Universidade Federal do Rio Grande do Sul

Comparação de ferramentas *in silico* para avaliação de patogenicidade de variantes *missense*

Pâmella Borges

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

Orientadora: Profa. Dra. Ursula Matte

Porto Alegre, Março de 2021

Instituições e Fontes Financiadoras

Este trabalho foi realizado no Laboratório Células, Tecidos e Genes no Centro de Pesquisa Experimental do Hospital de Clínicas de Porto Alegre.
As fontes financiadoras foram o Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e o Fundo de Incentivo à Pesquisa e Eventos (FIPE) do Hospital de Clínicas de Porto Alegre.

Agradecimentos

A Professora Dra. Ursula Matte pela oportunidade de desenvolver esse projeto, por toda a confiança e orientação ao longo de toda a minha iniciação científica e agora no mestrado, bem como por todas as contribuições para o meu crescimento intelectual e profissional.

A todos os colegas do laboratório Células, Tecidos e Genes, obrigada pelo companheirismo e ensinamentos.

Aos meus amigos, pelo apoio durante o percurso, pelos conselhos, conversas, ânimo e amizade ao longo de todos esses anos.

Aos meus pais, por todos os ensinamentos ao longo da vida, pelo apoio incondicional, pelas conversas, abraços e lágrimas derramadas em conjunto.

A minha irmã, por sempre estar comigo. Por todo carinho, aventuras, arteirices e momentos que compartilhamos ao longo desses quinze anos.

Sumário

Resumo

A análise de variantes representa um processo crítico no diagnóstico molecular e os programas *in silico* são especialmente usados quando nenhuma informação de literatura está disponível. Diferentes programas avaliam os possíveis efeitos gerados pela mutação, considerando critérios como conservação de aminoácidos e nucleotídeos, local e importância estrutural da alteração e fatores bioquímicos. Entretanto, esses critérios recebem pesos diferentes em cada programa e isso pode impactar diferentes grupos de proteínas de forma desigual. Portanto, saber qual programa é melhor para um gene específico representa uma maneira de aumentar a confiança na avaliação dos preditores. Porém, a obtenção desta informação implica em extensa revisão da literatura para avaliação dos programas. O processamento de linguagem natural, uma técnica de mineração de texto, pode ser empregado como forma de automatizar a busca na literatura de informações sobre as variantes e assim poder comparar os preditores com uma base maior de informações. Portanto, o objetivo deste trabalho é desenvolver uma ferramenta para comparar preditores *in silico* de acordo com o tipo de proteína. Uma revisão dos preditores mais e menos citados na literatura questiona os critérios de escolha das ferramentas para avaliar variantes *missense* e discorre sobre as características dos principais preditores. Para estabelecer o *workflow* para a ferramenta proposta e obter dados de validação, foi realizada a comparação de 34 ferramentas *in silico* utilizando dados curados manualmente para o gene *IDUA.* O desempenho dos preditores foi avaliado em dois grupos de variantes, um criado a partir de critérios mais rigorosos (108 variantes) e o outro a partir de critérios menos rigorosos (160 variantes). Os mesmos três preditores (BayesDel, PONP2 e ClinPred) apresentaram melhores desempenhos nos dois grupos e foram usados para avaliar 462 variantes de significado incerto. Finalmente, o *pipeline* de análise utilizado nesta comparação está sendo integrado com um algoritmo de mineração de texto, ainda em desenvolvimento, que realiza a extração automatizada das variantes relatadas na literatura com a sua interpretação clínica. Espera-se que a automatização de todo o processo possa ser usada para a escolha dos melhores preditores para cada situação específica.

Abstract

Variant analysis represents a critical process in molecular diagnosis and *in silico* programs are traditionally used when no literature information is available. Different programs evaluate the possible effects generated by the variant, considering criteria such as conservation of amino acids and nucleotides, location and structural importance of the alteration, and biochemical factors. However, these criteria are given different weights in each program and this can have an uneven impact on different groups of proteins. Therefore, knowing which program is best for a specific gene is a way to increase confidence in predictor evaluation. However, obtaining this information implies an extensive literature review to evaluate the programs. Natural language processing, a text mining technique, can be used as a way to automate the literature search for information about variants and thus allow the comparison of predictors with a larger informational base. Therefore, the aim of this work is to develop a tool to compare *in silico* predictors according to the protein type. A review of predictors' most and least cited in the literature question the criteria for choosing tools to assess missense variants and discuss the characteristics of the main predictors. To establish the workflow and obtain validation data for the proposed tool, 34 programs were compared *in silico* using manually cured data for the *IDUA* gene. The predictors' performance was evaluated in two groups of variants, one created stricter criteria (108 variants) and the other less stringent criteria (160 variants). The same three predictors (BayesDel, PONP2, and ClinPred) had the best performance in both groups and were used to evaluate 462 variants of uncertain significance. Finally, the analysis pipeline used in this comparison is being integrated with a text mining algorithm, still under development, which performs the automated extraction of the variants reported in the literature with its clinical interpretation. It is expected that the automation of the entire process can be used to choose the best predictors for each specific situation.

Introdução

O diagnóstico molecular é um conjunto de técnicas amplamente aplicadas, poderosas e sensíveis usadas para identificar marcadores biológicos em um genoma e proteoma (Choe et al. 2015). A análise de variantes, uma importante etapa do processo de diagnóstico molecular, apresenta crescente complexidade devido o avanço das técnicas moleculares como *whole-exome sequencing* (WES) e *whole-genome sequencing* (WGS) que geram um elevado número de dados para serem analisados, comparados e principalmente, interpretados.

As diretrizes e padrões para interpretação de variantes foram publicadas em 2015 quando o Colégio Americano de Genética Médica (ACMG) e a Associação de Patologia Molecular (AMP) se reuniram para compilar 28 regras baseadas nas experiências de cada laboratório (Richards et al. 2015). Em 2017, um grupo de pesquisadores insatisfeitos com aspectos das normas da ACMG-AMP, principalmente no que diz respeito à subjetividade da interpretação, revisou essas normas e apresentou mudanças na estrutura de avaliação, desenvolvendo o Sherloc (Nykamp et al. 2017). Apesar de apresentarem divergências, ambos os protocolos concordam que as evidências relatadas na literatura, embora necessitem ser tratadas com atenção, são indícios muito importantes na avaliação de uma variante.

Idealmente, o impacto funcional das variantes deveria ser determinado a partir de estudos experimentais, por exemplo, usando a mutagênese sítio dirigida. Além disso, estudos observacionais de análise de segregação em um número significativo de indivíduos também pode contribuir para essa avaliação. Mas considerando que as informações de um único genoma podem chegar a 200 GB e gerar um *Variant Call Format* (VCF) de 125 MB, com 3 milhões de variantes cada, é compreensível que a análise experimental não consiga acompanhar a descoberta e a anotação de novas variantes (Wong et al., 2019). Assim, é comum que não sejam encontrados relatos na literatura sobre variantes ou mesmo que exista divergência entre as interpretações. Nesses casos, avaliações *in silico,* apesar de não validadas clinicamente, são uma importante ferramenta para formar um nível de evidência (Richards et al. 2015; Nykamp et al. 2017). Inclusive, a análise *in silico* é empregada muitas vezes como a única ferramenta de avaliação de impacto das variantes.

A análise *in silico* é menos utilizada para variantes do tipo *nonsense*, para as quais existe certo consenso sobre a patogenicidade. Neste sentido é importante ressaltar que tal concepção pode estar incorreta, dependendo da localização da alteração. A existência de um códon de parada prematuro geralmente resulta em proteínas truncadas e rapidamente degradadas (Castiglia and Zambruno, 2010). Quando isso ocorre a montante do último éxon, é iniciado um conservado processo de vigilância celular que reconhece complexos de junção de exon ou protetores de ligação de RNA a jusante do ribossomo, chamado de processo de *nonsense mediated mRNA decay* (NMD), que degrada o mRNA. Apesar de conservada, existem mecanismos de escape da via, como variantes muito próximas ao códon de iniciação, que podem ter a tradução iniciada à jusante do códon de parada prematuro, ou a “regra dos 50-55 nucleotídeos” que diz que apenas variantes *nonsense* dentro dessa faixa na junção exon-exon são reconhecidas (Dyle et al, 2019; Lindeboom et al, 2016). Assim, nem todas as variantes *nonsense* podem ser tratadas como perda de função, pois além desses mecanismos, uma alteração no último exon pode não apresentar uma perda significativa para a funcionalidade da proteína.

Variantes de *splice*, sinônimas, *frameshift* e *in-frame* apresentam um crescimento no número de ferramentas de análise. Os preditores de *splice* costumam se basear no cálculo de entropia (Jian, 2013), dados de expressão e RNA-seq (Jaganathan et al., 2019). Deste grupo, tem o maior número de ferramentas específicas para avaliação. Existem poucos preditores específicos para as variantes sinônimas, *frameshift* e *in-frame*. A principal limitação das variantes sinônimas é a falta de dados experimentais de validação (Zeng and Bromberg, 2019). Já as variantes *in-frame* costumam ser avaliadas por programas que também avaliam variantes *missense*. Assim como as variantes *nonsense*, as de *frameshift* são pouco avaliadas e geralmente consideradas patogênicas pelo impacto causado na funcionalidade da proteína. Entretanto, também não são tratadas como perda de função quando presentes último exon (Lindeboom et al, 2016).

Já as variantes *missense* representam um desafio para a análise e não podem ser consideradas diretamente patogênicas (Nykamp et al. 2017). Para essas variantes é necessária uma avaliação por preditores computacionais, 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 and Thomas, 2016). Estratégias para avaliação da patogenicidade de variantes *missense* existem desde a década de 1970 e são o foco do presente trabalho.

Devido ao grande número de ferramentas disponíveis, o primeiro capítulo desta dissertação apresenta uma revisão de 34 preditores encontrados na literatura, comparando os mais e os menos utilizados, com base no número de citações. O capítulo estabelece um paralelo entre os dois grupos e avalia as estratégias utilizadas por cada preditor.

Neste contexto de ampla oferta de possibilidades, a escolha do preditor nem sempre segue parâmetros objetivos. No entanto, sabe-se que as performances dos preditores variam amplamente de acordo com a sequência proteica avaliada (Richards et al. 2015), tanto pelas estratégias utilizadas para comparação, quanto pelos grupos de treinamento dos algoritmos. Métodos diferentes geram resultados diferentes e existem diversas estratégias de aprendizado de máquina (*machine learning-ML*) disponíveis. A escolha do método utilizado deve variar de acordo com o problema analisado (Uçar et al, 2019). Outra importante etapa na elaboração de uma avaliação com ML é o conjunto de treinamento. Os dados presentes nesse conjunto devem ser independentes dos dados de validação para não gerar sobreajuste e influenciam diretamente no desempenho dos programas. Outro possível viés é a utilização de dados não balanceados. Uma boa representação dos dados permite que os algoritmos sejam treinados igualmente para checar todos os possíveis cenários, enquanto dados desbalanceados podem tendenciar a predição de um cenário sobre outro. Por exemplo, o maior número de variantes patogênicas no grupo de treinamento pode levar os preditores a classificar variantes benignas como patogênicas. Quanto melhor a representação dos dados, melhor o resultado final e, em casos de disparidade, deve-se utilizar alguma das estratégias disponíveis para ajustar os dados não balanceados (Uçar et al., 2019).

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. Como as proteínas podem ser agrupadas em famílias de acordo com as suas funções e estruturas, algo passível de se considerar é que proteínas da mesma família ou subfamília sejam avaliadas de forma parecida pelas ferramentas. Considerando as características de cada família proteica, preditores diferentes podem avaliar melhor um grupo em relação a outro devido às estratégias utilizadas na sua análise. Assim é interessante comparar não apenas as diferentes proteínas, mas se proteínas da mesma família ou subfamília apresentam similaridades de avaliação. Conhecer essas informações é importante para melhorar o desempenho e a confiança das avaliações *in silico* existentes, além de guiar novos programas.

