matrix protein	0,4329951637	0,05053888255	1100	0,2702196294	0,595770698	hij
fathmm-XF_coding_pred	extracellular matrix protein	0,3698619836	0,05053888255	1100	0,2070864493	0,5326375179	ijk
fathmm-MKL_coding_pred	extracellular matrix protein	0,3578448776	0,04818693382	1100	0,2026444953	0,51304526	jk
Eigen-phred_coding	extracellular matrix protein	0,2300921369	0,04818693382	1100	0,07489175453	0,3852925193	kl
Eigen-PC-phred_coding	extracellular matrix protein	0,1701880689	0,04818693382	1100	0,01498768653	0,3253884513	lm
GenoCanyon_score	extracellular matrix protein	0,1552267895	0,04818693382	1100	0,00002640707736	0,3104271718	lm
FATHMM_pred	extracellular matrix protein	0,1355650318	0,04818693382	1100	-0,01963535056	0,2907654142	lmn
MVP_score	extracellular matrix protein	0,1251729437	0,04818693382	1100	-0,03002743865	0,2803733261	lmn
MutPred_score	extracellular matrix protein	0,1114711382	0,04818693382	1100	-0,0437292442	0,2666715206	lmn
DANN_score	extracellular matrix protein	0,1029457809	0,04818693382	1100	-0,05225460147	0,2581461633	lmn

M-CAP_pred	extracellular matrix protein	0,06091874155	0,04818693382	1100	-0,09428164083	0,2161191239	mn
MPC_score	extracellular matrix protein	0	0,04818693382	1100	-0,1552003824	0,1552003824	no
CADD_phred	extracellular matrix protein	0	0,04818693382	1100	-0,1552003824	0,1552003824	no
GERP++_RS	extracellular matrix protein	0	0,04818693382	1100	-0,1552003824	0,1552003824	no
CADD_phred_hg19	extracellular matrix protein	0	0,04818693382	1100	-0,1552003824	0,1552003824	no
LINSIGHT	extracellular matrix protein	-0,003758452091	0,04818693382	1100	-0,1589588345	0,1514419303	no
integrated_fitCons_score	extracellular matrix protein	-0,1388256431	0,04818693382	1100	-0,2940260255	0,01637473929	op
HUVEC_fitCons_score	extracellular matrix protein	-0,1501098363	0,04818693382	1100	-0,3053102186	0,005090546104	p
GM12878_fitCons_score	extracellular matrix protein	-0,1531023235	0,04818693382	1100	-0,3083027058	0,002098058923	p
H1-hESC_fitCons_score	extracellular matrix protein	-0,1696473533	0,04818693382	1100	-0,3248477356	-0,0144469709	p

Figura 8 - Parte da tabela de comparações múltiplas par a par referentes à classe *extracellular matrix protein*

programs	PANTHER_CLASS	emmean	SE	df	lower.CL	upper.CL	.group
MetaRNN_pred	ion channel	0,6990962153	0,05650418845	1100	0,5171076354	0,8810847951	a
ClinPred_pred	ion channel	0,6801297372	0,05650418845	1100	0,4981411574	0,8621183171	ab
BayesDel_addAF_pred	ion channel	0,6529941221	0,05650418845	1100	0,4710055423	0,834982702	abc
SIFT4G_pred	ion channel	0,5848825401	0,06040551831	1100	0,3903285647	0,7794365156	abcd
PrimateAI_pred	ion channel	0,5253482193	0,05650418845	1100	0,3433596394	0,7073367991	bcde
REVEL_score	ion channel	0,52340013	0,05650418845	1100	0,3414115501	0,7053887099	bcde
LRT_pred	ion channel	0,516541451	0,05650418845	1100	0,3345528711	0,6985300309	bcdef
SIFT_pred	ion channel	0,4983531552	0,06524541682	1100	0,288210844	0,7084954663	bcdefg
VEST4_score	ion channel	0,4491325599	0,05650418845	1100	0,26714398	0,6311211397	defg
MutationAssessor_pred	ion channel	0,4456947395	0,07990898963	1100	0,1883240217	0,7030654573	cdefgh
BayesDel_noAF_pred	ion channel	0,4298189911	0,05650418845	1100	0,2478304113	0,611807571	defg
PROVEAN_pred	ion channel	0,4281096797	0,06524541682	1100	0,2179673685	0,6382519908	defgh
LIST-S2_pred	ion channel	0,4163573486	0,06040551831	1100	0,2218033731	0,610911324	defgh
Polyphen2_HVAR_pred	ion channel	0,3980863138	0,07147277313	1100	0,1678869456	0,628285682	defghi
Polyphen2_HDIV_pred	ion channel	0,3814566702	0,07147277313	1100	0,151257302	0,6116560384	efghi
MetaSVM_pred	ion channel	0,3473584131	0,05650418845	1100	0,1653698333	0,529346993	fghi
DEOGEN2_pred	ion channel	0,3471149503	0,07990898963	1100	0,08974423242	0,6044856681	efghij
MutationTaster_pred	ion channel	0,3412208413	0,05650418845	1100	0,1592322614	0,5232094211	ghi
fathmm-XF_coding_pred	ion channel	0,2478250234	0,05650418845	1100	0,06583644352	0,4298136032	hijk
fathmm-MKL_coding_pred	ion channel	0,2156294979	0,05650418845	1100	0,03364091802	0,3976180777	ijkl
MetaLR_pred	ion channel	0,2067608154	0,05650418845	1100	0,02477223552	0,3887493952	ijkl
MutPred_score	ion channel	0,151499588	0,05650418845	1100	-0,03048899186	0,3334881679	jklm
Eigen-phred_coding	ion channel	0,1350342711	0,05650418845	1100	-0,04695430873	0,317022851	klm

GenoCanyon_score	ion channel	0,1268507039	0,05650418845	1100	-0,05513787598	0,3088392837	klmn
MVP_score	ion channel	0,1261127773	0,05650418845	1100	-0,05587580261	0,3081013571	klmn
Eigen-PC-phred_coding	ion channel	0,1184700261	0,05650418845	1100	-0,06351855373	0,300458606	klmn
M-CAP_pred	ion channel	0,0585017845	0,05650418845	1100	-0,1234867954	0,2404903644	lmno
DANN_score	ion channel	0,02935042	0,05650418845	1100	-0,1526381599	0,2113389999	mno
FATHMM_pred	ion channel	0,01120448183	0,06524541682	1100	-0,1989378293	0,221346793	mno
LINSIGHT	ion channel	0,006721628	0,05650418845	1100	-0,1752669519	0,1887102079	mno
CADD_phred_hg19	ion channel	0	0,05650418845	1100	-0,1819885799	0,1819885799	mno
MPC_score	ion channel	0	0,05650418845	1100	-0,1819885799	0,1819885799	mno
CADD_phred	ion channel	0	0,05650418845	1100	-0,1819885799	0,1819885799	mno
GERP++_RS	ion channel	0	0,05650418845	1100	-0,1819885799	0,1819885799	mno
GM12878_fitCons_score	ion channel	-0,0403657135	0,05650418845	1100	-0,2223542934	0,1416228664	no
HUVEC_fitCons_score	ion channel	-0,06264207163	0,05650418845	1100	-0,2446306515	0,1193465082	o
H1-hESC_fitCons_score	ion channel	-0,06367524887	0,05650418845	1100	-0,2456638287	0,118313331	o
integrated_fitCons_score	ion channel	-0,1081963726	0,05650418845	1100	-0,2901849525	0,07379220723	o

Figura 9 - Parte da tabela de comparações múltiplas par a par referentes à classe *ion channel*

Na classe *microtubule binding protein* os preditores *BayesDel_addAF*, *ClinPred* e *MetaRNN* que apresentaram valores de *emmeans*, respectivamente de 0,90685 (0,64948 - 1,16422), 0,89898 (0,64161 - 1,15635) e 0,83886 (0,58148 - 1,09623 demonstraram estarem no mesmo grupo. É importante frisar que os preditores *REVEL* e *BayesDel_noAF* (preditor *BayesDel* que não utiliza frequência alélica) apresentaram diferença estatística mas ainda sim ficaram bem próximos (.group = abc), salientando também, os bons valores de *emmeans*, respectivamente, 0,80331 (0,54594 - 1,06068) e 0,78661 (0,52924 - 1,04398). Já na classe *serine/threonine protein kinase receptor proteins* os preditores *MetaRNN* e *ClinPred* (*emmeans* de 0,89889 (0,60171 - 1,19608) e 0,844145 (0,54696 - 1,14133), respectivamente) demonstraram estar no mesmo grupo, e, destaque para o *REVEL* que apresentou diferença estatística, e foi agrupado a outro grupo, porém, ficou próximo e bem pontuado, com *emmeans* de 0,82305 (0,52586 - 1,12024).

programs	PANTHER_CLASS	emmean	SE	df	lower.CL	upper.CL	.group
BayesDel_addAF_pred	microtubule binding protein	0,9068472002	0,07990898963	1100	0,6494764824	1,164217918	a
ClinPred_pred	microtubule binding protein	0,8989773175	0,07990898963	1100	0,6416065997	1,156348035	a
MetaRNN_pred	microtubule binding protein	0,8388553618	0,07990898963	1100	0,5814846439	1,09622608	ab
REVEL_score	microtubule binding protein	0,8033144768	0,07990898963	1100	0,5459437589	1,060685195	abc

BayesDel_noAF_pred	microtubule binding protein	0,7866082083	0,07990898963	1100	0,5292374904	1,043978926	abc
SIFT4G_pred	microtubule binding protein	0,727897787	0,1130083769	1100	0,3639206273	1,091874947	abcd
Polyphen2_HVAR_pred	microtubule binding protein	0,644291426	0,09227095335	1100	0,3471053196	0,9414775324	abcde
PROVEAN_pred	microtubule binding protein	0,6091529063	0,09227095335	1100	0,3119667999	0,9063390128	bcdef
Polyphen2_HDIV_pred	microtubule binding protein	0,5832386177	0,09227095335	1100	0,2860525112	0,8804247241	bcdef
DEOGEN2_pred	microtubule binding protein	0,5771135653	0,09227095335	1100	0,2799274589	0,8742996718	bcdef
VEST4_score	microtubule binding protein	0,5660290065	0,07990898963	1100	0,3086582887	0,8233997243	cdef
SIFT_pred	microtubule binding protein	0,5555535497	0,09227095335	1100	0,2583674432	0,8527396561	cdef
MutationAssessor_pred	microtubule binding protein	0,537273241	0,09227095335	1100	0,2400871346	0,8344593474	cdef
LRT_pred	microtubule binding protein	0,449393528	0,07990898963	1100	0,1920228102	0,7067642458	defg
MutationTaster_pred	microtubule binding protein	0,411248633	0,07990898963	1100	0,1538779152	0,6686193508	efg
LIST-S2_pred	microtubule binding protein	0,402528365	0,09227095335	1100	0,1053422586	0,6997144714	efghi
MetaSVM_pred	microtubule binding protein	0,3917779398	0,07990898963	1100	0,1344072219	0,6491486576	efgh
M-CAP_pred	microtubule binding protein	0,381834996	0,07990898963	1100	0,1244642782	0,6392057138	efghi
MetaLR_pred	microtubule binding protein	0,3597108458	0,07990898963	1100	0,1023401279	0,6170815636	fghi
MVP_score	microtubule binding protein	0,3451204638	0,07990898963	1100	0,08774974592	0,6024911816	fghij
fathmm-MKL_coding_pred	microtubule binding protein	0,205825595	0,07990898963	1100	-0,05154512283	0,4631963128	ghijk
FATHMM_pred	microtubule binding protein	0,185560054	0,09227095335	1100	-0,1116260524	0,4827461604	ghijkl
GenoCanyon_score	microtubule binding protein	0,162277216	0,07990898963	1100	-0,09509350183	0,4196479338	hijkl
PrimateAI_pred	microtubule binding protein	0,155088921	0,07990898963	1100	-0,1022817968	0,4124596388	hijklm
MutPred_score	microtubule binding protein	0,1421197475	0,07990898963	1100	-0,1152509703	0,3994904653	ijklmn
DANN_score	microtubule binding protein	0,1079681873	0,07990898963	1100	-0,1494025306	0,3653389051	jklmn
Eigen-phred_coding	microtubule binding protein	0,09690876325	0,07990898963	1100	-0,1604619546	0,3542794811	klmn
Eigen-PC-phred_coding	microtubule binding protein	0,0704498695	0,07990898963	1100	-0,1869208483	0,3278205873	klmn
fathmm-XF_coding_pred	microtubule binding protein	0,0362999235	0,07990898963	1100	-0,2210707943	0,2936706413	klmn
CADD_phred_hg19	microtubule binding protein	0	0,07990898963	1100	-0,2573707178	0,2573707178	klmn
MPC_score	microtubule binding protein	0	0,07990898963	1100	-0,2573707178	0,2573707178	klmn
CADD_phred	microtubule binding protein	0	0,07990898963	1100	-0,2573707178	0,2573707178	klmn
GERP++_RS	microtubule binding protein	0	0,07990898963	1100	-0,2573707178	0,2573707178	klmn
LINSIGHT	microtubule binding protein	-0,0034853785	0,07990898963	1100	-0,2608560963	0,2538853393	klmn
integrated_fitCons_score	microtubule binding protein	-0,06173352625	0,07990898963	1100	-0,3191042441	0,1956371916	lmn
HUVEC_fitCons_score	microtubule binding protein	-0,0886163315	0,07990898963	1100	-0,3459870493	0,1687543863	mn
GM12878_fitCons_score	microtubule binding protein	-0,0961924285	0,07990898963	1100	-0,3535631463	0,1611782893	n
H1-hESC_fitCons_score	microtubule binding protein	-0,09663170925	0,07990898963	1100	-0,3540024271	0,1607390086	n

Figura 10 - Parte da tabela de comparações múltiplas par a par referentes à classe *microtubule binding protein*

programs	PANTHER_CLASS	emmean	SE	df	lower.CL	upper.CL	.group
MetaRNN_pred	serine/threonine protein kinase receptor proteins	0,898892774	0,09227095335	1100	0,6017066676	1,19607888	a
ClinPred_pred	serine/threonine protein kinase receptor proteins	0,844144744	0,09227095335	1100	0,5469586376	1,14133085	ab
REVEL_score	serine/threonine protein kinase receptor proteins	0,8230498633	0,09227095335	1100	0,5258637569	1,12023597	abc
BayesDel_addAF_pred	serine/threonine protein kinase receptor proteins	0,7833325143	0,09227095335	1100	0,4861464079	1,080518621	abcd
BayesDel_noAF_pred	serine/threonine protein kinase receptor proteins	0,7145211887	0,09227095335	1100	0,4173350822	1,011707295	abcde
PROVEAN_pred	serine/threonine protein kinase receptor proteins	0,686022311	0,1130083769	1100	0,3220451513	1,049999471	abcdef
Polyphen2_HVAR_pred	serine/threonine protein kinase receptor proteins	0,611027861	0,1130083769	1100	0,2470507013	0,9750050207	abcdefg
SIFT4G_pred	serine/threonine protein kinase receptor proteins	0,5895770895	0,1130083769	1100	0,2255999298	0,9535542492	abcdefg
MutationAssessor_pred	serine/threonine protein kinase receptor proteins	0,567377067	0,1130083769	1100	0,2033999073	0,9313542267	bcdefg
Polyphen2_HDIV_pred	serine/threonine protein kinase receptor proteins	0,55974581	0,1130083769	1100	0,1957686503	0,9237229697	bcdefg
MetaSVM_pred	serine/threonine protein kinase receptor proteins	0,5472305473	0,09227095335	1100	0,2500444409	0,8444166538	cdefg
MetaLR_pred	serine/threonine protein kinase receptor proteins	0,506590632	0,09227095335	1100	0,2094045256	0,8037767384	defg
MutPred_score	serine/threonine protein kinase receptor proteins	0,4764513407	0,09227095335	1100	0,1792652342	0,7736374471	efgh
MutationTaster_pred	serine/threonine protein kinase receptor proteins	0,4694677873	0,09227095335	1100	0,1722816809	0,7666538938	efgh
fathmm-MKL_coding_pred	serine/threonine protein kinase receptor proteins	0,4661220043	0,09227095335	1100	0,1689358979	0,7633081108	efgh
PrimateAI_pred	serine/threonine protein kinase receptor proteins	0,4645632143	0,09227095335	1100	0,1673771079	0,7617493208	efgh
SIFT_pred	serine/threonine protein kinase receptor proteins	0,4539002665	0,1130083769	1100	0,08992310678	0,8178774262	efghi
fathmm-XF_coding_pred	serine/threonine protein kinase receptor proteins	0,4266384045	0,1130083769	1100	0,06266124478	0,7906155642	efghij
LRT_pred	serine/threonine protein kinase receptor proteins	0,407456747	0,09227095335	1100	0,1102706406	0,7046428534	fghi
LIST-S2_pred	serine/threonine protein kinase receptor proteins	0,39481352	0,1130083769	1100	0,03083636028	0,7587906797	efghijk

VEST4_score	serine/threonine protein kinase receptor proteins	0,322089947	0,09227095335	1100	0,02490384056	0,6192760534	ghijk
HUVEC_fitCons_score	serine/threonine protein kinase receptor proteins	0,1903205727	0,09227095335	1100	-0,1068655338	0,4875066791	hijkl
GM12878_fitCons_score	serine/threonine protein kinase receptor proteins	0,142701525	0,09227095335	1100	-0,1544845814	0,4398876314	ijkl
MVP_score	serine/threonine protein kinase receptor proteins	0,131372549	0,09227095335	1100	-0,1658135574	0,4285586554	ijkl
M-CAP_pred	serine/threonine protein kinase receptor proteins	0,1115731117	0,09227095335	1100	-0,1856129948	0,4087592181	jkl
integrated_fitCons_score	serine/threonine protein kinase receptor proteins	0,1007502417	0,09227095335	1100	-0,1964358648	0,3979363481	kl
DEOGEN2_pred	serine/threonine protein kinase receptor proteins	0,1	0,1130083769	1100	-0,2639771597	0,4639771597	ijkl
Eigen-PC-phred_coding	serine/threonine protein kinase receptor proteins	0,07142857133	0,09227095335	1100	-0,2257575351	0,3686146778	kl
Eigen-phred_coding	serine/threonine protein kinase receptor proteins	0,07142857133	0,09227095335	1100	-0,2257575351	0,3686146778	kl
FATHMM_pred	serine/threonine protein kinase receptor proteins	0,048389111	0,1130083769	1100	-0,3155880487	0,4123662707	kl
GenoCanyon_score	serine/threonine protein kinase receptor proteins	0,02979507567	0,09227095335	1100	-0,2673910308	0,3269811821	l
DANN_score	serine/threonine protein kinase receptor proteins	0,02380952367	0,09227095335	1100	-0,2733765828	0,3209956301	l
CADD_phred_hg19	serine/threonine protein kinase receptor proteins	0,002506265667	0,09227095335	1100	-0,2946798408	0,2996923721	l
CADD_phred	serine/threonine protein kinase receptor proteins	0,002506265667	0,09227095335	1100	-0,2946798408	0,2996923721	l
MPC_score	serine/threonine protein kinase receptor proteins	0,002506265667	0,09227095335	1100	-0,2946798408	0,2996923721	l
GERP++_RS	serine/threonine protein kinase receptor proteins	0,002506265667	0,09227095335	1100	-0,2946798408	0,2996923721	l
LINSIGHT	serine/threonine protein kinase receptor proteins	0	0,09227095335	1100	-0,2971861064	0,2971861064	l
H1-hESC_fitCons_score	serine/threonine protein kinase receptor proteins	-0,06698117833	0,09227095335	1100	-0,3641672848	0,2302049281	l

Figura 11 - Parte da tabela de comparações múltiplas par a par referentes à classe *serine/threonine protein kinase receptor proteins*

Outra classe avaliada foi a *transmembrane signal receptor*, onde os preditores *BayesDel_addAF*, *MetaRNN*, *REVEL*, *BayesDel_noAF*, *PROVEAN*, *ClinPred* e *VEST4* (*emmeans* de 0,80253 (0,50534 - 1,09972), 0,76287 (0,46568 - 1,060051812),

0,75283 (0,45565 - 1,05002), 0,73484 (0,43766 - 1,03203), 0,70843 (0,34445 - 1,07240), 0,70173 (0,40454 - 0,99891), 0,68519 (0,38801 - 0,98238), respectivamente) demonstraram estarem vinculados a um mesmo grupo para a classe corrente. Já na última classe avaliada, a *ubiquitin-protein ligase*, os preditores *ClinPred*, *MetaRNN* e *BayesDel_addAF* (*emmeans* de 0,80628 (0,50909 - 1,10346), 0,7754 (0,47822 - 1,07259) e 0,71427 (0,41709 - 1,01145), respectivamente) demonstraram estar no mesmo grupo, sem diferença estatística significativa.