Para realizar essa comparação é necessária a construção de um banco de dados de variantes com significado conhecido e subsequentes testes e avaliações de performance nos diversos preditores. Visando padronizar a realização dessas análises, o segundo capítulo desta dissertação apresenta uma comparação de 51 predições para 160 variantes do gene *IDUA* curadas manualmente da literatura, bem como a avaliação de 426 variantes de significado incerto encontradas em bancos de dados populacionais pelos preditores com melhores desempenhos.

No entanto, para gerar o banco de variantes com significado conhecido, como feito neste trabalho, é necessário que cada pesquisador leia e avalie um grande número de artigos relacionados ao gene de interesse. Isso torna a criação do banco algo trabalhoso e, principalmente, demorado. Realizar essa curadoria manualmente para um grande número de genes em um curto período de tempo é impossível. Portanto, uma estratégia de automatização é necessária.

Com o desenvolvimento da ciência da computação, diversas tarefas e processos foram automatizados. A tradução e interpretação de uma linguagem natural é um processo complexo que está em difusão desde os anos 1950. Muitas estratégias já foram desenvolvidas para realizar essa tarefa, mas uma em especial vem ganhando destaque ao longo dos anos: o *deep learning*. O deep learning tem como ideia o aprendizado pelo modelo de representações intermediárias úteis, que apresentam vários níveis de representação para serem otimizados (Hirschberg and Manning, 2015).

O processamento de linguagem natural (*natural language processing-NLP*) apresenta diversas etapas e métodos de mineração de texto para aprender, compreender e produzir conteúdo de linguagem humana (Esteva et al, 2019), extraindo não somente as informações relevantes para o usuário, mas também significado essas informações contextualmente. Considerando essa estratégia, o terceiro e último capítulo da dissertação apresenta uma aplicação do processo na busca de automatizar a comparação e escolha dos preditores. O trabalho está em desenvolvimento e busca avaliar as performances dos preditores em diferentes genes, tentando entender se existem estratégias mais adequadas para diferentes grupos proteicos.

Objetivo Geral

O objetivo do trabalho é fazer a comparação da performance de preditores de variantes *missense* entre e intra diferentes grupos protéicos, utilizando como base um banco de variantes curadas.

Objetivos específicos

1. Realizar uma revisão da literatura dos preditores de variantes missense mais e menos citados na literatura.
2. Estabelecer as etapas de comparação de desempenho dos preditores com um grupo de variantes do gene *IDUA*;
3. Automatizar a criação das bases de dados para comparar as predições de diferentes ferramentas entre e intra grupos proteicos utilizando um algoritmo de processamento de linguagem natural.

Capítulo 1

No capítulo é apresentado um artigo de revisão sobre preditores de variantes *missenses*. O artigo foca nos principais preditores encontrados na literatura, baseado no número de citações e estabelece um paralelo entre os preditores mais e menos citados.

O artigo está em fase final de elaboração para submissão no periódico *Bioinformatics*.

***In silico* tools for predicting pathogenicity of missense variants: are the most cited, the most accurate?**

Pâmella Borges[1,2,3], Ursula Matte[1,2,3,4].

Affiliations:

[1]Cell, Tissue and Gene Laboratory, Clinicas Hospital of Porto Alegre (HCPA), Rio Grande do Sul, Brazil

[2]Bioinformatics Core, Experimental Research Centre, HCPA, Rio Grande do Sul, Brazil.

[3]Graduate Programme in Genetics and Molecular Biology, Federal University of Rio Grande do Sul (UFRGS).

[4]Department of Genetics, UFRGS, Porto Alegre, Brazil.

Correspondence: pamella.bor@gmail.com

Introduction

The advance of genomic analysis contributed to generating new information and bioinformatics consolidation. With the development of DNA-sequencing technologies and the growing number of new variants starting in the 1990s and exploding in the 2000s with large-scale projects, the in silico predictors gained space and particular relevance. The discovery of a large number of variants modified how we process and understand modern genetics, opening new investigative routes on variant interpretation and how they relate to our health.

Variant interpretation is directly associated with human genetics, and there is particular interest in distinguishing functionally neutral variants from those that contribute to disease (Ng PC, Henikoff S, 2002). Several variants may alter inherited traits by affecting gene transcription, pre-mRNA splicing, or protein translation, thus impacting protein expression or function (Castiglia and Zambruno, 2010). The most complex variant type to assign a functional effect to is a missense mutation, in which a change of amino acid is caused by a single nucleotide substitution (Zhang et al., 2012). Indeed, tools for classifying these variants started to be developed in the early 1970s. The Grantham score, developed in 1974, compares the composition, polarity, and molecular volume differences between amino acids. Estimating the extent to which observed exchanges could be explained by conservation, the article presents a matrix with fixed values for each amino acid substitution overall proteins and beginning generalized comparison and evaluation of missense variants (Grantham, 1974).

In the 2000s, bioinformatics methods were developed considering evolutionary conservation, structural effects, or a combination of both (Tang and Thomas, 2016). The first predictors were based on alignments and differed in punctuation matrices and the probability determination (SIFT (Ng PC et al., 2001), PolyPhen (Sunyaev S et al., 2001), and PANTHER (Thomas PD et al., 2003). These programs relied on the availability of sequences in other species, which was not the case in the early days of genomics. Shortly after, structural predictors were introduced (MAPP) (Stone and Sidow, 2005). However, this type of predictor presented some limitations because they depended on the three-dimensional protein structure, which was also scarce (Tang and Thomas, 2016). Combined methods were developed to circumvent these problems, with the need for a combined analysis of two types of data. The machine learning algorithms appeared as a crucial informatic method for selecting the best ways to predict the variants' pathogenicity based on the training group, considering combination methods.

We currently have several prediction algorithms available to choose from when evaluating the pathogenicity of a missense variant. So much so that we often find ourselves questioning which tool best predicts outcomes for the protein of interest.

Often authors choose predictors previously used, which become the most commonly cited tools. We compared the total number of citations of the original articles of 34 predictors [SIFT (Kumar P et al., 2009; Ng PC et al., 2003; Ng PC et al., 2001), SIFT4G (Vaser R et al., 2016), PolyPhen2 (Adzhubei I et al., 2013; Adzhubei IA et al., 2010; Sunyaev S et al., 2001), LRT (Chun S et al., 2009), MutationAssessor (Reva B et al., 2011; Reva B et al., 2007), FATHMM (Shihab HA et al., 2014; Shihab HA et al., 2013; Shihab HA et al., 2013), MetaSVM/LR (Dong C et al., 2015), CADD (Rentzsch P et al., 2019; Kircher M et al., 2014), VEST (Carter H et al., 2013), PROVEAN (Choi Y et al., 2015; Choi Y et al., 2012], fitConsx4 (Gulko B et al, 2015), REVEL (Ioannidis NM et al., 2016), DANN (Quang D et al., 2015), MutationTaster2 (Schwarz JM et al., 2010], M-CAP (Jagadeesh KA et al., 2016), LINSIGHT (Huang YF et al., 2017), MutPred (Li B et al., 2009), PrimateAI (Sundaram L et al., 2019), BayesDel (Feng BJ, 2017), ClinPred (Alirezaie N et al., 2018), LIST-S2 (Malhis N et al., 2020), GenoCanyon (Lu Q et al., 2015), Eigen and Eigen-PC (Ionita-Laza I et al., 2016), PhD-SNP (Capriotti E et al., 2006), PANTHER (Mi H et al., 2013; Thomas PD et al., 2003), SNPs&GO (Capriotti E et al., 2013), PSNPE (Bendl J et al., 2016; Bendl J et al., 2014), FunSeq2 (Fu Y et al., 2014), GWAVAE (Ritchie GR et al., 2014), SuSPect (Yates CM et al., 2014), PMut (López-Ferrando V et al., 2017; Ferrer-Costa C et al., 2005), CONDEL (González-Pérez A et al., 2011), PON-P2 (Niroula A et al., 2015), SNAP2 (Bromberg Y et al., 2007)] in Pubmed, Google Scholar and Web of Science from 2001 to 2020. The sum of articles citing each predictor in all databases was considered to calculate the percentage of citations for each predictor. For instance, if a given predictor has 10, 20, and 30 citations in each database, it has 60 citations in total, even though they might be repeated across databases. If the sum of all citations for all predictors in all databases is 600, then in this particular example, the predictor would have 10% of citations.

Among predictors with the highest number of citations, PolyPhen and SIFT standout with 26.22 and 25.42% of citations, whereas the next are CADD with 9.29%, PANTHER 7.90%, PROVEAN with 5.55%, and MutationTaster2 with 4.86%. All others have less than 4% of citations. On the other end are LISTS2 with 0.002%, PrimateAI 0.20%, GenoCanyon 0.17%, ClinPred 0.07% and BayesDel 0.04%. All other predictors have more than 0.20%. The absolute number of citations and the percentage for each predictor are present in Table 1 for the predictor with the highest and lowest percentages. A complete list can be found in supplementary table 1.

As expected, the overall number of citations is higher for older predictors. To better evaluate more recent predictors, only citations from 2015 to 2020 were compared (Supplementary figure 1). Four manuscripts describing three predictors (PolyPhen2, 2010; SIFT, 2009; SIFT, 2003; CADD, 2014) have over 250 citations, as shown in Figure 1a. These three predictors belong to the six most cited groups, considering all periods (Figure 1b). Again, Polyphen and SIFT are among the first described predictors, showing that predictors are continually cited from their publication. Figure 1c shows the citation over time for the group of less cited articles.

Table 1: Articles per year for most and least cited predictors comparing 34 predictors in three different databases (Pubmed, Google Scholar, and Web of Science).

Predictors	Year	Titles	Pubmed	Google Scholar	Web of science	% of total cited
Articles of the most cited predictors						
SIFT	2001	Predicting deleterious amino acid substitutions	1,025	2,374	1,628	4.520
PolyPhen	2001	Prediction of deleterious human alleles	436	1,210	878	2.270
SIFT	2003	SIFT: Predicting amino acid changes that affect protein function	1,874	4,588	3,081	8.581
PANTHER	2003	PANTHER: A Library of Protein Families and Subfamilies Indexed by Function	1,110	2,543	1,697	4.811
SIFT	2009	Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm	2,826	5,825	4,236	11.588
Polyphen2	2010	A method and server for predicting damaging missense mutations	5,069	10,298	7,479	20.543
MutationTaster2	2010	MutationTaster evaluates disease-causing potential of sequence alterations	1,118	2,361	1,766	4.716
PROVEAN	2012	Predicting the Functional Effect of Amino Acid Substitutions and Indels	903	2,011	899	3.429
Polyphen2	2013	Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2	939	2,022	-	2.662
PANTHER	2013	PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees	727	1445	1009	2.860
CADD	2014	A general framework for estimating the relative pathogenicity of human genetic variants	2,010	3,885	2,698	7.727
PROVEAN	2015	PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels	454	1,098	746	2.066
CADD	2019	CADD: predicting the deleteriousness of variants throughout the human genome	304	701	441	1.300
Articles of the least cited predictors						
GenoCanyon	2015	A Statistical Framework to Predict Functional Non-Coding Regions in the Human Genome Through Integrated Analysis of Annotation Data	46	90	53	0.170
BayesDel	2017	PERCH: A Unified Framework for Disease Gene Prioritization	13	19	16	0.043
PrimateAI	2018	Predicting the clinical impact of human mutation with deep neural networks	44	106	62	0.191
ClinPred	2018	ClinPred: Prediction Tool to Identify Disease-Relevant Nonsynonymous Single-Nucleotide Variants	16	36	26	0.070
LIST-S2	2020	LIST-S2: taxonomy based sorting of deleterious missense mutations across species	1	1	0	0.002

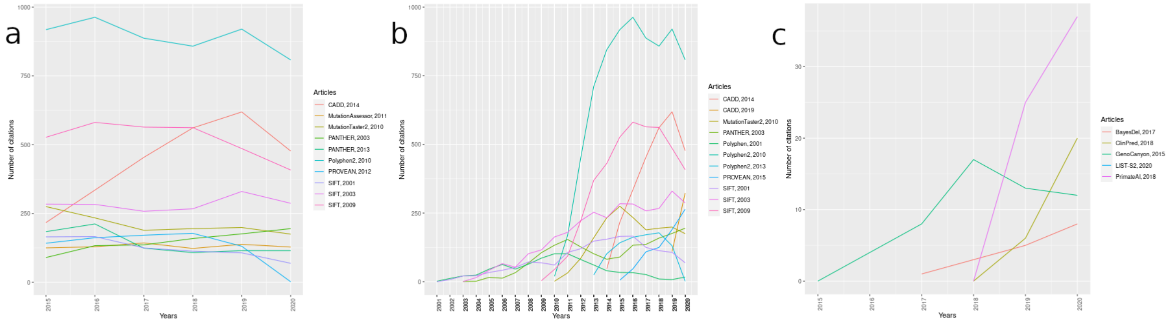


Figure 1: Number of citations over time: a) Most cited articles; b) all articles belong to six predictors more cited; c) all articles belong to five predictors less cited. Note the difference in both the x and y-axis in graphs.

A text mining R script from Edureka's Data Science was applied to all 46 articles describing the *in silico* predictors. The most frequent seven words were present over 200 times (Figure 2a): *substitution*, *sift*, *score*, *annotation*, *deleterious*, *code*, and *SNPs*. The appearance of Sift highlights its use as a reference tool. A word cloud was created for words with frequencies higher than 60 (Figure 2b). Interestingly, the term "PolyPhen" is present in the word cloud but in the group with frequencies lower than 100 times.

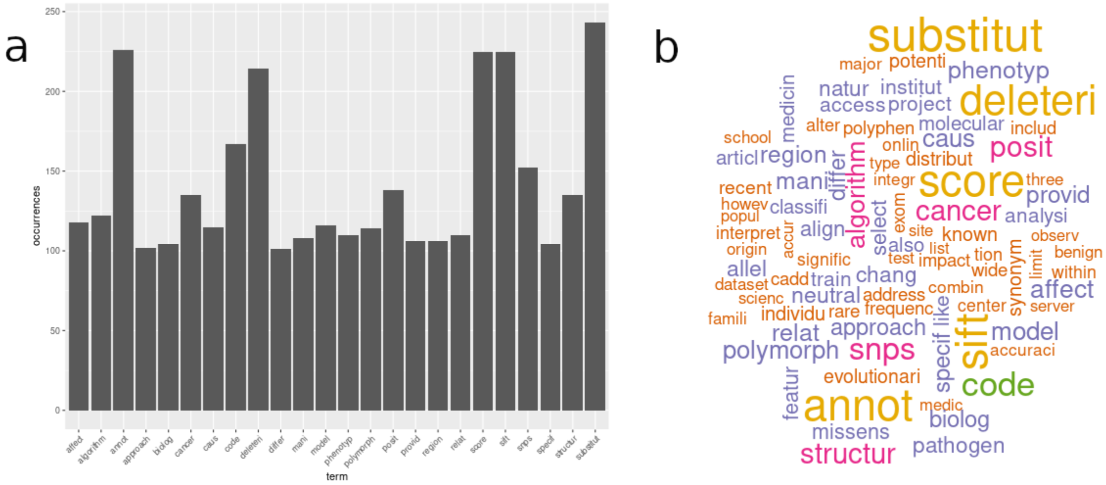


Figure 2: Text mining analysis from all predictor articles. a) Words with frequency higher than 100 times. b) Wordcloud with all terms with frequency higher than 60 times.

Although we cannot directly infer that these predictors are the most and least used in clinical practice, we may assume that the number of citations may correlate to clinical use. Indeed, laboratory clinical reports often present predictions based on SIFT and PolyPhen data, thus contributing to their popularity. As the choice of tools is a critical point in bioinformatics analysis, especially in the absence of literature information supporting the analysis, the recurrence of the same predictors leads us to question why this occurs.

Table 2 summarizes the 11 predictors shown before with information about each one's development to understand the continuous preference better. The most popular tools have a web interface, while PrimateAI, BayesDel, and ClinPred need local installation in the UNIX environment, thus requiring bioinformatics skills . This little difference puts these tools at a disadvantage because the user needs to install and run the script using a code line.

CADD, LIST-S2, GenoCanyon, and PrimateAI do not have an established cut-off point. This is debatable, as some authors believe there is no ideal cut-off for binarization deleterious/benign (LIST-S2). Some reasons they present is the loss of information represented for binary cut-off and that other factors influence the cut-off choice, as the severity of the phenotype, the inheritance pattern of the disease or available for curation or experimental follow-up of variants (CADD), beyond the fact that the output is an imperfect probability, that may vary from gene to gene. LIST-S2 recommends in its documentation that each researcher determines the cut-off value that is best suited to each analysis.

Also, the programs present diversified inputs. The user can obtain predictions starting with chromosomal position and nucleotide change or with protein position and amino acid change. Depending on the program, this choice influences the result. MutationTaster2 also asks for the transcript id and some bases flanking the variant position. About robust analysis, all predictors realize batch upload.

Table 2: Summary of the top seven predictors cited in the literature.

Name	Complete name	Training data	Information used	Prediction model	Score
PolyPhen2 (HDIV)	Polymorphism Phenotyping v2	5564 Mendelian disease mutations and 7539 divergence SNVs from close mammalian homolog proteins	eight sequence-based and three structure-based predictive features	naive Bayes classifier	Tolerated: 0.0 to 0.15 Possibly damaging: 0.15 to 0.85 Damaging: 0.85 to 1.0
PolyPhen2 (HVAR)		22196 disease associated SNVs and 21119 common SNVs	same as above	same as above	
SIFT	Sorting Intolerant From Tolerant	1750 deleterious and 2254 tolerant nsSNVs of E. coli LacI gene	sequence homology based on PSI-BLAST	position specific scoring matrix	Deleterious: 0.0 to 0.05 Tolerated (benign): 0.05 to 1.0
MutationTaster2	-	SNVs from 1000 G (1000 Genomes Project), HGMD	conservation, splice site, mRNA features, protein features; regulatory features	naive Bayes classifier	D: disease causing (>0.5) A: disease causing automatic N: polymorphism (<0.5) P: polymorphism automatic
PROVEAN	Protein Variation Effect Analyzer	SNVs from UniProt/HUMSAVAR	sequence homology	Delta alignment score	deleterious < -2.5 < neutral
CADD	Combined Annotation Dependent Depletion	16,627,775 “observed” variants and 49,407,057 “simulated” variants	63 annotations (949 features)	linear kernel support vector machine	Literature used >20 as cut off for consider pathogenic
PANTHER	Protein ANalysis THrough Evolutionary Relationships	HGMD and dbSNP	evolutionarily related sequences	HMM modeling of protein families	Neutral: 0 to 0.5 Disease: 0.5 to 1
PrimateAI	-	dataset of ~380,000 common missense variants from humans and six non-human primate species	36 layers of convolutions, consisting of roughly 400,000 trainable parameters	deep neural networks	0 (less pathogenic) to 1 (more pathogenic)
BayesDel	-	39,395 pathogenic variants and 39,978 neutral rom ClinVar and UniProtKB	combined multiple deleteriousness predictors to create an overall score	naïve Bayesian	Universal cutoff value (0.069 with MaxAF, -0.057 without MaxAF)
ClinPred	-	11,082 variants from ClinVar, with 7,059 labeled as benign and 4,023 labeled as pathogenic	Incorporates allele frequencies from gnomAD and 16 individual prediction scores	random forest (cforest) and gradient boosted decision tree (xgboost) models	Benign: 0 to 0.5 Pathogenic: 0.5 to 1
LIST-S2	Local Identity and Shared Taxa - species specific	26,708 benign and 20,015 deleterious	ExAC, gnomAD, UniProt (been associated with diseases and cancer), ClinVar	High Local identity Pairwise Sequence Alignment to all protein sequences; identifies the most relevant homologies; estimates the potential deleteriousness of mutations based on taxonomy distance of species	Deleterious < 0.84 Benign > 0.85
GenoCanyon	-	Genomic data for all the 22 annotations were downloaded from the UCSC Genome Browser. 12,801,840 locations were used to estimate the parameters.	49 parameters	unsupervised statistical learning	cutoff point 0.5

An important aspect that contributes to some predictors being more used than others is the ACMG-AMP recommendation (Richards et al., 2015), which cites what they considered the “most used predictors” in 2015. All top predictors found in our analysis are present in this list. As seen in Figure 1b, the number of citations grows before 2015, although it is impossible to rule out the visibility that being part of this list gives to these programs or measure the impact on citations directly.

Something also relevant is the citation of papers describing the first algorithms by subsequent programs. This happens in two cases: predictors use others for statistical comparisons and to demonstrate the new algorithms' performance. For example, BayesDel includes three of the most cited predictors (PolyPhen2 (Adzhubei et al., 2010), SIFT (Kumar et al., 2009) and Mutation Taster (Schwarz et al., 2010)) to create its overall score, and ClinPred uses SIFT, PolyPhen-2, MutationAssessor, PROVEAN, and CADD in feature analysis. Other predictors have similar comparisons and contribute to increasing the number of citations of previous and more recognized programs.

Tools use different strategies to predict the pathogenicity of variants. Some predictors agree in the method of choice, as PolyPhen2, MutationTaster2, and BayesDel. These machine learning algorithms predict pathogenicity based on the Naive Bayes classifier, a probabilistic model contingent on Bayes theorem that calculates the probability of some “a” event happening, given that “b” has occurred.

PolyPhen2 is based on several features comprising the sequence, phylogenetic and structural information characterizing the substitution. This supervised machine-learning has two models: HumDiv and HumVar. The differences between training datasets make HumDiv preferred for evaluating rare alleles, dense mapping of regions identified by genome-wide association studies, and natural selection analysis, while HumVar is better for Mendelian diseases. The user can choose, or the program will run HumDiv as default. Both report the probability that a given variant is damaging, estimate false and true positive rates, and give a qualitative prediction, as benign, possibly damaging, or probably damaging (Adzhubei I et al., 2013; Adzhubei IA et al., 2010; Sunyaev S et al., 2001).

MutationTaster2 calculates the alteration probabilities to be either a disease mutation or a harmless polymorphism considering evolutionary conservation, splice-site changes, loss of protein features, and changes that might affect the amount of mRNA. The probability varies from 0 to 1, and values close to 1 indicate 'high security' of the prediction (Schwarz JM et al., 2010). BayesDel creates an overall score combining six deleteriousness predictors and three conservation scores considered mutually independent. The scores are calculated as a weighted product of likelihood ratios, and in the end, the model was optimized for the area under the receiver operating characteristic curve (Feng BJ, 2017).

SIFT (Sorting Intolerant From Tolerant) evaluates position-specific information obtained for alignment to predict pathogenicity. It is based on conserved sequences in a protein family and the type of amino acid change. Conserved positions are considered as intolerant to most changes, although poorly conserved positions have different scores. All values obtained are directly related to the diversity of the sequences in the alignment and how similar or not are the amino acid changes (Kumar P et al., 2009; Ng PC et al., 2003; Ng PC et al., 2001), SIFT4G (Vaser R et al., 2016).