programs	PANTHER_CLASS	emmean	SE	df	lower.CL	upper.CL	.group
BayesDel_addAF_pred	transmembrane signal receptor	0,802531397	0,09227095335	1100	0,5053452906	1,099717503	a
MetaRNN_pred	transmembrane signal receptor	0,7628657053	0,09227095335	1100	0,4656795989	1,060051812	a
REVEL_score	transmembrane signal receptor	0,752834467	0,09227095335	1100	0,4556483606	1,050020573	a
BayesDel_noAF_pred	transmembrane signal receptor	0,7348443227	0,09227095335	1100	0,4376582162	1,032030429	a
PROVEAN_pred	transmembrane signal receptor	0,708425503	0,1130083769	1100	0,3444483433	1,072402663	ab
ClinPred_pred	transmembrane signal receptor	0,7017268447	0,09227095335	1100	0,4045407382	0,9989129511	ab
VEST4_score	transmembrane signal receptor	0,6851953603	0,09227095335	1100	0,3880092539	0,9823814668	ab
SIFT_pred	transmembrane signal receptor	0,6217003235	0,1130083769	1100	0,2577231638	0,9856774832	abc
Polyphen2_HVAR_pred	transmembrane signal receptor	0,6204795207	0,09227095335	1100	0,3232934142	0,9176656271	abc
SIFT4G_pred	transmembrane signal receptor	0,6088403273	0,09227095335	1100	0,3116542209	0,9060264338	abc
MutationAssessor_pred	transmembrane signal receptor	0,578887823	0,09227095335	1100	0,2817017166	0,8760739294	abc
Polyphen2_HDIV_pred	transmembrane signal receptor	0,5473712797	0,09227095335	1100	0,2501851732	0,8445573861	abcd
LIST-S2_pred	transmembrane signal receptor	0,4886185245	0,1130083769	1100	0,1246413648	0,8525956842	abcde
DEOGEN2_pred	transmembrane signal receptor	0,4405316903	0,09227095335	1100	0,1433455839	0,7377177968	bcdef
PrimateAI_pred	transmembrane signal receptor	0,4369929453	0,09227095335	1100	0,1398068389	0,7341790518	bcdef
LRT_pred	transmembrane signal receptor	0,381604086	0,09227095335	1100	0,08441797956	0,6787901924	cdefg
MutationTaster_pred	transmembrane signal receptor	0,3775510203	0,09227095335	1100	0,08036491389	0,6747371268	cdefg
M-CAP_pred	transmembrane signal receptor	0,3763024013	0,09227095335	1100	0,07911629489	0,6734885078	cdefg
MVP_score	transmembrane signal receptor	0,2794762777	0,09227095335	1100	-0,01770982878	0,5766623841	defgh
MetaSVM_pred	transmembrane signal receptor	0,2485910913	0,09227095335	1100	-0,04859501511	0,5457771978	efgh
GenoCanyon_score	transmembrane signal receptor	0,2345325307	0,09227095335	1100	-0,06265357578	0,5317186371	efgh
MutPred_score	transmembrane signal receptor	0,1880429093	0,09227095335	1100	-0,1091431971	0,4852290158	efghi
fathmm-MKL_coding_pred	transmembrane signal receptor	0,185941043	0,09227095335	1100	-0,1112450634	0,4831271494	efghi
fathmm-XF_coding_pred	transmembrane signal receptor	0,169556105	0,09227095335	1100	-0,1276300014	0,4667422114	efghi
Eigen-phred_coding	transmembrane signal receptor	0,1587301587	0,09227095335	1100	-0,1384559478	0,4559162651	fghi

MetaLR_pred	transmembrane signal receptor	0,151132994	0,09227095335	1100	-0,1460531124	0,4483191004	fghi
Eigen-PC-phred_coding	transmembrane signal receptor	0,1394557823	0,09227095335	1100	-0,1577303241	0,4366418888	ghij
DANN_score	transmembrane signal receptor	0,1317155067	0,09227095335	1100	-0,1654705998	0,4289016131	ghij
CADD_phred_hg19	transmembrane signal receptor	0	0,09227095335	1100	-0,2971861064	0,2971861064	hijk
CADD_phred	transmembrane signal receptor	0	0,09227095335	1100	-0,2971861064	0,2971861064	hijk
MPC_score	transmembrane signal receptor	0	0,09227095335	1100	-0,2971861064	0,2971861064	hijk
GERP++_RS	transmembrane signal receptor	0	0,09227095335	1100	-0,2971861064	0,2971861064	hijk
LINSIGHT	transmembrane signal receptor	-0,003401360667	0,09227095335	1100	-0,3005874671	0,2937847458	hijk
integrated_fitCons_score	transmembrane signal receptor	-0,010433019	0,09227095335	1100	-0,3076191254	0,2867530874	hijk
FATHMM_pred	transmembrane signal receptor	-0,022363351	0,1130083769	1100	-0,3863405107	0,3416138087	hijk
GM12878_fitCons_score	transmembrane signal receptor	-0,09816849833	0,09227095335	1100	-0,3953546048	0,1990176081	ijk
HUVEC_fitCons_score	transmembrane signal receptor	-0,1453623757	0,09227095335	1100	-0,4425484821	0,1518237308	jk
H1-hESC_fitCons_score	transmembrane signal receptor	-0,1693136227	0,09227095335	1100	-0,4664997291	0,1278724838	k

Figura 12 - Parte da tabela de comparações múltiplas par a par referentes à classe *transmembrane signal receptor*

programs	PANTHER_CLASS	emmean	SE	df	lower.CL	upper.CL	.group
ClinPred_pred	ubiquitin-protein ligase	0,8062776643	0,09227095335	1100	0,5090915579	1,103463771	a
MetaRNN_pred	ubiquitin-protein ligase	0,775403347	0,09227095335	1100	0,4782172406	1,072589453	a
BayesDel_addAF_pred	ubiquitin-protein ligase	0,7142664887	0,09227095335	1100	0,4170803822	1,011452595	ab
REVEL_score	ubiquitin-protein ligase	0,6888632673	0,09227095335	1100	0,3916771609	0,9860493738	abc
MetaSVM_pred	ubiquitin-protein ligase	0,67322515	0,09227095335	1100	0,3760390436	0,9704112564	abc
VEST4_score	ubiquitin-protein ligase	0,669294262	0,09227095335	1100	0,3721081556	0,9664803684	abc
BayesDel_noAF_pred	ubiquitin-protein ligase	0,6454328363	0,09227095335	1100	0,3482467299	0,9426189428	abcd
SIFT_pred	ubiquitin-protein ligase	0,620010379	0,1130083769	1100	0,2560332193	0,9839875387	abcdefg
MetaLR_pred	ubiquitin-protein ligase	0,6022685683	0,09227095335	1100	0,3050824619	0,8994546748	abcde
MutationTaster_pred	ubiquitin-protein ligase	0,594094394	0,09227095335	1100	0,2969082876	0,8912805004	abcdef
PROVEAN_pred	ubiquitin-protein ligase	0,5319627917	0,09227095335	1100	0,2347766852	0,8291488981	abcdefg
SIFT4G_pred	ubiquitin-protein ligase	0,4511660643	0,09227095335	1100	0,1539799579	0,7483521708	bcdefgh
MutationAssessor_pred	ubiquitin-protein ligase	0,409284177	0,1130083769	1100	0,04530701728	0,7732613367	bcdefghij
fathmm-MKL_coding_pred	ubiquitin-protein ligase	0,3987160073	0,09227095335	1100	0,1015299009	0,6959021138	cdefghi
Polyphen2_HDIV_pred	ubiquitin-protein ligase	0,3299622175	0,1130083769	1100	-0,03401494222	0,6939393772	defghijkl
FATHMM_pred	ubiquitin-protein ligase	0,3253994613	0,09227095335	1100	0,02821335489	0,6225855678	efghijk
LIST-S2_pred	ubiquitin-protein ligase	0,323788071	0,09227095335	1100	0,02660196456	0,6209741774	efghijk

PrimateAI_pred	ubiquitin-protein ligase	0,287230904	0,09227095335	1100	-0,009955202444	0,5844170104	ghijklm
LRT_pred	ubiquitin-protein ligase	0,262256678	0,1130083769	1100	-0,1017204817	0,6262338377	fghijklm
Eigen-phred_coding	ubiquitin-protein ligase	0,2229526383	0,09227095335	1100	-0,07423346811	0,5201387448	hijklm
MVP_score	ubiquitin-protein ligase	0,2190071157	0,09227095335	1100	-0,07817899078	0,5161932221	hijklm
MutPred_score	ubiquitin-protein ligase	0,2088970873	0,09227095335	1100	-0,08828901911	0,5060831938	hijklmn
Eigen-PC-phred_coding	ubiquitin-protein ligase	0,208039202	0,09227095335	1100	-0,08914690444	0,5052253084	hijklmn
fathmm-XF_coding_pred	ubiquitin-protein ligase	0,1749609185	0,1130083769	1100	-0,1890162412	0,5389380782	hijklmn
DEOGEN2_pred	ubiquitin-protein ligase	0,163653834	0,1130083769	1100	-0,2003233257	0,5276309937	hijklmn
HUVEC_fitCons_score	ubiquitin-protein ligase	0,1486209887	0,09227095335	1100	-0,1485651178	0,4458070951	ijklmn
Polyphen2_HVAR_pred	ubiquitin-protein ligase	0,1219790245	0,1130083769	1100	-0,2419981352	0,4859561842	hijklmn
M-CAP_pred	ubiquitin-protein ligase	0,1107881137	0,09227095335	1100	-0,1863979928	0,4079742201	ijklmn
GenoCanyon_score	ubiquitin-protein ligase	0,09822260833	0,09227095335	1100	-0,1989634981	0,3954087148	jklmn
DANN_score	ubiquitin-protein ligase	0,081122807	0,09227095335	1100	-0,2160632994	0,3783089134	jklmn
integrated_fitCons_score	ubiquitin-protein ligase	0,07234980233	0,09227095335	1100	-0,2248363041	0,3695359088	klmn
CADD_phred_hg19	ubiquitin-protein ligase	0	0,09227095335	1100	-0,2971861064	0,2971861064	lmn
MPC_score	ubiquitin-protein ligase	0	0,09227095335	1100	-0,2971861064	0,2971861064	lmn
CADD_phred	ubiquitin-protein ligase	0	0,09227095335	1100	-0,2971861064	0,2971861064	lmn
GERP++_RS	ubiquitin-protein ligase	0	0,09227095335	1100	-0,2971861064	0,2971861064	lmn
LINSIGHT	ubiquitin-protein ligase	-0,006215463333	0,09227095335	1100	-0,3034015698	0,2909706431	mn
GM12878_fitCons_score	ubiquitin-protein ligase	-0,08866714667	0,09227095335	1100	-0,3858532531	0,2085189598	n
H1-hESC_fitCons_score	ubiquitin-protein ligase	-0,09042384433	0,09227095335	1100	-0,3876099508	0,2067622621	n

Figura 13 - Parte da tabela de comparações múltiplas par a par referentes à classe *ubiquitin protein ligase*

Dentre os preditores supracitados, frente a cada grupo gênico avaliado, foi mensurado a média de valor de *kappa* e a média de acurácia para eles, ou seja, foi realizado a média desses valores para cada preditor frente aos genes que fazem parte desse grupo. Na classe *ATP-binding cassette (ABC) transporter* o preditor *BayesDel_addAF* obteve uma média de acurácia e valor de *kappa* de 0,842 e 0,399, respectivamente. Enquanto o preditor *ClinPred* obteve 0, 839 e 0,428, e o *Meta_RNN* obteve 0,793 e 0,404, para média de acurácia e valor de *kappa* respectivamente. Para a classe *chromatin/chromatin-binding. or -regulatory protein* o preditor *Meta_RNN* teve uma média de acurácia de 0,932 e uma média de *kappa* de 0,706, enquanto as médias

para o *REVEL* foram de 0,940 e 0,850, respectivamente, para o *BayesDel_addAF* foram de 0,908 e 0,761, respectivamente, e para o *ClinPred* os valores foram 0,838 e 0,618. Já, para a classe *extracellular matrix protein*, *ClinPred* obteve médias de 0,968 e 0,896, enquanto para o *Meta_RNN* foram 0,972 e 0,882 e para o *BayesDel_addAF* as médias ficaram em 0,969 e 0,867. A outra classe que também foi avaliada foi a de *ion channel* onde o preditor *Meta_RNN* obteve médias de acurácia e valor de *kappa* de 0,941 e 0,738, respectivamente, enquanto isso, o preditor *ClinPred* obteve 0,927 e 0,721 e o preditor *BayesDel_addAF* obteve 0,934 e 0,692 de médias. Na classe *microtubule binding protein* o preditor *BayesDel_addAF* obteve as médias de acurácia e valor de *kappa* de 0,942 e 0,849, respectivamente, enquanto o preditor *ClinPred* obteve 0,954 e 0,866, o preditor *Meta_RNN* teve médias de 0,936 e 0,821, o preditor *REVEL* obteve 0,899 e 0,755, e, por final, o preditor *BayesDel_noAF* obteve 0,915 e 0,754, respectivamente. Na classe *serine/threonine protein kinase receptor proteins* os preditores melhores elencados foram o *Meta_RNN*, *ClinPred* e *REVEL* que obtiveram médias de acurácia e valor de *kappa* de 0,979 e 0,891, 0,961 e 0,865, 0,960 e 0,786, respectivamente. Já para a classe da *transmembrane signal receptor* foram os preditores *BayesDel_addAF* (acurácia = 0,908 e *kappa* = 0,779), *Meta_RNN* (0,899 e 0,779), *REVEL* (0,846 e 0,674), *BayesDel_noAF* (0,880 e 0,711), *PROVEAN* (0,570 e 0,470), *ClinPred* (0,884 e 0,729) e *VEST4* (0,891 e 0,722). Por fim, a classe *ubiquitin-protein ligase*, o preditor *ClinPred* obteve as médias de acurácia e valor de *kappa* de 0,954 e 0,808, enquanto *Meta_RNN* teve 0,934 e 0,725 e o *BayesDel_addAF* registrou 0,947 e 0,672, respectivamente. Os dados podem ser vistos na figura 15.