Integrating diverse annotations correlated with molecular functionality and pathogenicity and trained by a support vector machine, Combined Annotation–Dependent Depletion (CADD) combines them to a single measure (C score) for each variant. Considering that variants that reduce organismal fitness are depleted by natural selection, CADD compares the annotation of simulated variants with nearly fixed alleles in humans (Rentzsch P et al., 2019; Kircher M et al., 2014).

PANTHER uses a statistical model for scoring the “functional likelihood” of amino acid substitutions using evolutionarily related sequences. It calculates the score of a single amino acid at a particular position or the likelihood of converting one amino acid to another and compares it with scores from known variants to determine the pathogenicity (Mi H et al., 2013; Thomas PD et al., 2003).

PROVEAN (Protein Variation Effect Analyzer) predicts a delta score computed from a substitution matrix (information on the substitution frequency and chemical properties of 20 amino acid residues), gap penalties, percent identity threshold for sequence clustering, number of top clusters generated, and can also be determined by the neighborhood that surrounds the site of variation. The web server supports PROVEAN Protein, Protein Batch, and Genome Variants functions (Choi Y et al., 2015; Choi Y et al., 2012).

Random Forest is a machine learning classifier that uses a large number of individual trees. Each tree has a class prediction, and the choice of model happens for the class with the most votes. Although not frequently used in the most cited predictors, this method has become common in newer predictors. ClinPred uses gradient boosted decision tree (xgboost) models trained using either balanced or equal weights. The prediction used 13 individual prediction scores and three conservations scores (SIFT, PolyPhen, LRT, MutationAssessor, PROVEAN, CADD, DANN, PhastCons, fitCons, GERP, PhyloP, and SiPhy) plus allele frequencies of variants from gnomAD (Alirezaie N et al., 2018).

Two versions of the deep learning network were trained to discriminate variants in PrimateAI: one with common unlabeled variants, as a control for aligning ability between the species and humans, trinucleotide context, and sequencing coverage, and one with the full benign labeled dataset. Both have human variants, but the second also has six non-human primate species. As input, the program receives a sequence with a length of 51 amino acids centered at a variant, used to generate matrices from multiple sequence alignments of 99 vertebrates. The output is solvent accessibility networks and secondary structure with the missense variant substituted at the central position (Sundaram L et al., 2019).

LIST-S2 (Local Identity and Shared Taxa - species-specific) have three modules assembled hierarchically: position mutation module, position module, and mutation module. The first determines if a mutation occurs in homolog close or distantly related. The second establishes if the variation occurs in related species. The third estimates the likelihood of changing the reference by the variant. The final score combines the weighted scores from the two first modules with the third module. The features that determine the score are: variant share among taxa, average values of the top one-third of the shared taxa profile vector, average variant shared taxa of all 19 possible variations, and the general amino acid swap-ability matrix (Malhis N et al., 2020).

GenoCanyon treats the conservation measures, and the biochemical signals as consequences of genomic function, where 1 is functional, and 0 is not. It models, as function consequences, 22 conditionally independent annotations selected due to their functional impacts that are relatively well studied and easier to model and correspond to either conservation score or biochemical activity. The prediction score is the posterior probability of the functional potential at this location (Lu Q et al., 2015).

It is not straightforward to detect a trend on why some predictors are used more often than others. To be web-based or a local program does not seem to matter. As the Naive Bayes method, the same prediction models can be used by the most and least cited programs. The training set and the information used do not differ among programs and can even be the same for different predictors. However, unexpectedly, results can be quite contradictory. A recent study from our group evaluated 3,040 missense variants in five predictors (Polyphen2, MutPred, PROVEAN, SIFT, and REVEL), and only 44.54% of them had a total consensus in the five programs, whereas 31.84% had one disagreement and 23.62% had two disagreements (Borges et al., 2019).

In the end, it is not clear if the selection of a predictor is a conscious choice made by each researcher based on objective criteria or the following of a trend in the literature, set by the earlier use of predictors. Given the absence of consensus among predictors (Borges et al., 2019; Guidugli et al., 2018; Rodrigues et al., 2015) and the lack of studies comparing predictor's performance for specific proteins, the second hypothesis seems to be more likely.

References

Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet. 2013;Chapter 7:Unit7.20. doi:10.1002/0471142905.hg0720s76

Adzhubei IA, Schmidt S, Peshkin L, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010;7(4):248-249. doi:10.1038/nmeth0410-248

Alirezaie N, Kernohan KD, Hartley T, Majewski J, Hocking TD. ClinPred: Prediction Tool to Identify Disease-Relevant Nonsynonymous Single-Nucleotide Variants. Am J Hum Genet. 2018;103(4):474-483. doi:10.1016/j.ajhg.2018.08.005

Bendl J, Musil M, Štourač J, Zendulka J, Damborský J, Brezovský J. PredictSNP2: A Unified Platform for Accurately Evaluating SNP Effects by Exploiting the Different Characteristics of Variants in Distinct Genomic Regions. PLoS Comput Biol. 2016;12(5):e1004962. Published 2016 May 25. doi:10.1371/journal.pcbi.1004962

Bendl J, Stourac J, Salanda O, et al. PredictSNP: robust and accurate consensus classifier for prediction of disease-related mutations. PLoS Comput Biol. 2014;10(1):e1003440. doi:10.1371/journal.pcbi.1003440

Borges P, Pasqualim G, Giugliani R, Vairo F, Matte U. Estimated prevalence of mucopolysaccharidoses from population-based exomes and genomes. Orphanet J Rare Dis. 2020;15(1):324. Published 2020 Nov 18. doi:10.1186/s13023-020-01608-0

Bromberg Y, Rost B. SNAP: predict effect of non-synonymous polymorphisms on function. Nucleic Acids Res. 2007;35(11):3823-3835. doi:10.1093/nar/gkm238

Capriotti E, Calabrese R, Casadio R. Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. Bioinformatics. 2006;22(22):2729-2734. doi:10.1093/bioinformatics/btl423

Capriotti E, Calabrese R, Fariselli P, Martelli PL, Altman RB, Casadio R. WS-SNPs&GO: a web server for predicting the deleterious effect of human protein variants using functional annotation. BMC Genomics. 2013;14 Suppl 3(Suppl 3):S6. doi:10.1186/1471-2164-14-S3-S6

Carter H, Douville C, Stenson PD, Cooper DN, Karchin R. Identifying Mendelian disease genes with the variant effect scoring tool. BMC Genomics. 2013;14 Suppl 3(Suppl 3):S3. doi:10.1186/1471-2164-14-S3-S3

Castiglia D, Zambruno G. Mutation mechanisms. Dermatol Clin. 2010;28(1):17-22. doi:10.1016/j.det.2009.10.002

Choi Y, Chan AP. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. Bioinformatics. 2015;31(16):2745-2747. doi:10.1093/bioinformatics/btv195

Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the functional effect of amino acid substitutions and indels. PLoS One. 2012;7(10):e46688. doi:10.1371/journal.pone.0046688

Chun S, Fay JC. Identification of deleterious mutations within three human genomes. Genome Res. 2009;19(9):1553-1561. doi:10.1101/gr.092619.109

Dong C, Wei P, Jian X, et al. Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. Hum Mol Genet. 2015;24(8):2125-2137. doi:10.1093/hmg/ddu733

Feng BJ. PERCH: A Unified Framework for Disease Gene Prioritization. Hum Mutat. 2017;38(3):243-251. doi:10.1002/humu.23158

Ferrer-Costa C, Gelpí JL, Zamakola L, Parraga I, de la Cruz X, Orozco M. PMUT: a web-based tool for the annotation of pathological mutations on proteins. Bioinformatics. 2005;21(14):3176-3178. doi:10.1093/bioinformatics/bti486

Fu Y, Liu Z, Lou S, et al. FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. Genome Biol. 2014;15(10):480. doi:10.1186/s13059-014-0480-5

González-Pérez A, López-Bigas N. Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. Am J Hum Genet. 2011;88(4):440-449. doi:10.1016/j.ajhg.2011.03.004

Grantham R. Amino acid difference formula to help explain protein evolution. Science. 1974;185(4154):862-864. doi:10.1126/science.185.4154.862

Guidugli L, Shimelis H, Masica DL, et al. Assessment of the Clinical Relevance of BRCA2 Missense Variants by Functional and Computational Approaches. Am J Hum Genet. 2018;102(2):233-248. doi:10.1016/j.ajhg.2017.12.013

Gulko B, Hubisz MJ, Gronau I, Siepel A. A method for calculating probabilities of fitness consequences for point mutations across the human genome. Nat Genet. 2015;47(3):276-283. doi:10.1038/ng.3196

Huang YF, Gulko B, Siepel A. Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. Nat Genet. 2017;49(4):618-624. doi:10.1038/ng.3810

Ioannidis NM, Rothstein JH, Pejaver V, et al. REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. Am J Hum Genet. 2016;99(4):877-885. doi:10.1016/j.ajhg.2016.08.016

Ionita-Laza I, McCallum K, Xu B, Buxbaum JD. A spectral approach integrating functional genomic annotations for coding and noncoding variants. Nat Genet. 2016;48(2):214-220. doi:10.1038/ng.3477

Jagadeesh KA, Wenger AM, Berger MJ, et al. M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. Nat Genet. 2016;48(12):1581-1586. doi:10.1038/ng.3703

Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014;46(3):310-315. doi:10.1038/ng.2892

Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc. 2009;4(7):1073-1081. doi:10.1038/nprot.2009.86

Li B, Krishnan VG, Mort ME, et al. Automated inference of molecular mechanisms of disease from amino acid substitutions. Bioinformatics. 2009;25(21):2744-2750. doi:10.1093/bioinformatics/btp528

López-Ferrando V, Gazzo A, de la Cruz X, Orozco M, Gelpí JL. PMut: a web-based tool for the annotation of pathological variants on proteins, 2017 update. Nucleic Acids Res. 2017;45(W1):W222-W228. doi:10.1093/nar/gkx313

Lu Q, Hu Y, Sun J, Cheng Y, Cheung KH, Zhao H. A statistical framework to predict functional non-coding regions in the human genome through integrated analysis of annotation data. Sci Rep. 2015;5:10576. Published 2015 May 27. doi:10.1038/srep10576

Malhis N, Jacobson M, Jones SJM, Gsponer J. LIST-S2: taxonomy based sorting of deleterious missense mutations across species. Nucleic Acids Res. 2020;48(W1):W154-W161. doi:10.1093/nar/gkaa288

Mi H, Muruganujan A, Thomas PD. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. Nucleic Acids Res. 2013;41(Database issue):D377-D386. doi:10.1093/nar/gks1118

Ng PC, Henikoff S. Accounting for human polymorphisms predicted to affect protein function. Genome Res. 2002;12(3):436-446. doi:10.1101/gr.212802

Ng PC, Henikoff S. Predicting deleterious amino acid substitutions. Genome Res. 2001;11(5):863-874. doi:10.1101/gr.176601

Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003;31(13):3812-3814. doi:10.1093/nar/gkg509

Niroula A, Urolagin S, Vihinen M. PON-P2: prediction method for fast and reliable identification of harmful variants. PLoS One. 2015;10(2):e0117380. Published 2015 Feb 3. doi:10.1371/journal.pone.0117380

Quang D, Chen Y, Xie X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. Bioinformatics. 2015;31(5):761-763. doi:10.1093/bioinformatics/btu703

Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. 2019;47(D1):D886-D894. doi:10.1093/nar/gky1016

Reva B, Antipin Y, Sander C. Determinants of protein function revealed by combinatorial entropy optimization. Genome Biol. 2007;8(11):R232. doi:10.1186/gb-2007-8-11-r232

Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res. 2011;39(17):e118. doi:10.1093/nar/gkr407

Richards S, Aziz N, Bale 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. Genet Med. 2015;17(5):405-424. doi:10.1038/gim.2015.30

Ritchie GR, Dunham I, Zeggini E, Flicek P. Functional annotation of noncoding sequence variants. Nat Methods. 2014;11(3):294-296. doi:10.1038/nmeth.2832

Rodrigues C, Santos-Silva A, Costa E, Bronze-da-Rocha E. Performance of In Silico Tools for the Evaluation of UGT1A1 Missense Variants. Hum Mutat. 2015;36(12):1215-1225. doi:10.1002/humu.22903

Schwarz JM, Rödelsperger C, Schuelke M, Seelow D. MutationTaster evaluates disease-causing potential of sequence alterations. Nat Methods. 2010;7(8):575-576. doi:10.1038/nmeth0810-575

Shihab HA, Gough J, Cooper DN, Day IN, Gaunt TR. Predicting the functional consequences of cancer-associated amino acid substitutions. Bioinformatics. 2013;29(12):1504-1510. doi:10.1093/bioinformatics/btt182

Shihab HA, Gough J, Cooper DN, et al. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. Hum Mutat. 2013;34(1):57-65. doi:10.1002/humu.22225

Shihab HA, Gough J, Mort M, Cooper DN, Day IN, Gaunt TR. Ranking non-synonymous single nucleotide polymorphisms based on disease concepts. Hum Genomics. 2014;8(1):11. Published 2014 Jun 30. doi:10.1186/1479-7364-8-11

Stone EA, Sidow A. Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. Genome Res. 2005;15(7):978-986. doi:10.1101/gr.3804205

Sundaram L, Gao H, Padigepati SR, et al. Predicting the clinical impact of human mutation with deep neural networks (published correction appears in Nat Genet. 2019 Feb;51(2):364]. Nat Genet. 2018;50(8):1161-1170. doi:10.1038/s41588-018-0167-z

Sunyaev S, Ramensky V, Koch I, Lathe W 3rd, Kondrashov AS, Bork P. Prediction of deleterious human alleles. Hum Mol Genet. 2001;10(6):591-597. doi:10.1093/hmg/10.6.591

Tang H, Thomas PD. Tools for Predicting the Functional Impact of Nonsynonymous Genetic Variation. Genetics. 2016;203(2):635-647. doi:10.1534/genetics.116.190033

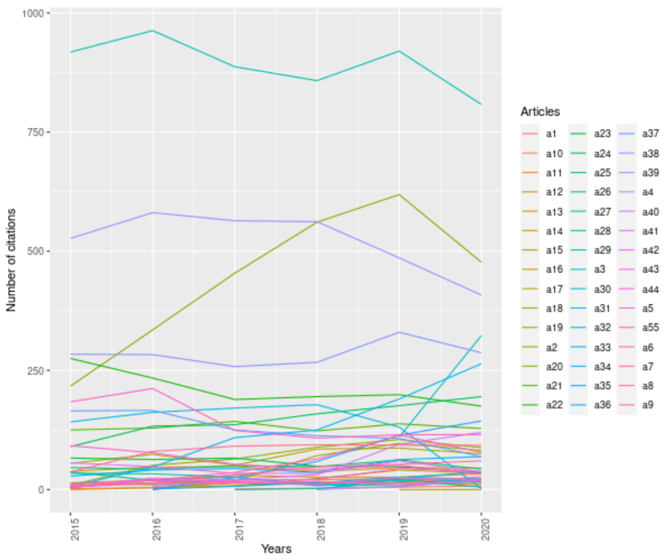
Thomas PD, Campbell MJ, Kejariwal A, et al. PANTHER: a library of protein families and subfamilies indexed by function. Genome Res. 2003;13(9):2129-2141. doi:10.1101/gr.772403

Vaser R, Adusumalli S, Leng SN, Sikic M, Ng PC. SIFT missense predictions for genomes. Nat Protoc. 2016;11(1):1-9. doi:10.1038/nprot.2015.123

Yates CM, Filippis I, Kelley LA, Sternberg MJ. SuSPect: enhanced prediction of single amino acid variant (SAV) phenotype using network features. J Mol Biol. 2014;426(14):2692-2701. doi:10.1016/j.jmb.2014.04.026

Zhang Z, Miteva MA, Wang L, Alexov E. Analyzing effects of naturally occurring missense mutations. Comput Math Methods Med. 2012;2012:805827. doi:10.1155/2012/805827

Supplementary Figure 1: Number of citations overtime for all predictor articles.
a1=BayesDel, 2017; a2=CADD, 2014; a3=CADD, 2019; a4=ClinPred, 2018; a5=CONDEL, 2011; a6=DANN, 2015; a7=Eigen and Eigen-PC, 2016; a8=FATHMM1, 2013; a9=FATHMM, 2013; a10=FATHMM, 2014; a11=fitCons x 4, 2015; a12=FunSeq2, 2014; a13=GenoCanyon, 2015; a14=GWAVAE, 2014; a15=LINSIGHT, 2017; a16=LIST-S2, 2020; a17=LRT, 2009; a18=M-CAP, 2016; a19=MetaSVM/LR, 2015; a20=MutationAssessor, 2007; a21=MutationAssessor, 2011; a22=MutationTaster2, 2010; a23=MutPred, 2009; a24=PANTHER, 2003; a25=PhD-SNP, 2006; a26=PMut, 2005; a27=PMut, 2017; a28=PolyPhen, 2001; a29=Polyphen2, 2010; a30=PON-P2, 2015; a31=PrimateAI, 2018; a32=PROVEAN, 2012; a33=PROVEAN, 2015; a34=PSNPE, 2014; a35=PSNPE, 2016; a36=REVEL, 2016; a37=SIFT, 2001; a38=SIFT, 2003; a39=SIFT, 2009; a40=SIFT4G, 2016; a41=SNAP2, 2007; a42=SNPs&GO, 2013; a43=SuSPect, 2014; a44=VEST, 2013; a55=PANTHER, 2013



Supplementary Table 1: A complete list of articles of 34 predictors with total citation number in three different databases (Pubmed, Google Scholar, and Web of Science) and the percentage of citations for each predictor.

Predictors	Year	Titles	Pubmed	Google Scholar	Web of science	% of total cited
Polyphen2	2010	A method and server for predicting damaging missense mutations	5069	10298	7479	20.543
SIFT	2009	Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm	2826	5825	4236	11.588
SIFT	2003	SIFT: Predicting amino acid changes that affect protein function	1874	4588	3081	8.581
CADD	2014	A general framework for estimating the relative pathogenicity of human genetic variants	2010	3885	2698	7.727
PANTHER	2003	PANTHER: A Library of Protein Families and Subfamilies Indexed by Function	1110	2543	1697	4.811
MutationTaster2	2010	MutationTaster evaluates disease-causing potential of sequence alterations	1118	2361	1766	4.716
SIFT	2001	Predicting deleterious amino acid substitutions	1025	2374	1628	4.520
PROVEAN	2012	Predicting the Functional Effect of Amino Acid Substitutions and Indels	903	2011	899	3.429
PANTHER	2013	PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees	727	1445	1009	2.860
MutationAssessor	2011	Predicting the functional impact of protein mutations: application to cancer genomics	694	1420	953	2.758
Polyphen2	2013	Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2	939	2022	-	2.662
PolyPhen	2001	Prediction of deleterious human alleles	436	1210	878	2.270
PROVEAN	2015	PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels	454	1098	746	2.066
LRT	2009	Identification of deleterious mutations within three human genomes	394	764	536	1.523
FATHMM	2013	Predicting the Functional, Molecular, and Phenotypic Consequences of Amino Acid Substitutions using Hidden Markov Models	371	751	516	1.473
CONDEL	2011	Improving the Assessment of the Outcome of Nonsynonymous SNVs with a Consensus Deleteriousness Score, Condel	379	708	502	1.429
SNAP2	2007	SNAP: predict effect of non-synonymous polymorphisms on function	318	721	507	1.390
MutPred	2009	Automated inference of molecular mechanisms of disease from amino acid substitutions	285	704	490	1.330
CADD	2019	CADD: predicting the deleteriousness of variants throughout the human genome	304	701	441	1.300
PhD-SNP	2006	Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information	240	631	438	1.177
MetaSVM/LR	2015	Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies	301	579	407	1.157
REVEL	2016	REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants	232	504	341	0.968
PMut	2005	PMUT: a web-based tool for the annotation of pathological mutations on proteins	214	488	368	0.962
DANN	2015	DANN: a deep learning approach for annotating the pathogenicity of genetic variants	208	486	310	0.903

GWAVAE	2014	Functional annotation of noncoding sequence variants	215	400	266	0.792
SIFT4G	2016	SIFT missense predictions for genomes	182	414	278	0.786
PSNPE	2014	PredictSNP: Robust and Accurate Consensus Classifier for Prediction of Disease-Related Mutations	158	397	286	0.756
M-CAP	2016	M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity	173	377	269	0.736
Eigen and Eigen-PC	2016	A spectral approach integrating functional genomic annotations for coding and noncoding variants	161	320	204	0.616
VEST	2013	Identifying Mendelian disease genes with the Variant Effect Scoring Tool	124	254	161	0.485
MutationAssessor	2007	Determinants of protein function revealed by combinatorial entropy optimization	116	245	155	0.464
FunSeq2	2014	FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer	125	234	148	0.456
fitCons x 4	2015	A method for calculating probabilities of fitness consequences for point mutations across the human genome	97	174	107	0.340
LINSIGHT	2017	Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data	92	179	103	0.336
FATHMM	2013	Predicting the functional consequences of cancer-associated amino acid substitutions	82	171	102	0.319
SuSPect	2014	SuSPect: Enhanced Prediction of Single Amino Acid Variant (SAV) Phenotype Using Network Features	58	151	93	0.272
SNPs&GO	2013	WS-SNPs&GO: a web server for predicting the deleterious effect of human protein variants using functional annotation	62	139	93	0.264
PON-P2	2015	PON-P2: Prediction Method for Fast and Reliable Identification of Harmful Variants	59	132	97	0.259
FATHMM	2014	Ranking non-synonymous single nucleotide polymorphisms based on disease concepts	41	107	64	0.191
PrimateAI	2018	Predicting the clinical impact of human mutation with deep neural networks	44	106	62	0.191
PSNPE	2016	PredictSNP2: A Unified Platform for Accurately Evaluating SNP Effects by Exploiting the Different Characteristics of Variants in Distinct Genomic Regions	49	93	70	0.191
GenoCanyon	2015	A Statistical Framework to Predict Functional Non-Coding Regions in the Human Genome Through Integrated Analysis of Annotation Data	46	90	53	0.170
PMut	2017	PMut: a web-based tool for the annotation of pathological variants on proteins, 2017 update	20	75	38	0.120
ClinPred	2018	ClinPred: Prediction Tool to Identify Disease-Relevant Nonsynonymous Single-Nucleotide Variants	16	36	26	0.070
BayesDel	2017	PERCH: A Unified Framework for Disease Gene Prioritization	13	19	16	0.043
LIST-S2	2020	LIST-S2: taxonomy based sorting of deleterious missense mutations across species	1	1	0	0.002

Capítulo 2

O capítulo é uma análise comparativa de 33 preditores e um escore de conservação avaliados em um grupo de 160 variantes *missense* relatadas na literatura para o gene *IDUA*. Os preditores que obtiveram melhor desempenho foram utilizados para avaliar 426 variantes de significado incerto reportadas em bancos de dados populacionais (ExAC v0.3.1, gnomAD v2.0.2, ABraOM, LOVD, 1,000 genomes (1,000 Genomes Project Consortium), dbSNP, Human Genome Mutation Database (HGMD) and ClinVar). Além de investigar as variantes do gene *IDUA*, o trabalho também serviu para estabelecer os parâmetros para busca na literatura e análises estatísticas realizadas no capítulo 3.