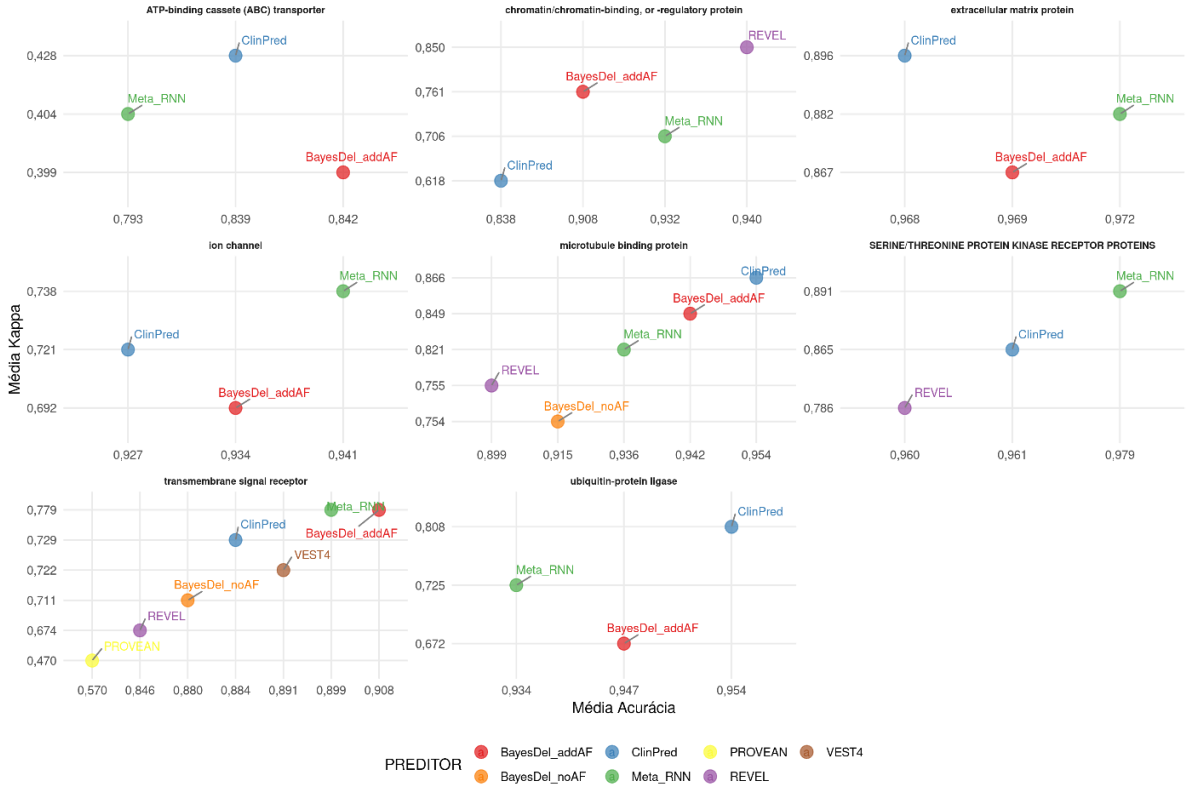

Figura 14 - Gráfico das médias de acurácia e *kappa* para cada preditor em sua determinada classe

4. DISCUSSÃO

A integração da análise *in silico*, diagnóstico molecular e insights clínicos está revolucionando o diagnóstico e a classificação de doenças genéticas. À medida que o cenário de informações genômicas e ferramentas computacionais se expande, a abordagem multidisciplinar será fundamental para desbloquear uma compreensão mais profunda dos mecanismos de doenças genéticas e orientar tratamentos personalizados. Portanto, estratégias objetivas para escolher preditores incluem avaliar seu desempenho para cada gene individual, e essa abordagem pode ajudar a melhorar a confiabilidade das análises *in silico* (Borges, 2021). Nesse contexto, é indispensável saber se a escolha de um preditor de variante apresenta diferenças em grupos de proteínas específicas e esse é o presente objetivo deste trabalho.

Neste trabalho comparamos 38 preditores de variantes frente a 39 genes agrupados em 8 grupos gênicos diferentes com o intuito de avaliar uma possível relação entre qualidade de assertividade de predição em diferentes grupos gênicos. A escolha do banco de dados *dbNSFP* foi feita devido a facilidade de ter todas as predições agrupadas em um único lugar, juntamente com a classificação de um

repositório de confiança, como o *ClinVar*. O *dbNSFP* na sua versão v4.2a apresenta muitos dados faltantes de predições para determinados preditores frente algumas variantes, e isso torna-se uma dificuldade para avaliações estatísticas devido a não igualdade de número dados para cada preditor. Portanto, em alguns casos, os preditores podem ter tido números diferentes de classificações entre eles, o que pode influenciar de alguma maneira nos cálculos.

O *ClinVar* é uma base de dados de fácil trabalho, porém, devido a diversas classificações de variantes, foi utilizado inicialmente para verificar a quantidade de variantes para fazer um corte nos dados, resgatando apenas genes com uma boa quantidade de variantes para termos estatísticas mais seguras. Porém, algumas das classificações encontradas podem possuir *'conflicting interpretations'* mesmo não estando classificadas como tal nos dados. Por exemplo, quando fomos verificar a quantidade de variantes *'pathogenic'* e *'benign'* estavamos agrupando apenas as *'pathogenic'* e/ou '*likely pathogenic'* e '*benign*' e/ou '*likely benign*', porém algumas variantes tinham várias classificações diferentes, que poderiam ser definidas como uma não concordância, porém sem apresentar o termo *'conflicting interpretations'*. Portanto, no nosso caso, embora alguma dessas classificações dentro de uma variante fossem *'pathogenic'* ou *'benign'* não eram computadas, pois não demonstraram confiança na classificação. Nesse contexto, talvez a utilização desse modelo de corte tenha reduzido bastante o número de genes do genoma que apresentavam mais de 100 variantes. Após resgatar os dados desses genes do corte de 100 ou mais variantes no *dbNSFP* verificamos que as quantidades de variantes que computamos do *ClinVar* não estavam exatamente iguais às registradas do *dbNSFP*, e, portanto, a partir desse corte inicial, optamos por fazer o mesmo corte para esses genes, mas a partir dos dados do *ClinVar* contidos no *dbNSFP,* o que gerou uma redução maior no número de genes a serem avaliados.

Outro corte realizado foi devido a classificação dos genes em grupos. No contexto desse trabalho é importante agrupar as proteínas em grupos, e a opção escolhida foi agrupar conforme suas classes proteicas. Para isso, a utilização do PANTHER é de extrema importância, visto que a ferramenta apresenta uma classificação pronta para tal. Como mencionado, optamos por utilizar apenas grupos que tinham 3 ou mais genes classificados, dos selecionados. Esse novo filtro dos dados apresentou poucas classes para serem avaliadas, portanto, ao analisar todos os genes que tiveram uma classificação *Protein Class* do *PANTHER*, verificou-se que

alguns poderiam ser agrupados em coletivos maiores. Essa separação dos genes em classes diminuiu mais ainda nosso conjunto de dados, mas, notamos uma melhor confiabilidade nas estatísticas geradas, e, mesmo tendo um baixo número amostral, nossos resultados demonstram boa credibilidade.

A avaliação dos resultados podemos perceber uma grande repetição de preditores com bons valores estatísticos, o *ClinPred*, *BayesDel_addAF* e o *Meta_RNN* se destacaram em todos os 8 grupos gênicos que avaliamos, apresentando bons valores de *emmeans*, média de acurácia e média de *kappa*. Embora tenha ocorrido essa recorrência, podemos perceber com base nos agrupamentos, que existem diferenças entre os melhores preditores para cada grupo. Por exemplo para *ATP-binding cassete (ABC) transporter* os 10 melhores preditores elencados apresentam bastante variações estatísticas, analisando os agrupamentos realizado pelas letras, notamos variações de ‘a’, até ‘abcdef’. E as variações continuam para *chromatin/chromatin-binding. or -regulatory protein* onde as variações dos agrupamentos por letras variam de ‘a’, até ‘cdefg’, o mesmo acontece com *extracellular matrix protein* com variações de ‘a’ até ‘defg’, *ion channel* com variações de ‘a’ até ‘cdefgh’, *microtubule binding protein* com divergências de ‘a’ até ‘bcdef’, *serine/threonine protein kinase receptor proteins* com variações de ‘a’ até ‘bcdefg’ , *transmembrane signal receptor* obteve a menor variação, mas ainda sim obteve indo de ‘a’ até ‘abc’ e por final *ubiquitin-protein ligase* indo de ‘a’ até ‘abcdef’.

Como já mencionado, embora tivemos uma recorrência nos preditores mais bem elencados para cada um dos grupos gênicos, podemos concluir que existe sim diferença entre os melhores preditores para cada grupo. Além disso, também podemos analisar que os preditores *H1-hESC_fitCons* e *GM12878_fitCons* estiveram quase sempre elencados nos piores preditores de cada grupo.

Sabendo que o preditor *ClinPred* é um algoritmo que funciona com base em *machine learning* aplicando principalmente conceitos de *RandonForest* (ALIREZAIE et al., 2018), que o *BayesDel_addAF* utiliza métodos rebuscados de estatística para pontuar as classificações como abordagem *naïve Bayesian* (FENG, 2017) e o *Meta_RNN* que utiliza tecnicas de *deep learning* para classificar as variantes, além de que, são algoritmos mais atuais, podemos supor que métodos analíticos mais atuais, podem apresentar uma melhor análise para classificação de variantes. Também podemos supor que técnicas que utilizam dados mais específicos como *H1-hESC_fitCons* e *GM12878_fitCons* que são preditores fitCons que utilizam dados

referentes às células-tronco embrionárias humanas H1 (H1 hESCs) e células linfoblastóides (GM12878), respectivamente, podem não ter boa previsão em determinados grupos, como os que foram avaliados por esse trabalho.

Apesar da importância da utilização de preditores para fazer validação, é recomendado que as análises *in silico* sejam utilizadas em conjunto com outras evidências, como dados funcionais e informações clínicas, para uma classificação mais confiável das variantes *missense* (RICHARDS et al., 2015). Uma abordagem multidisciplinar, que incorpore tanto a análise molecular quanto a interpretação clínica, é fundamental para a correta identificação de variantes em pacientes com doenças genéticas.

Portanto, o presente trabalho teve como objetivo verificar se os preditores de variantes genéticas podem variar sua assertividade conforme aplicados em grupos gênicos diferentes, e, conclui-se que, embora exista recorrência entre os melhores preditores para cada grupo, existe diferença na assertividade para cada grupo.

Ademais, o trabalho mostra um método de avaliar como preditores se saem em diferentes situações que as variantes se encontram, que pode ser posteriormente aplicada em outros grupos, diferentes de gênicos do *PANTHER CLASS*.

Em resumo, o estudo investiga como diferentes preditores de variantes genéticas funcionam em grupos de genes específicos, mostrando a importância de abordagens analíticas para fazer a escolha de determinado preditor na hora de classificar uma variante, e não apenas escolher de maneira aleatória ou por reconhecimento dos mesmos. Portanto, o presente trabalho pretende auxiliar usuários de programas que predizem o significado de variantes, a escolher com maior confiabilidade qual utilizar, e, consequentemente, auxiliar na correta classificação de variantes genéticas.

REFERÊNCIAS

ALIREZAIE, N. et al. ClinPred: Prediction Tool to Identify Disease-Relevant Nonsynonymous Single-Nucleotide Variants. **The American Journal of Human Genetics**, v. 103, n. 4, p. 474–483, out. 2018.

AUTON, A. et al. A global reference for human genetic variation. **Nature**, v. 526, n. 7571, p. 68–74, 1 out. 2015.

CASTIGLIA, D.; ZAMBRUNO, G. Mutation Mechanisms. **Dermatologic Clinics**, v. 28, n. 1, p. 17–22, jan. 2010.

CHEN, S. et al. A genome-wide mutational constraint map quantified from variation in 76,156 human genomes. [s.d.].

CHOE, H. et al. Molecular Diagnostics. **Journal of the American Academy of Orthopaedic Surgeons**, v. 23, p. S26–S31, abr. 2015.

CLAUSSNITZER, M. et al. A brief history of human disease genetics. **Nature**, v. 577, n. 7789, p. 179–189, 9 jan. 2020.

FAIRLEY, S. et al. The International Genome Sample Resource (IGSR) collection of open human genomic variation resources. **Nucleic Acids Research**, v. 48, n. D1, p. D941–D947, 8 jan. 2020.

FENG, B.-J. PERCH: A Unified Framework for Disease Gene Prioritization. **Human Mutation**, v. 38, n. 3, p. 243–251, mar. 2017.

GULLEY, M. L. et al. Clinical Laboratory Reports in Molecular Pathology. **Archives of Pathology & Laboratory Medicine**, v. 131, n. 6, p. 852–863, 1 jun. 2007.

LANDRUM, M. J. et al. ClinVar: improvements to accessing data. **Nucleic Acids Research**, v. 48, n. D1, p. D835–D844, 8 jan. 2020.

LAPPALAINEN, T. et al. Genomic Analysis in the Age of Human Genome Sequencing. **Cell**, v. 177, n. 1, p. 70–84, mar. 2019.

LEK, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. **Nature**, v. 536, n. 7616, p. 285–291, 17 ago. 2016.

LINDEBOOM, R. G. H.; SUPEK, F.; LEHNER, B. The rules and impact of nonsense-mediated mRNA decay in human cancers. **Nature Genetics**, v. 48, n. 10, p. 1112–1118, 12 out. 2016.

LIU, X. et al. dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. **Genome Medicine**, v. 12, n. 1, p. 103, 2 dez. 2020.

MCLAREN, W. et al. The Ensembl Variant Effect Predictor. **Genome Biology**, v. 17, n. 1, p. 122, 6 dez. 2016.

MI, H. et al. PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API. **Nucleic Acids Research**, v. 49, n. D1, p. D394–D403, 8 jan. 2021.

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

NYKAMP, K. et al. Sherloc: a comprehensive refinement of the ACMG–AMP variant classification criteria. **Genetics in Medicine**, v. 19, n. 10, p. 1105–1117, out. 2017.

PÂMELLA BORGES. **Comparação de ferramentas in silico para avaliação de patogenicidade de variantes missense**. Porto Alegre : Universidade Federal do Rio Grande do Sul, 2021.

RICHARDS, S. et al. 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. **Genetics in Medicine**, v. 17, n. 5, p. 405–424, maio 2015.

SHERRY, S. T. dbSNP: the NCBI database of genetic variation. **Nucleic Acids Research**, v. 29, n. 1, p. 308–311, 1 jan. 2001.

STONE, E. A.; SIDOW, A. Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. **Genome Research**, v. 15, n. 7, p. 978–986, jul. 2005.

SUNYAEV, S. Prediction of deleterious human alleles. **Human Molecular Genetics**, v. 10, n. 6, p. 591–597, 1 mar. 2001.

TANG, H.; THOMAS, P. D. Tools for Predicting the Functional Impact of Nonsynonymous Genetic Variation. **Genetics**, v. 203, n. 2, p. 635–647, 1 jun. 2016.

THOMAS, P. D. et al. PANTHER: A Library of Protein Families and Subfamilies Indexed by Function. **Genome Research**, v. 13, n. 9, p. 2129–2141, set. 2003.

UÇAR, M. K. et al. The Effect of Training and Testing Process on Machine Learning in Biomedical Datasets. **Mathematical Problems in Engineering**, v. 2020, p. 1–17, 13 maio 2020.

WANG, K.; LI, M.; HAKONARSON, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. **Nucleic Acids Research**, v. 38, n. 16, p. e164–e164, 1 set. 2010.

WONG, Y. K. E. et al. The applications of big data in molecular diagnostics. **Expert Review of Molecular Diagnostics**, v. 19, n. 10, p. 905–917, 3 out. 2019.

5. ARTIGO – ALGORITMO DE COMPARAÇÃO DE PREDITORES PARA ANÁLISE DE DOENÇAS GENÉTICAS

Artigo a ser enviado para a revista *Genetics and Molecular Biology*

Matheus Pereira Mai [1] e Ursula Matte [2, 3, 4]

1. Programa de Graduação em Biotecnologia/Bioinformática, UFRGS, Porto Alegre, RS, Brasil.
2. Laboratório de células, Tecidos e Genes, HCPA, Porto Alegre, RS, Brasil.
3. Núcleo de Bioinformática, HCPA, Porto Alegre, RS, Brasil.
4. Programa de Pós-Graduação em Genética e Biologia Molecular, UFRGS, Porto Alegre, RS, Brasil.

ABSTRACT:

This study compares the performance of 38 variant predictors in 39 genes grouped into 8 categories. Using the dbNSFP database and ClinVar for predictions and classifications, our results show that ClinPred, BayesDel_addAF, and Meta_RNN consistently outperform other predictors, although there are differences in particular gene classes. This highlights the importance of assessing which predictor presents the best results on a case-by-case basis with a set of curated data.

Keywords: Variant Predictors. Genetic variants. Gene groups.

Introduction

Molecular analysis of genetic diseases plays a key role in disease diagnosis, revealing the intricate relationship between sequence variations and predisposition to or occurrence of disease. This process provides a potent tool to identify underlying disease mechanisms and develop new prevention and treatment strategies (CLAUSSNITZER et al., 2020). Advances in whole-genome sequencing have allowed the detection of a broad spectrum of genetic variants across the genome, thus aiding in the research of rare and common diseases (LAPPALAINEN et al., 2019). Notably, technological advances in bioinformatics and analytical techniques, facilitated by substantial collaborative projects (AUTON et al., 2015; LEK et al., 2016), have driven the identification of disease-causing genes and the elucidation of associated biological pathways, thus strengthening precision medicine and genomics (CLAUSSNITZER et al., 2020; LAPPALAINEN et al., 2019).

The classification of genetic variants constitutes a critical step in the molecular diagnostic process, distinguishing variants based on their clinical relevance (RICHARDS et al., 2015). This differentiation is crucial, ranging from highly probable pathogenic changes to probably benign changes (RICHARDS et al., 2015). The advent of advanced sequencing technologies led to the refinement of classification methods, guided by standardized criteria. Notably, two prominent classification protocols, the American College of Medical Genetics and Genomics (ACMG) and Association for Molecular Pathology (AMP) guidelines (RICHARDS et al., 2015), along with the Sherloc protocol (NYKAMP et al., 2017), emerged to provide a systematic framework for classifying variants.

Missense variants, which are single nucleotide alterations that lead to amino acid exchange, often pose challenges in classification, unlike other types of alterations, due to their uncertain pathogenicity (NYKAMP et al., 2017). Such variants often fall into the Variant of Uncertain Meaning (VUS) category. Although experimental validation of variant impacts is ideal, the large amount of genomic data makes comprehensive experimental validation unfeasible (WONG et al., 2019). Therefore, *in silico* evaluations gained prominence as a valuable tool, providing additional evidence to assist in the classification of variants (NYKAMP et al., 2017; RICHARDS et al., 2015).

The scenario of computational variant prediction tools has grown exponentially over the years, supported by data and bioinformatics algorithms. These predictors, for example SIFT (NG, 2003), PolyPhen (SUNYAEV, 2001) and PANTHER (THOMAS et al., 2003), have progressed from basic alignment-based approaches to other programs that incorporate structural information and learning techniques. machine (BORGES, 2021). However, choosing which predictor to employ remains a challenge. Protocols such as ACMG-AMP provide recommendations on the use of tools, but indicate the most cited ones, such as PolyPhen and SIFT, without another evaluation metric for this indication (BORGES, 2021).

A major problem is that protein sequence variations generate different performances for each predictor (RICHARDS et al., 2015). Thus, it is interpreted that the adequacy of a predictor depends on the specific problem to be addressed (UÇAR et al., 2020). This diversity often results in divergent results, requiring strategies to balance these results (UÇAR et al., 2020). PANTHER gene classification based on evolutionary relationships and functional groups provides a valuable resource for variant analysis (MI et al., 2021). When combined with *in silico* prediction tools, this classification can reinforce the robustness of the analysis and classification of variants.

Methods

Dataset

Data was obtained from *dbNSFP* (Database for Non-Synonymous Functional Predictions), version *v4.2a* (LIU et al., 2020) through the link https://drive.google.com/file/d/1OfAx1SrJVPPNYWfDpQd43S9Aqro3C6aA/view, on April 2021. Variant annotation, ClinVar classification and the in-built prediction for 38 tools were used. The prediction tools were: SIFT_pred, SIFT4G_pred, Polyphen2_HDIV_pred, Polyphen2_HVAR_pred, LRT_pred, MutationTaster_pred, MutationAssessor_pred, FATHMM_pred, PROVEAN_pred, VEST4_score, MetaSVM_pred, MetaLR_pred, MetaRNN_pred, M-CAP_pred, REVEL_score, MutPred_score, MVP_score, MPC_score, PrimateAI_pred, DEOGEN2_pred, BayesDel_addAF_pred, BayesDel_noAF_pred, ClinPred_pred, LIST-S2_pred, CADD_phred, CADD_phred_hg19, DANN_score, fathmm-MKL_coding_pred, fathmm-XF_coding_pred, Eigen-phred_coding, Eigen-PC-phred_coding,

GenoCanyon_score, integrated_fitCons_score, GM12878_fitCons_score, H1-hESC_fitCons_score, HUVEC_fitCons_score, LINSIGHT. Due to the size of .tsv files, information was retrieved via *bash Linux*.