O artigo está em fase de formatação para envio para o periódico *Molecular Genetics and Metabolism*.

Which is the best *in silico* program for the missense variations in *IDUA* gene? A comparison of 33 programs plus a conservation score and evaluation of 586 missense variants.

Pâmella Borges[1,2,3], Gabriela Pasqualim[5], Ursula Matte[1,2,3,4].

Affiliations:
[1]Cell, Tissue and Gene Laboratory, Clinicas Hospital of Porto Alegre (HCPA), Rio Grande do Sul, Brazil
[2]Bioinformatics Core, Experimental Research Centre, HCPA, Rio Grande do Sul, Brazil.
[3]Graduate Programme in Genetics and Molecular Biology, Federal University of Rio Grande do Sul (UFRGS).
[4]Department of Genetics, UFRGS, Porto Alegre, Brazil.
[5]Genetics Laboratory, Biological Sciences Institute, Federal University of Rio Grande (FURG)

Correspondence: pamella.bor@gmail.com

Abstract

Mucopolysaccharidosis type I (MPS I) is an autosomal recessive disease characterized by the deficiency of alpha-L-iduronidase (*IDUA*), an enzyme involved in glycosaminoglycan (GAG) degradation. More than 200 disease-causing variants have been reported and characterized in the *IDUA* gene. The gene also has several variants of unknown significance (VUS) and literature conflicting interpretations of pathogenicity. This study evaluated 586 variants obtained from the literature review, five population databases, in addition to dbSNP, HGMD, and ClinVar. For the variants described in the literature, two datasets were created based on the strength of the criteria. The stricter criteria subset had 108 variants with expression study, analysis of healthy controls, and/or complete gene sequence. The less stringent criteria subset had additional 52 variants found in the literature review, HGMD or ClinVar, and in dbSNP with allele frequency higher than 0.001. The other 426 variants were considered VUS. The two strength criteria datasets were used to evaluate 33 programs plus a conservation score. BayesDel (addAF and noAF), PONP2 (genome and protein), and ClinPred algorithms showed the best sensitivity, specificity, accuracy, and kappa value for both criteria subsets. The VUS variants were evaluated with these five algorithms. Based on results, 122 variants had total consensus among the five predictors, with 57 classified as predicted deleterious and 65 as predicted neutral. For variants not included in PONP2, 88 variants were considered deleterious and 92 neutral by all other predictors. The remaining 124 did not obtain a consensus among predictors.

Keywords: Mucopolysaccharidosis type I (MPS I), missense variants, *in silico* predictions, VUS classifications.

Introduction

Mucopolysaccharidosis type I (MPS I) is an autosomal recessive disease characterized by the deficiency of alpha-L-iduronidase (*IDUA*) involved in glycosaminoglycan (GAG) degradation (Scott 1991). This deficiency leads to progressive lysosomal accumulation of heparan and dermatan sulfate and causes a gradual deterioration of cells and tissues that culminates in early death in severe cases (Lehman et al., 2011). MPS I has considerable phenotypic variation, with an extensive range of clinical manifestations and well-defined extreme phenotypes. Scheie syndrome (MPS I-S; OMIM# 607016) is the attenuated phenotype and includes somatic involvement, while Hurler syndrome (MPS I-H; OMIM# 607014) is the severe phenotype with important neurological impairment, among other features (Kubaski et al., 2020). All phenotypes exhibit excessive GAG accumulation and excretion in urine and are undistinguishable by routine biochemical tests (Lehman et al., 2011; Viana et al., 2011).

More than 200 disease-causing variants have been reported and characterized in the *IDUA* gene (Bertola et al., 2011). In a 2019 study with data from the MPS I Registry, nonsense and missense variants corresponded, respectively, to 56.5% and 33.6% of the reported variants (Clarke et al., 2019). Attenuated cases present at least one allele with residual activity, generally due to missense variants, regardless of the other allele and genotype-phenotype correlation has been established for some missense pathogenic variants (Fuller et al. 2005). Non-disease causing missense variants, such as p.Arg105Gln, p.Gln63Pro (Scott et al., 1991), p.His33Gln (Scott et al., 1992), and p.Ala361Thr (Scott et al., 1993), have also been described in the literature.

The broader use of massive parallel genetic sequencing increased the list of variants of unknown significance (VUS). Functional molecular assessments do not accompany the pace of detection of new genetic variants. Most variants present in Exome Aggregation Consortium (ExAC) and The Genome Aggregation Database (gnomAD) (Lek et al., 2016; Karczewski et al., 2020) have not been assessed for their pathogenicity. Therefore, research and clinical laboratories use *in silico* strategies to help understand the biological significance of VUS. These methods are already considered in ACMG standard guidelines (Richards et al., 2015) to indicate some supporting evidence level when clinical information is insufficient or nonexistent. Clinical laboratories also created their guideline on variant interpretation, named Sherloc (semiquantitative, hierarchical evidence-based rules for locus interpretation) (Keith Nykamp et al., 2017).

Even though computational analysis is often used, results must be viewed with caution. Not only do different programs have discordant results for the same gene, but algorithms may have different values of accuracy, specificity, and sensitivity depending on the characteristics of the gene or protein. Therefore, ideally, a performance assessment should be performed for each gene/protein to choose the best algorithm for variant prioritization. However, this also needs reliable standards as calibrators - and literature and curated databases also show divergence.

This study aims to compare *in silico* predictors using two datasets of variants with different degrees of confidence. Using the best predictors indicated by these two datasets, we evaluated the VUS present in the *IDUA* gene in population databases.

Methods

Curated variant selection:

We created a database with missense variants described in the literature, in curated databases, and in population databases with frequencies greater than 0.001. First, we performed a manual review of all missense variants in the *IDUA* gene published between 1991 and 2019. According to the variant classification methods in each manuscript, variants from the literature were divided into two subsets (stricter or less stringent evidence). Evidence was considered stricter if at least one of the following was performed: expression study, evaluation of healthy controls, or complete gene sequence corroborating the pathogenic or non-pathogenic disease-causing variant status. The subset of variants with less stringent criteria comprised all variants in the stricter criteria subset plus the rest of missense variants described in the literature, variants from HGMD (Stenson et al. 2014) and Clinvar (with their classifications) (Landrum et al. 2014), and variants in population databases with allele frequencies greater than 0.001. These two subsets were selected to evaluate the prediction programs' characteristics and to compare the correlation between variants' predictions and literature information. Variants that do not have any of these criteria were considered VUS.

In silico programs:

We analyzed 33 prediction algorithms and one conservation score. For those predictors with more than one training set, such as PolyPhen HDIV and HVAR, each training set was evaluated separately. So, in total we had 51 predictors: SIFT (protein data training) (Kumar et al., 2009), SIFT4G (Vaser et al., 2016), Polyphen2 (HDIV and HVAR) (Adzhubei et al., 2013), LRT (Chun; Fay, 2009), MutationTaster2 (Schwarz et al., 2010), MutationAssessor (Reva et al., 2007), FATHMM (Coding Variants-Weighted, MKL coding, and XF coding) (Shihab et al., 2013), MetaSVM/LR (Dong et al., 2015), CADD (GRCh37/hg19 and GRCh38/hg38) (Kircher et al., 2014), VEST4 (Carter et al., 2013), PROVEAN (protein data training) (Choi et al., 2012), fitCons x4 (Gulko et al., 2015), LINSIGHT (Huang et al., 2017), M-CAP (Jagadeesh et al., 2016), REVEL (Ioannidis et al., 2016), MutPred (Li et al., 2009), PrimateAI (Sundaram et al., 2019), BayesDel (addAF and noAF) (Feng, 2017), ClinPred (Alirezaie et al., 2018) and LIST-S2 (Malhis et al., 2020). We also tested GERP++ conservation score (Davydov et al., 2010) from dbNSFP v4.1a, a database developed for functional prediction and annotation of all potential non-synonymous single-nucleotide variants (nsSNVs) in the human genome (Liu et al., 2020).

The prediction of PhD-SNP (Capriotti et al., 2006), PANTHER (Thomas et al., 2003), SNPs&GO (Capriotti et al., 2013), PredictSNP (Bendl et al., 2016), CADD 1.2, DANN (Quang et al., 2015), FATHMM (Coding Variants - Unweighted), FunSeq2 (Fu et al., 2014), GWAVAE 1.0 (Ritchie et al., 2014), SuSPect (Yates et al., 2014), PMut (Ferrer-Costa et al., 2005), CONDEL (González-Pérez et al., 2011), PROVEAN (genome data training), SIFT (genome data training), PON-P2 (identifier, protein and genome data training) (Niroula et al., 2015) and MutPred were obtained from the web-based application. The variant classifiers were used when provided by the algorithm. The scores of VEST4, REVEL, MutPred, CADD_raw, CADD_phred, integrated_fitCons, SusPect, and GERP++_NR were transformed in binary classification. The cutoff of 0.5 was applied for SuSPect and VEST4, 0.75 for MutPred and REVEL, 20 for CADD_phred, zero for CADD_raw, 0.4 for fitCons x4, and 0.047 for GERP++ as suggested by the authors.

Variants of unknown significance:

All missense variants in canonical *IDUA* sequence present in ExAC v0.3.1 (Lek et al. 2016), gnomAD v2.0.2 (Karczewski et al. 2019), ABraOM (Naslavsky et al. 2017), LOVD (Fokkema et al. 2005), 1,000 genomes (1,000 Genomes Project Consortium), dbSNP (Sherry et al. 2001) with frequencies lower than 0.0001, plus variants in the Human Genome Mutation Database (HGMD) (Stenson et al. 2014) with classification conflict and in ClinVar (Landrum et al. 2014) those without classification were considered VUS. These variants were merged in a single database to remove duplicates and exclude those included in the datasets previously used to compare the algorithms.

Statistical analysis:

The statistical analysis was performed using SPSS (Statistical Package for the Social Sciences) and python algorithms. The sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), accuracy, true positive rate (TPR), false-positive rate (FPR), and the Fisher's exact test were calculated on python with libraries matplotlib.pyplot (Hunter, 2007), sklearn.metrics (Pedregosa et al., 2011), pandas (The pandas development team, 2020), and NumPy (Harris, 2020). The kappa value was generated with SPSS 18.03.

Results

A total of 586 unique variants were analyzed in this study obtained according to the workflow presented in Figure 1. Each database's contribution can be seen in supplementary figure 1. dbSNP (Sherry et al. 2001) and gnomAD v2.0.2 (Karczewski et al. 2019) databases had the larger number of variants, with 363 and 316, respectively, being 83 and 86 exclusives. ExAC v0.3.1 (Lek et al., 2016) contributed with 266 variants, being only six exclusives. LOVD (Fokkema et al. 2005) presents 44 variants, and three were exclusive, whereas HGMD (Stenson et al. 2014) and ClinVar (Landrum et al. 2014) contributed with 3 and 19 exclusive variants, respectively, from a total of 136 and 131. ABraOM (Naslavsky et al. 2017) and 1,000 genomes (1,000 Genomes Project Consortium) presented 19 and 47 variants, respectively, but none was private.

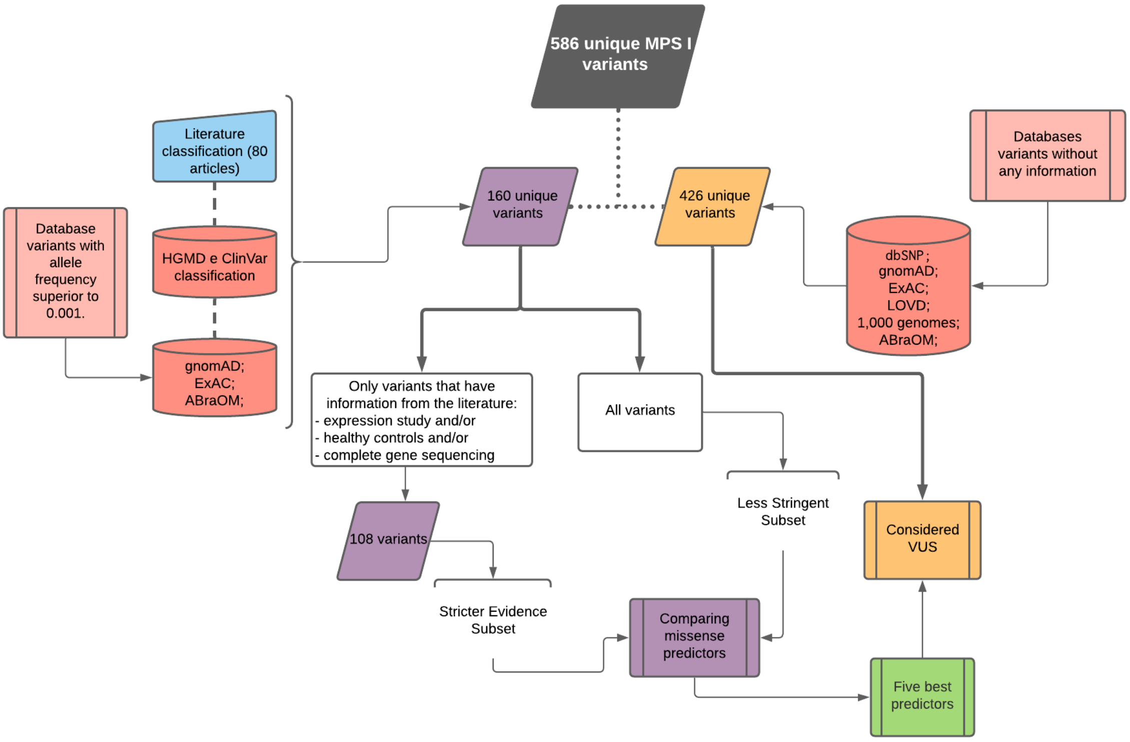

586 unique MPS I variants
Literature classification (80 articles)
Database variants with allele frequency superior to 0.001.
HGMD e ClinVar classification
gnomAD; ExAC; ABraOM;
160 unique variants
426 unique variants
Databases variants without any information
dbSNP; gnomAD; ExAC; LOVD; 1,000 genomes; ABraOM;
Only variants that have information from the literature:
- expression study and/or
- healthy controls and/or
- complete gene sequencing
All variants
Less Stringent Subset
Considered VUS
108 variants
Stricter Evidence Subset
Comparing missense predictors
Five best predictors

Figure 1: Workflow chart showing variant retrieval and curation.

First, 145 variants manually retrieved from the literature were combined with variants in curated databases and population databases with frequencies higher than 0.001. This formed a set of 160 unique variants used to compare the algorithms. Another 426 variants were obtained from population databases and considered VUS.

According to the type of evidence used for their description, variants in the first set were divided into two subgroups. Out of the 145 variants from the literature, 108 had at least one of three measures that were considered stricter evidence criteria (Figure 2). In this group of variants with stricter evidence, 91 were disease-causing, and of these, 19 variants do not have expression studies, 48 variants were not analyzed in healthy controls, and 50 variants were not described in studies with complete gene sequencing (Supplementary Table 1). Of the 17 non-disease-causing variants in the group with stricter evidence, only five were not analyzed by expression studies (Supplementary Table 2).

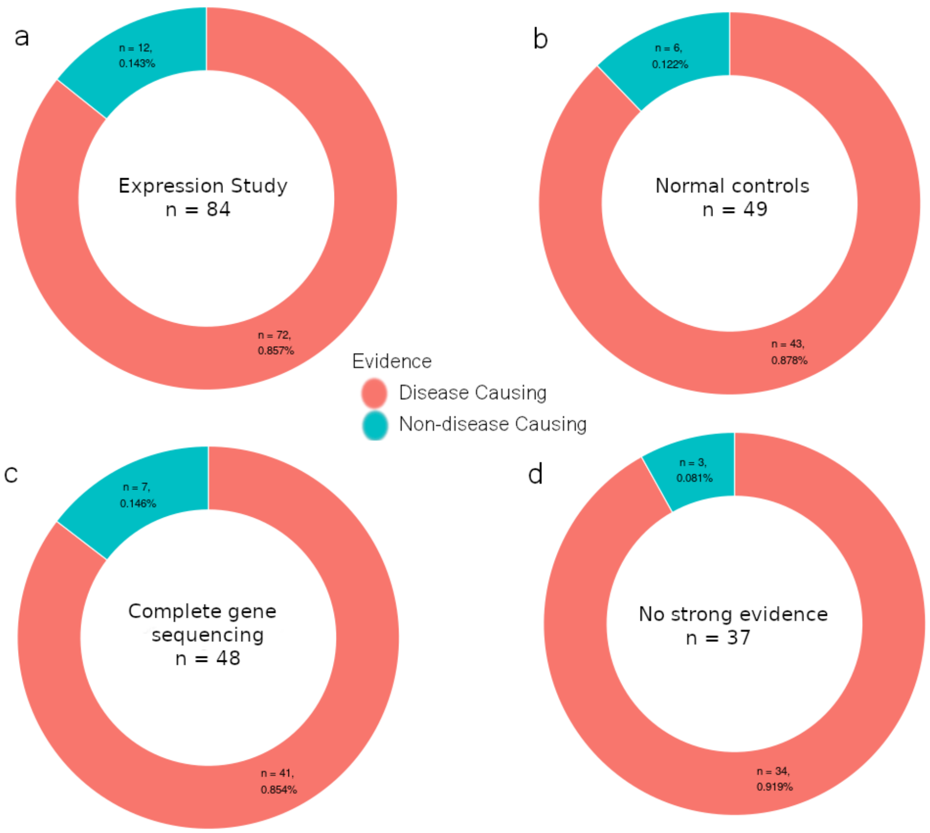


Figure 2: Percentage of disease-causing and non-disease-causing variants in each evidence criteria: variants with expression study (a), comparison with normal controls (b), complete gene sequencing (c), and absence of stricter evidence (d).

The 160 variants (26 predicted neutral and 134 predicted deleterious) in the less stringent criteria subset and 108 variants (17 benign and 91 pathogenic) in the stricter criteria subset were used for evaluating 33 prediction algorithms plus one conservation score. As one program may present more than one training dataset, a total of 51 estimates were obtained. SIFT, PROVEAN, PolyPhen2, BayesDel, CADD, FATHMM, fitCons, MutPred, and PON-P2 were evaluated for every available training set.

For the stricter criteria subset, only BayesDel (addAF and noAF), PONP2 (genome, protein, and identifier), and ClinPred presented accuracy higher than 90% and kappa value higher than 0.6, being PONP2 (genome database), ClinPred and BayesDel (addAF) the ones with the best relation between sensitivity and specificity and higher kappa value (0.692, 0.719 and 0.821) (Supplementary Table 3). One PPV could not be calculated because FunSeq classified all variants as neutral. Three algorithms (integrated_fitCons, GM12878_fitCons and M-CAP) classified all variants as deleterious and did not present NPV. Kappa value also could not be calculated for these four predictors.

The lowest sensitivities (between 0 and 0.3) were observed in PrimeAI and SusPect predictors. Excluding predictors that have maximum sensitivity and minimal specificity, the algorithms Polyphen2 (HDIV), MutationTaster, MutationAssessor, VEST4, DEOGEN2, BayesDel (addAF and noAF), ClinPred, CADD (raw_hg38, phred_hg38, raw_hg19, phred_hg19), FATHMM (Coding Variants - Weighted), H1hESC_fitCons, GERP++, CONDEL, and PON-P2 (identifier, protein, and genome) present large sensitivity (over 90%). Excluding FunSeq, only SNPs&GO have specificity higher than 90%, and 14 algorithms have specificity between 80 and 90% (Supplementary Table 3).

The less stringent criteria subset showed similar patterns as the stricter criteria subset despite obtaining a general reduction in the calculated values, except for the PON-P2 (identifier) algorithm that showed an increased sensitivity. The same four algorithms classified all variants as only neutral or deleterious. In this subset, no algorithm had specificity higher than 90%, and nine algorithms had specificity between 80 and 90%, including PrimateAI and SNPs&GO (Supplementary Figure 2a and 2b). In this subset, PONP2 (genome database), ClinPred and BayesDel (addAF) obtained accuracy higher than 90% (0.92, 0.91 and 0.93) and kappa value higher than 0.6 (0.666, 0.680 and 0.743) (Figure 3a and 3b). All sensitivity, specificity, accuracy, PPV, NPV, FPR, and Kappa values are displayed in Supplementary Tables 3 and 4 for the stricter and less stringent criteria subsets.

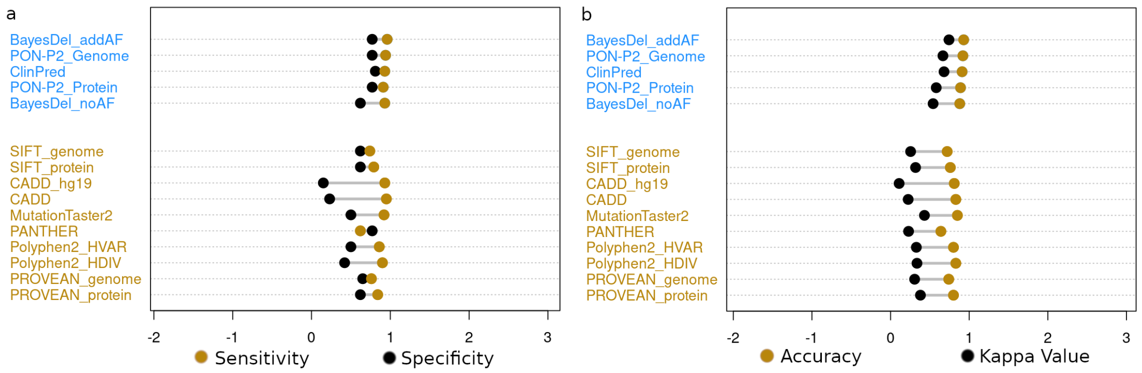


Figure 3: The sensitivity and specificity (a), and the accuracy and kappa value (b) for the top five classifiers in blue (BayesDel-addAF, PONP2-genome, ClinPred, PONP2-protein, and BayesDel-noAF algorithms) and the top six cited in yellow (SIFT, CADD, MutationTaster2, PANTHER, PolyPhen2, and Provean) for the less stringent criteria subset.

The Fisher's Exact Test was performed to test if less stringent criteria and stricter criteria subsets present statistical differences in predictors' performance. The ratio of hits and errors for each program was compared between less stringent criteria and stricter subsets, and none presented statistically significant values (Figure 4a). When we compared the same subset estimates, both subsets have the same pattern with several p-values lower than 0.05, as shown in Figure 4b for the less stringent criteria subset.