Gene selection

Given the uneven number of variants among different genes, with many genes presenting few variants, only genes with at least 100 variants considered pathogenic or benign were included. First, we downloaded ClinVar data through FTP, and used a python algorithm to sort genes with at least 100 variants with any of the given classifications: "*Pathogenic", "Likely Pathogenic", "Benign"* and *"Likely Pathogenic".* We retrieved 149 genes that were then searched on *dbNSFP*, but only 86 of them had concordant data.

Data cleaning

Since *dbNSFP* uses the prediction score for each *in silico* tool, we standardized that any score that could be considered pathogenic, according to the tools' documentation, would be equal to 1, whereas any benign score would be equal to 0. The same standardization was applied to ClinVar classification. If values were "*pathogenic", "likely pathogenic", "drug response", "risk factor",* or any combination of that, it would be considered 1. If values were *"benign", "likely benign", "protective"*, it would be 0. In the case of contradictory classification, for example, *"pathogenic and likely benign"* for the same variant, it was considered a VUS and excluded from analysis.

Gene classification

Genes were grouped using PANTHER (Protein ANalysis Through Evolutionary Relationships) Protein Class (PC). Some classes were further grouped to increase the number of genes in each class, as shown in table 1.

Table 1. Grouping of PANTHER class.

Original PANTHER CLASS	New PANTHER CLASS
Extracellular matrix structural protein	*Extracellular matrix protein*
Voltage-gated ion channel	*Ion channel*
Microtubule binding motor protein	*Microtubule binding protein*
Non motor microtubule binding protein	
Serine/Threonin protein kinase receptor	*Serine/Threonin protein kinase*
Non-receptor Serine/Threonin protein kinase	

Statistical analysis

Sensitivity, specificity, and accuracy were evaluated using an algorithm in python, considering ClinVar classification as the golden standard. Kappa value was calculated using the scikit-learn library, with the cohen_kappa_score method.

The comparison between *in silico* tools and PC-based gene groups was performed in R using the Youden index. Initially, a linear regression model was tested, using the *lm()* function, in which we considered the response variable Youden against the predicting variables *PANTHER_CLASS* and *programs*. We then used the *emmeans* package to calculate the estimated marginal means (EMM). Multiple comparison effects were adjusted using the Benjamini-Hochberg method and results were visualized and interpreted with the *cld()* function that groups statistically similar results with common letters.

Results

After restricting our analysis to genes with at least 100 variants classified in *ClinVar* and *dbNSFP* we obtained 86 genes that fulfilled this criterion. After classifying these genes in PANTHER, only 39 genes were present in classes with more than 3 genes. Therefore, the 39 genes included in this analysis were: *FBN1, BRCA1, DMD, COL4A5, SCN1A, USH2A, COL3A1, ABCA4, COL1A1, COL1A2, SCN2A, MECP2, SCN5A, CDKL5, KCNH2, ABCC6, NIPBL, DNAH5, BMPR2, COL2A1, COL7A1, DNAH11, SCN8A, CACNA1A, COL4A3, ACVRL1, DYNC2H1, SPAST, COL6A3,*

BRAF, GRIN2B, VWF, COL4A4, CHD7, ABCB1, CACNA1H, GRIN2A, VHL, KCNT1, MBD5 e FBN2. Out of these, the gene with the higher number of variants was *FBN1* with 1298 variants, being 1283 "*pathogenic*" and 15 "*benign*". On the other hand, the gene with least variants was *FBN2,* with 101, being 51 "*pathogenic*" and 50 "*benign*".

These 39 genes were grouped in 8 classes: extracellular matrix protein, ion channel, chromatin/chromatin-binding, or -regulatory protein, microtubule binding protein, serine/threonine protein kinase receptor proteins, transmembrane signal receptor, ATP-binding cassette (ABC) transporter e ubiquitin-protein ligase. The class *extracellular matrix protein* had 11 genes, whereas four classes had only three genes, as shown in the table 2.

Table 2. Number of genes and variants per PANTHER class.

PANTHER CLASS	**Number of genes**	**Mean number of pathogenic variants**	**Mean number of benign variants**
Extracellular matrix protein	11	214.00	12.67
Ion channel	8	101.50	60.00
Chromatin/chromatin-binding, or -regulatory protein	4	316.27	30.82
Microtubule binding protein	4	195.75	43.50
Serine/Threonin protein kinase receptor proteins	3	125.50	43.00
Transmembrane signal receptor	3	152.00	20.33
ATP-binding cassette (ABC) transporter	3	159.33	60.67
Ubiquitin-protein ligase	3	493.00	106.00

Figure 1 shows the output of statistical comparisons for one class of genes (ATP-binding cassette transporter). A large overlap between predictors can be seen, as represented by same letters in column ".group". Still, the best predictor for this class of genes would be *BayesDel_addAF* (the *BayesDel* predictor that uses also allele frequency), *ClinPred* and *MetaRNN* with *emmeans* values of 0.75956 (0.46238 – 1.05675), 0.75661 (0.45942 – 1.05379) and 0.72163 (0.42445 – 1.01882), respectively.

programs	PANTHER_CLASS	emmean	SE	df	lower.CL	upper.CL	.group
BayesDel_addAF_pred	ATP-binding cassette (ABC) transporter	0,759565999	0,0922709534	1100	0,4623798926	1,0567521054	a
ClinPred_pred	ATP-binding cassette (ABC) transporter	0,756607616	0,0922709534	1100	0,4594215096	1,0537937224	a
MetaRNN_pred	ATP-binding cassette (ABC) transporter	0,7216317173	0,0922709534	1100	0,4244456109	1,0188178238	ab
PROVEAN_pred	ATP-binding cassette (ABC) transporter	0,709071589	0,1130083769	1100	0,3450944293	1,0730487487	abc
REVEL_score	ATP-binding cassette (ABC) transporter	0,6701221027	0,0922709534	1100	0,3729359962	0,9673082091	abc
BayesDel_noAF_pred	ATP-binding cassette (ABC) transporter	0,6633011993	0,0922709534	1100	0,3661150929	0,9604873058	abc
VEST4_score	ATP-binding cassette (ABC) transporter	0,6499017687	0,0922709534	1100	0,3527156622	0,9470878751	abc
MetaSVM_pred	ATP-binding cassette (ABC) transporter	0,6073654477	0,0922709534	1100	0,3101793412	0,9045515541	abcd
DEOGEN2_pred	ATP-binding cassette (ABC) transporter	0,5710447945	0,1130083769	1100	0,2070676348	0,9350219542	abcde
Polyphen2_HDIV_pred	ATP-binding cassette (ABC) transporter	0,5220517087	0,0922709534	1100	0,2248656022	0,8192378151	abcdef
SIFT_pred	ATP-binding cassette (ABC) transporter	0,5190463835	0,1130083769	1100	0,1550692238	0,8830235432	abcdef
SIFT4G_pred	ATP-binding cassette (ABC) transporter	0,508308004	0,0922709534	1100	0,2111218976	0,8054941104	abcdef
MutationTaster_pred	ATP-binding cassette (ABC) transporter	0,5080042103	0,0922709534	1100	0,2108181039	0,8051903168	abcdef
Polyphen2_HVAR_pred	ATP-binding cassette (ABC) transporter	0,4986636203	0,0922709534	1100	0,2014775139	0,7958497268	abcdef
MetaLR_pred	ATP-binding cassette (ABC) transporter	0,4836255157	0,0922709534	1100	0,1864394092	0,7808116221	abcdef
LIST-S2_pred	ATP-binding cassette (ABC) transporter	0,459283821	0,0922709534	1100	0,1620977146	0,7564699274	bcdefg
MutationAssessor_pred	ATP-binding cassette (ABC) transporter	0,4255549413	0,0922709534	1100	0,1283688349	0,7227410478	cdefg
fathmm-XF_coding_pred	ATP-binding cassette (ABC) transporter	0,3492943343	0,0922709534	1100	0,0521082279	0,6464804408	defgh
fathmm-MKL_coding_pred	ATP-binding cassette (ABC) transporter	0,3398294993	0,0922709534	1100	0,0426433929	0,6370156058	defghi
GenoCanyon_score	ATP-binding cassette (ABC) transporter	0,268091125	0,0922709534	1100	-0,029094981	0,5652772314	efghij
Eigen-phred_coding	ATP-binding cassette (ABC) transporter	0,2553880977	0,0922709534	1100	-0,041798009	0,5525742041	efghij
LRT_pred	ATP-binding cassette (ABC) transporter	0,240158456	0,0922709534	1100	-0,05702765	0,5373445624	fghij
Eigen-PC-phred_coding	ATP-binding cassette (ABC) transporter	0,1708836137	0,0922709534	1100	-0,126302493	0,4680697201	ghijk
MVP_score	ATP-binding cassette (ABC) transporter	0,1070652127	0,0922709534	1100	-0,190120894	0,4042513191	hijkl
DANN_score	ATP-binding cassette (ABC) transporter	0,073380172	0,0922709534	1100	-0,223805934	0,3705662784	hijkl
PrimateAI_pred	ATP-binding cassette (ABC) transporter	0,055222876	0,0922709534	1100	-0,24196323	0,3524089824	ijkl
M-CAP_pred	ATP-binding cassette (ABC) transporter	0,0182057895	0,1130083769	1100	-0,34577137	0,3821829492	ijkl
LINSIGHT	ATP-binding cassette (ABC) transporter	9,214851E-15	0,0922709534	1100	-0,297186106	0,2971861064	jkl
CADD_phred	ATP-binding cassette (ABC) transporter	1,44329E-15	0,0922709534	1100	-0,297186106	0,2971861064	jkl
CADD_phred_hg19	ATP-binding cassette (ABC) transporter	3,330669E-16	0,0922709534	1100	-0,297186106	0,2971861064	jkl
MPC_score	ATP-binding cassette (ABC) transporter	-5,55112E-16	0,0922709534	1100	-0,297186106	0,2971861064	jkl
GERP++_RS	ATP-binding cassette (ABC) transporter	-2,10942E-15	0,0922709534	1100	-0,297186106	0,2971861064	jkl
FATHMM_pred	ATP-binding cassette (ABC) transporter	-0,030543406	0,1130083769	1100	-0,394520566	0,3334337537	jkl
integrated_fitCons_score	ATP-binding cassette (ABC) transporter	-0,07367416	0,0922709534	1100	-0,370860266	0,2235119468	kl
HUVEC_fitCons_score	ATP-binding cassette (ABC) transporter	-0,076166034	0,0922709534	1100	-0,37335214	0,2210200728	kl
MutPred_score	ATP-binding cassette (ABC) transporter	-0,110218309	0,0922709534	1100	-0,407404415	0,1869677978	kl
H1-hESC_fitCons_score	ATP-binding cassette (ABC) transporter	-0,120507795	0,0922709534	1100	-0,417693901	0,1766783118	kl
GM12878_fitCons_score	ATP-binding cassette (ABC) transporter	-0,122208656	0,0922709534	1100	-0,419394762	0,1749774504	l

Figure 1 – Output of the par-au-par comparison for the *ATP-binding cassette (ABC) transporter* class.

For the *chromatin/chromatin-binding or -regulatory protein* class, *MetaRNN* and *REVEL* had the highest *emmeans,* with 0.88729 (0.62992 – 1.14466) and 0.84761 (0.59024 – 1.10498), respectively. It is worth noticing that *BayesDel_addAF* and *ClinPred,* although statistically different, had similar *emmeans* values, respectively 0.75818 (0.50081 – 1.01555) and 0.74431 (0.48694 – 1.00168).

For the class *extracellular matrix protein* predictors *ClinPred* and *MetaRNN* had *emmeans* values of 0.91120 (0.75600 – 1.0664) and 0.89346 (0.73826 – 1.04866), respectively. Again, *BayesDel_addAF,* although with a statistical difference, had good *emmeans* values of 0.87059 (0.71539 – 1.02579). For *ion channel* the best *emmeans* values were for predictors *MetaRNN* and *ClinPred* with, respectively 0.6991 (0.51711 – 0.88108) and 0.68013 (0.49814 – 0.86212). *BayesDel_addAF* is in the same situation as for the latter class, with *emmeans* of 0.65299 (0.47101 – 0.83499).

In class *microtubule binding protein, BayesDel_addAF*, *ClinPred, and MetaRNN* had *emmeans* values of 0.90685 (0.64948 – 1.16422), 0.89898 (0.64161 – 1.15635), and 0.83886 (0.58148 – 1.09623), respectively. For class *serine/threonine protein kinase receptor proteins* predictors *MetaRNN* and *ClinPred,* with *emmeans* of 0.89889 (0.60171 – 1.19608) and 0.844145 (0.54696 – 1.14133), respectively, were the best whereas *REVEL,* although with a statistical difference, had *emmeans* of 0.82305 (0.52586 – 1.12024).

The class with the largest number of predictors without statistical difference was the *transmembrane signal receptor*, with the following *emmeans: BayesDel_addAF* 0.80253 (0.50534 – 1.09972), *MetaRNN* 0.76287 (0.46568 – 1.060051812), *REVEL* 0.75283 (0.45565 – 1.05002), *BayesDel_noAF* 0.73484 (0.43766 – 1.03203), *PROVEAN* 0.70843 (0.34445 – 1.07240), *ClinPred* 0.70173 (0.40454 – 0.99891), and *VEST4* 0.68519 (0.38801 – 0.98238). Finally, for the class *ubiquitin-protein ligase* predictors *ClinPred*, *MetaRNN, and BayesDel_addAF* presented the best performance, with *emmeans* of 0.80628 (0.50909 – 1.10346), 0.7754 (0.47822 – 1.07259), and 0.71427 (0.41709 – 1.01145), respectively.

The predictors' accuracy and kappa value for the different gene classes were analyzed (figure 2). For *ATP-binding cassette (ABC) transporter, BayesDel_addAF* had a mean accuracy and *kappa* value of 0.842 and 0.399, whereas these values were 0.839 and 0.428 for *ClinPred* and 0.793 and 0.404 for *Meta_RNN.* For the class *chromatin/chromatin-binding. or -regulatory protein* predictor *Meta_RNN* had mean accuracy of 0.932 and mean *kappa* 0.706. For *REVEL,* mean accuracy was 0.940 and mean *kappa was* 0.850, whereas for *BayesDel_addAF* they were 0.908 and 0.761, and for *ClinPred* 0.838 and 0.618.

For the *extracellular matrix protein* class, *ClinPred* has means of 0.968 and 0.896, whereas *Meta_RNN* had 0.972 and 0.882 and *BayesDel_addAF* 0.969 and 0.867. For class *ion channel*, predictor *Meta_RNN* had means of accuracy and *kappa*

value of 0.941 and 0.738, respectively, whereas predictor *ClinPred* had 0.927 and 0.721 and *BayesDel_addAF* 0.934 and 0.692. For *microtubule binding protein* predictor *BayesDel_addAF* had mean accuracy and *kappa* value of 0.942 and 0.849, predictor *ClinPred* had 0.954 and 0.866, and *Meta_RNN* 0.936 and 0.821.

In the class *serine/threonine protein kinase* the best ranked predictors, *Meta_RNN*, *ClinPred, and REVEL,* had the following means of accuracy and *kappa* values: 0.979 and 0.891; 0.961 and 0.865; 0.960 and 0.786, respectively. For *transmembrane signal receptor* the results were: *BayesDel_addAF* (accuracy = 0.908 and *kappa* = 0.779), *Meta_RNN* (0.899 and 0.779), *REVEL* (0.846 and 0.674), *BayesDel_noAF* (0.880 and 0.711), *PROVEAN* (0.570 and 0.470), *ClinPred* (0.884 and 0.729), and *VEST4* (0.891 and 0.722). Finally, for class *ubiquitin-protein ligase*, predictor *ClinPred* had means of accuracy and *kappa* value of 0.954 and 0.808, whereas *Meta_RNN* had 0.934 and 0.725 and *BayesDel_addAF* 0.947 and 0.672.

Figure 2 – Accuracy and *kappa* values for best-ranked predictors in each gene class.

Discussion

In this work, we compared 38 predictors of variants against 39 genes grouped in 8 different gene groups in order to evaluate a possible relationship between the quality of prediction assertiveness in different gene groups. The choice of the dbNSFP database was made due to the ease of having all the predictions grouped in a single tool, together with the classification of a trusted repository, such as ClinVar. The dbNSFP v4.2a version presents many missing data of predictions for certain predictors against some variants, and this becomes a difficulty for statistical evaluations due to the non-equal amount of data for each predictor.

ClinVar is a curated database, however, some of the classifications may have 'conflicting interpretations', even though they are not classified as such. The need for a precise classification led to a reduced number of genes with at least 100 classified variants. The inconsistency with dbNSFP further reduced the number of genes considered in this study. But it was the Protein Class classification in PANTHER that cut the number of genes down to 39. Since our hypothesis was that predictors would perform differently according to gene categories, we had to group genes and the use of an existing tool seemed an unbiased strategy. However, we needed groups with at least three genes. This led to a reduced dataset, but in favor of better statistical analysis.

Our results show a repetition of predictors with good statistical values, as ClinPred, BayesDel_addAF, and Meta_RNN stood out in all 8 gene classes that we evaluated, showing good values of *emmeans*, accuracy and *kappa* values. Although this recurrence has occurred, there are differences between the best predictors for each class. In addition, we can also analyze that the predictors H1-hESC_fitCons and GM12878_fitCons were almost always listed as the worst predictors of each group. However, we also noticed an overlap between predictors with similar statistical significance. For example, for *ATP-binding cassette (ABC) transporter*, the 10 best listed predictors show a lot of statistical overlap, as predictors were grouped from 'a' to 'abcdef', where equal letters represent lack of statistical difference.

The best ranked predictors were ClinPred, BayesDel_addAF, and Meta_RNN. ClinPred predictor is an algorithm that works based on machine learning mainly applying concepts from Random Forest (ALIREZAIE et al., 2018). BayesDel_addAF uses sophisticated statistical methods to score the classifications with a naïve

Bayesian approach (FENG, 2017), whereas Meta_RNN uses deep learning techniques to classify variants. All these are recently developed algorithms, with more up-to-date analytical methods, relying on artificial intelligence, which seems to offer a better analysis for classifying variants. We can also assume that predictors that use more specific data are not as suitable for general predictions. This is the case of H1-hESC_fitCons and GM12878_fitCons, which are fitCons predictors that use data referring to H1 human embryonic stem cells (H1 hESCs) and lymphoblastoid cells (GM12878), respectively.

Despite the importance of using predictors for validation, it is recommended that *in silico* analyzes be used in conjunction with other evidence, such as functional data and clinical information, for a more reliable classification of missense variants (RICHARDS et al., 2015). A multidisciplinary approach, which incorporates both molecular analysis and clinical interpretation, is essential for the correct identification of variants in patients with genetic diseases.

References

ALIREZAIE, N. et al. ClinPred: Prediction Tool to Identify Disease-Relevant Nonsynonymous Single-Nucleotide Variants. **The American Journal of Human Genetics**, v. 103, n. 4, p. 474–483, out. 2018.

AUTON, A. et al. A global reference for human genetic variation. **Nature**, v. 526, n. 7571, p. 68–74, 1 out. 2015.

CASTIGLIA, D.; ZAMBRUNO, G. Mutation Mechanisms. **Dermatologic Clinics**, v. 28, n. 1, p. 17–22, jan. 2010.

CHEN, S. et al. A genome-wide mutational constraint map quantified from variation in 76,156 human genomes. [s.d.].

CHOE, H. et al. Molecular Diagnostics. **Journal of the American Academy of Orthopaedic Surgeons**, v. 23, p. S26–S31, abr. 2015.

CLAUSSNITZER, M. et al. A brief history of human disease genetics. **Nature**, v. 577, n. 7789, p. 179–189, 9 jan. 2020.

FAIRLEY, S. et al. The International Genome Sample Resource (IGSR) collection of open human genomic variation resources. **Nucleic Acids Research**, v. 48, n. D1, p. D941–D947, 8 jan. 2020.

FENG, B.-J. PERCH: A Unified Framework for Disease Gene Prioritization. **Human Mutation**, v. 38, n. 3, p. 243–251, mar. 2017.

GULLEY, M. L. et al. Clinical Laboratory Reports in Molecular Pathology. **Archives of Pathology & Laboratory Medicine**, v. 131, n. 6, p. 852–863, 1 jun. 2007.

LANDRUM, M. J. et al. ClinVar: improvements to accessing data. **Nucleic Acids Research**, v. 48, n. D1, p. D835–D844, 8 jan. 2020.

LAPPALAINEN, T. et al. Genomic Analysis in the Age of Human Genome Sequencing. **Cell**, v. 177, n. 1, p. 70–84, mar. 2019.

LEK, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. **Nature**, v. 536, n. 7616, p. 285–291, 17 ago. 2016.

LINDEBOOM, R. G. H.; SUPEK, F.; LEHNER, B. The rules and impact of nonsense-mediated mRNA decay in human cancers. **Nature Genetics**, v. 48, n. 10, p. 1112–1118, 12 out. 2016.

LIU, X. et al. dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. **Genome Medicine**, v. 12, n. 1, p. 103, 2 dez. 2020.

MCLAREN, W. et al. The Ensembl Variant Effect Predictor. **Genome Biology**, v. 17, n. 1, p. 122, 6 dez. 2016.

MI, H. et al. PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API. **Nucleic Acids Research**, v. 49, n. D1, p. D394–D403, 8 jan. 2021.

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

NYKAMP, K. et al. Sherloc: a comprehensive refinement of the ACMG–AMP variant classification criteria. **Genetics in Medicine**, v. 19, n. 10, p. 1105–1117, out. 2017.

PÂMELLA BORGES. **Comparação de ferramentas in silico para avaliação de patogenicidade de variantes missense**. Porto Alegre : Universidade Federal do Rio Grande do Sul, 2021.

RICHARDS, S. et al. 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. **Genetics in Medicine**, v. 17, n. 5, p. 405–424, maio 2015.

SHERRY, S. T. dbSNP: the NCBI database of genetic variation. **Nucleic Acids Research**, v. 29, n. 1, p. 308–311, 1 jan. 2001.

STONE, E. A.; SIDOW, A. Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. **Genome Research**, v. 15, n. 7, p. 978–986, jul. 2005.

SUNYAEV, S. Prediction of deleterious human alleles. **Human Molecular Genetics**, v. 10, n. 6, p. 591–597, 1 mar. 2001.

TANG, H.; THOMAS, P. D. Tools for Predicting the Functional Impact of Nonsynonymous Genetic Variation. **Genetics**, v. 203, n. 2, p. 635–647, 1 jun. 2016.

THOMAS, P. D. et al. PANTHER: A Library of Protein Families and Subfamilies Indexed by Function. **Genome Research**, v. 13, n. 9, p. 2129–2141, set. 2003.

UÇAR, M. K. et al. The Effect of Training and Testing Process on Machine Learning in Biomedical Datasets. **Mathematical Problems in Engineering**, v. 2020, p. 1–17, 13 maio 2020.

WANG, K.; LI, M.; HAKONARSON, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. **Nucleic Acids Research**, v. 38, n. 16, p. e164–e164, 1 set. 2010.

WONG, Y. K. E. et al. The applications of big data in molecular diagnostics. **Expert Review of Molecular Diagnostics**, v. 19, n. 10, p. 905–917, 3 out. 2019.

APÊNDICE A: Informações do dbNSFP v4.2a sobre os preditores (retirado do arquivo README.txt)

Preditor	Informações
SIFT_pred	If SIFTori is smaller than 0.05 (rankscore>0.39575) the corresponding nsSNV is predicted as "D(amaging)"; otherwise it is predicted as "T(olerated)".
SIFT4G_pred	If SIFT4G is < 0.05 the corresponding nsSNV is predicted as "D(amaging)"; otherwise it is predicted as "T(olerated)".
Polyphen2_HDIV_pred	Polyphen2 prediction based on HumDiv, "D" ("probably damaging", HDIV score in [0.957,1] or rankscore in [0.55859,0.91137]), "P" ("possibly damaging", HDIV score in [0.454,0.956] or rankscore in [0.37043,0.55681]) and "B" ("benign", HDIV score in [0,0.452] or rankscore in [0.03061,0.36974]). Score cutoff for binary classification is 0.5 for HDIV score or 0.38028 for rankscore, i.e. the prediction is "neutral" if the HDIV score is smaller than 0.5 (rankscore is smaller than 0.38028), and "deleterious" if the HDIV score is larger than 0.5 (rankscore is larger than 0.38028)
Polyphen2_HVAR_pred	Polyphen2 prediction based on HumVar, "D" ("probably damaging", HVAR score in [0.909,1] or rankscore in [0.65694,0.97581]), "P" ("possibly damaging", HVAR in [0.447,0.908] or rankscore in [0.47121,0.65622]) and "B" ("benign", HVAR score in [0,0.446] or rankscore in [0.01493,0.47076]). Score cutoff for binary classification is 0.5 for HVAR score or 0.48762 for rankscore, i.e. the prediction is "neutral" if the HVAR score is smaller than 0.5 (rankscore is smaller than 0.48762), and "deleterious" if the HVAR score is larger than 0.5 (rankscore is larger than 0.48762).
LRT_pred	LRT prediction, D(eleterious), N(eutral) or U(nknown), which is not solely determined by the score.
MutationTaster_pred	MutationTaster prediction, "A" ("disease_causing_automatic"), "D" ("disease_causing"), "N" ("polymorphism") or "P" ("polymorphism_automatic"). The score cutoff between "D" and "N" is 0.5 for MTnew and 0.31733 for the rankscore
MutationAssessor_pred	MutationAssessor's functional impact of a variant predicted functional, i.e. high ("H") or medium ("M"), or predicted non-functional, i.e. low ("L") or neutral ("N"). The MAori score cutoffs between "H" and "M", "M" and "L", and

	"L" and "N", are 3.5, 1.935 and 0.8, respectively. The rankscore cutoffs between "H" and "M", "M" and "L", and "L" and "N", are 0.9307, 0.52043 and 0.19675, respectively
FATHMM_pred	If a FATHMMori score is <=-1.5 (or rankscore >=0.81332) the corresponding nsSNV is predicted as "D(AMAGING)"; otherwise it is predicted as "T(OLERATED)".
PROVEAN_pred	If PROVEANori <= -2.5 (rankscore>=0.54382) the corresponding nsSNV is predicted as "D(amaging)"; otherwise it is predicted as "N(eutral)".
VEST4_score	VEST 4.0 score. Score ranges from 0 to 1. The larger the score the more likely the mutation may cause functional change.
MetaSVM_pred	Prediction of our SVM based ensemble prediction score,"T(olerated)" or "D(amaging)". The score cutoff between "D" and "T" is 0. The rankscore cutoff between "D" and "T" is 0.82257.
MetaLR_pred	Prediction of our MetaLR based ensemble prediction score,"T(olerated)" or "D(amaging)". The score cutoff between "D" and "T" is 0.5. The rankscore cutoff between "D" and "T" is 0.81101.
MetaRNN_pred	Prediction of our MetaRNN based ensemble prediction score,"T(olerated)" or "D(amaging)". The score cutoff between "D" and "T" is 0.5. The rankscore cutoff between "D" and "T" is 0.6149.
M-CAP_pred	Prediction of M-CAP score based on the authors' recommendation, "T(olerated)" or "D(amaging)". The score cutoff between "D" and "T" is 0.025.
REVEL_score	REVEL is an ensemble score based on 13 individual scores for predicting the pathogenicity of missense variants. Scores range from 0 to 1. The larger the score the more likely the SNP has damaging effect.
MutPred_score	General MutPred score. Scores range from 0 to 1. The larger the score the more likely the SNP has damaging effect
MVP_score	A pathogenicity prediction score for missense variants using deep learning approach. The range of MVP score is from 0 to 1. The larger the score, the more likely the variant is pathogenic. The authors suggest thresholds of 0.7 and 0.75 for separating damaging vs tolerant variants in constrained genes (ExAC pLI >=0.5) and non-constrained genes (ExAC pLI<0.5), respectively.

MPC_score	A deleteriousness prediction score for missense variants based on regional missense constraint. The range of MPC score is 0 to 5. The larger the score, the more likely the variant is pathogenic.
PrimateAI_pred	Prediction of PrimateAI score based on the authors' recommendation, "T(olerated)" or "D(amaging)". The score cutoff between "D" and "T" is 0.803.
DEOGEN2_pred	Prediction of DEOGEN2 score based on the authors' recommendation, "T(olerated)" or "D(amaging)". The score cutoff between "D" and "T" is 0.5.
BayesDel_addAF_pred	Prediction of BayesDel_addAF score based on the authors' recommendation, "T(olerated)" or "D(amaging)". The score cutoff between "D" and "T" is 0.0692655.
BayesDel_noAF_pred	Prediction of BayesDel_noAF score based on the authors' recommendation, "T(olerated)" or "D(amaging)". The score cutoff between "D" and "T" is -0.0570105.
ClinPred_pred	Prediction of ClinPred score based on the authors' recommendation, "T(olerated)" or "D(amaging)". The score cutoff between "D" and "T" is 0.5.
LIST-S2_pred	Prediction of LIST-S2 score based on the authors' recommendation, "T(olerated)" or "D(amaging)". The score cutoff between "D" and "T" is 0.85.
CADD_phred	CADD phred-like score. This is phred-like rank score based on whole genome CADD raw scores. Please refer to Kircher et al. (2014) Nature Genetics 46(3):310-5 for details. The larger the score the more likely the SNP has damaging effect. Please note the following copyright statement for CADD: "CADD scores (http://cadd.gs.washington.edu/) are Copyright 2013 University of Washington and Hudson-Alpha Institute for Biotechnology (all rights reserved) but are freely available for all academic, non-commercial applications.
CADD_phred_hg19	CADD phred-like score using the hg19 model. This is phred-like rank score based on whole genome CADD raw scores. Please refer to Kircher et al. (2014) Nature Genetics 46(3):310-5 for details. The larger the score the more likely the SNP has damaging effect. Please note the following copyright statement for CADD: "CADD scores (http://cadd.gs.washington.edu/) are Copyright 2013 University of Washington and Hudson-Alpha Institute for Biotechnology (all rights reserved) but are freely available for all academic, non-commercial applications.

DANN_score	DANN is a functional prediction score retrained based on the training data of CADD using deep neural network. Scores range from 0 to 1. A larger number indicate a higher probability to be damaging.
Fathmm-MKL_coding_pred	If a fathmm-MKL_coding_score is >0.5 (or rankscore >0.28317) the corresponding nsSNV is predicted as "D(AMAGING)"; otherwise it is predicted as "N(EUTRAL)".
Fathmm-XF_coding_pred	If a fathmm-XF_coding_score is >0.5, the corresponding nsSNV is predicted as "D(AMAGING)"; otherwise it is predicted as "N(EUTRAL)"
Eigen-phred_coding	Eigen score in phred scale.
Eigen-PC_phred_coding	Eigen PC score in phred scale.
GenoCanyon_score	A functional prediction score based on conservation and biochemical annotations using an unsupervised statistical learning. (doi:10.1038/srep10576)
Integrated_fitCons_score	fitCons score predicts the fraction of genomic positions belonging to a specific function class (defined by epigenomic "fingerprint") that are under selective pressure. Scores range from 0 to 1, with a larger score indicating a higher proportion of nucleic sites of the functional class the genomic position belong to are under selective pressure, therefore more likely to be functional important. Integrated (i6) scores are integrated across three cell types (GM12878, H1-hESC and HUVEC)
GM12878_fitCons_score	fitCons score predicts the fraction of genomic positions belonging to a specific function class (defined by epigenomic "fingerprint") that are under selective pressure. Scores range from 0 to 1, with a larger score indicating a higher proportion of nucleic sites of the functional class the genomic position belong to are under selective pressure, therefore more likely to be functional important. GM12878 fitCons scores are based on cell type GM12878. More details can be found in doi:10.1038/ng.3196.
H1-hESC_fitCons_score	fitCons score predicts the fraction of genomic positions belonging to a specific function class (defined by epigenomic "fingerprint") that are under selective pressure. Scores range from 0 to 1, with a larger score indicating a higher proportion of nucleic sites of the functional class the genomic position belong to are under selective pressure, therefore more likely to be functional important. GM12878 fitCons scores are based on cell type H1-hESC. More details can be found in doi:10.1038/ng.3196.
HUVEC_fitCons_score	fitCons score predicts the fraction of genomic positions belonging to a specific function class (defined by epigenomic "fingerprint") that are under selective pressure. Scores range from 0 to 1, with a larger score indicating a higher proportion of nucleic sites of the functional class the genomic position belong to are under

	selective pressure, therefore more likely to be functional important. GM12878 fitCons scores are based on cell type HUVEC. More details can be found in doi:10.1038/ng.3196.
LINSIGHT	"The LINSIGHT score measures the probability of negative selection on noncoding sites" Details refer to doi:10.1038/ng.3810.
GERP++_RS	GERP++ RS score, the larger the score, the more conserved the site. Scores range from -12.3 to 6.17.