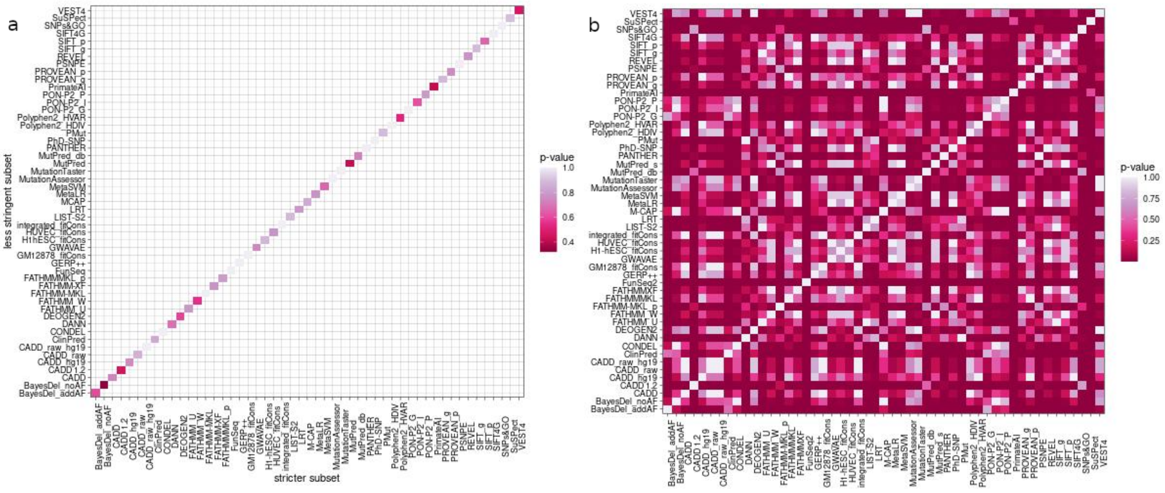


Figure 4: P-value of Fisher’s Exact Test comparing less stringent criteria and stricter subset (a) and the 51 estimates in a less stringent subset (b).

Not all 51 estimates were obtained for all 160 variants. MutationAssessor, LRT, PrimateAI, PANTHER, GWAVAE, PMut, M-CAP, MutPred, and all three PON-P2 algorithms did not return a predicted classification for some variants (Figure 5). All three PON-P2 training sets were the predictors that contained the most unclassified variants, followed by MutPred and predictions obtained from dbNSFP. The algorithms LRT (2), MutationAssessor (3), and PrimateAI (3) failed to classify variants in the first amino acid (MutationAssessor and PrimateAI) or at the end of the protein (LRT).

For the stricter criteria subset, all programs fail to report more deleterious variants except for M-CAP. MutationAssessor, PrimateAI, PANTHER present the fewest number of unclassified variants, and only for deleterious. MutPred dbNSFP produces a larger number of unclassified variants both neutral and deleterious. For the less stringent criteria subset, MutPred dbNSFP increased the number of unclassified variants, exceeding the other programs (Figure 4). LRT and PMut had one neutral and one deleterious uncategorized variant, respectively, in this subset. M-CAP continued to show more neutral (8) than deleterious (2) variants unclassified.

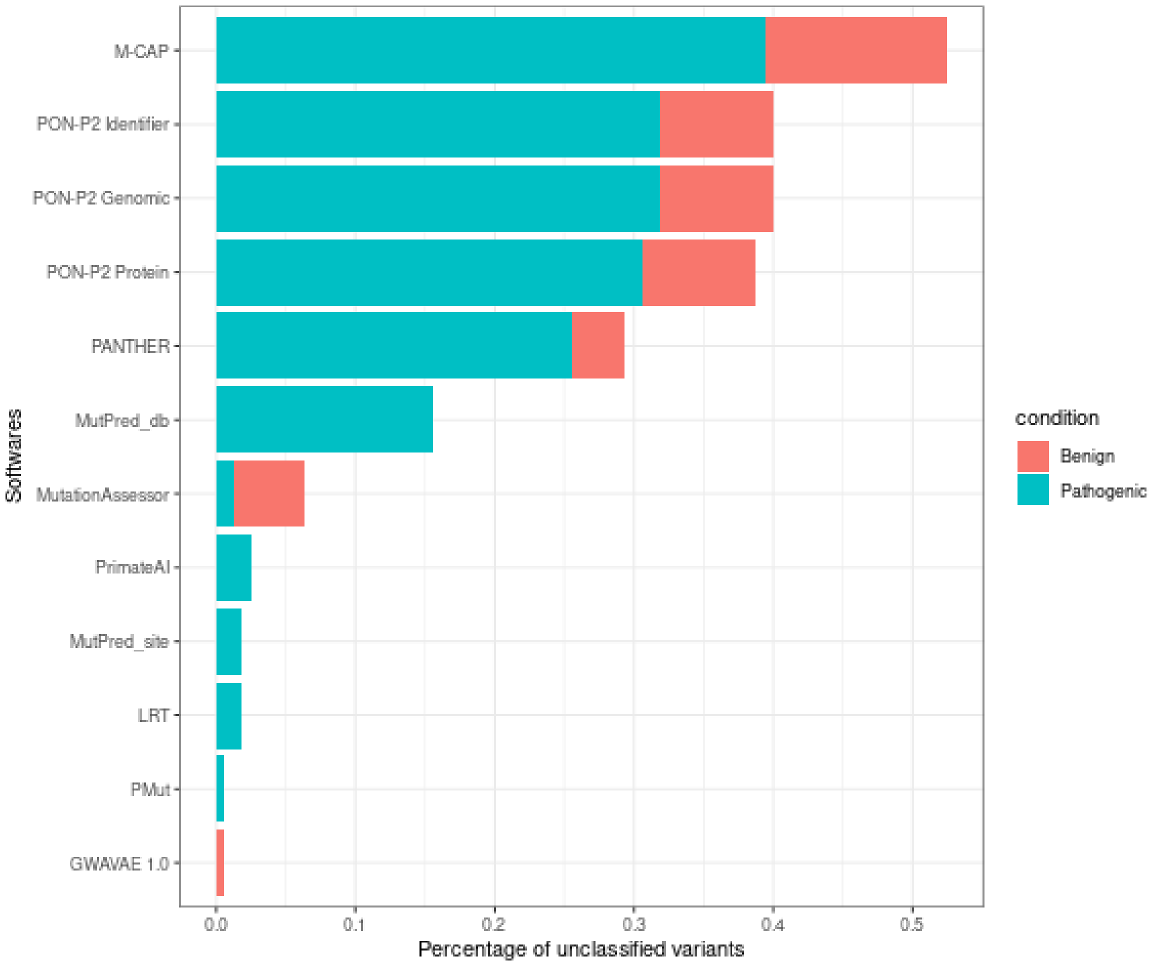


Figure 5: Number of unclassified variants per software for the less stringent criteria subset.

In silico VUS classification:

Based on values present in both evaluation subsets, the 426 VUS were classified using the best five predictors: BayesDel (addAF and noAF), PONP2 (genome and protein), and ClinPred algorithms. PONP2 (genome and protein) is the only of these five predictors that do not classify every variant, with both failing to classify 267 variants plus six unclassified variants exclusive to PONP2-genome and other six exclusives to PONP2-protein. Out of the 426 variants, 57 obtained a total consensus of the five programs as predicted deleterious and 65 as predicted neutral. For variants not included in PONP2, 88 variants were considered deleterious and 92 neutral by all other predictors. The remaining 124 did not obtain a consensus among predictors (Figure 6).

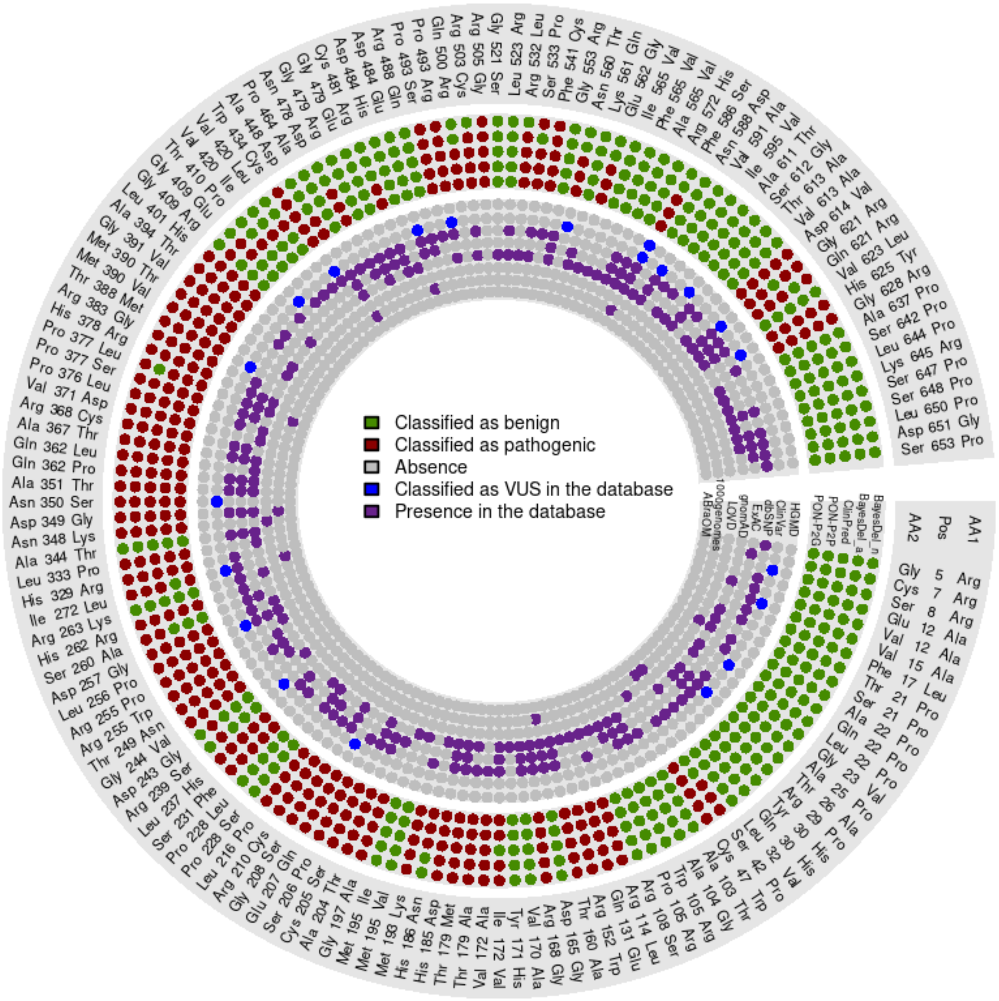


Figure 6: VUS classified by all best softwares.

Discussion

In this study, we evaluated the prediction of 33 softwares plus a conservation score for missense variants in the *IDUA* gene. Two datasets were created based on literature information and public databases: the first dataset was used to evaluate the best predictors for missense *IDUA* variants. The second dataset comprised 426 VUS that were evaluated by the five best-performing algorithms. For the first dataset, two subsets were separated based on standards: modifications with specific literature information as stricter criteria subset and all variants present in literature review plus databases with variant classification and high allele frequency. These variants were included to increase the amount of non-disease-causing mutations in the curated dataset.

The subsets did not demonstrate a notable difference, although the less stringent criteria subset presents lower overall values. The difference in performance may be explained by the lower classification confidence of the less stringent criteria subset. While the stricter criteria subsets represent a supervised subset and include variants with a high confidence level of categorization, the less stringent criteria and more flexible subset may contain incorrect classification. That may be due to the relatively small number of variants introduced in the less stringent criteria subset (52 added to the 108 in the stricter subset).

Despite that, both comparison groups present the same predictors with the most satisfactory performances. BayesDel, the best performance predictor, is a metascore that combines deleteriousness predictors in the naïve Bayesian approach and uses ClinVar variants as standard to determine cutoff value. For this predictor, the set that integrates maximum and minor allele frequency across populations (addAF) presents superior performance than without (noAF) (Feng, 2017). ClinPred had the second-highest value in the kappa test, and either uses ClinVar as a training dataset and combines two machine learning algorithms: random forest (cforest) and gradient boosted decision tree (xgboost) models (Alirezaie et al., 2018). PON-P2 uses variation data from VariBench to train a random forest selection features predictor for pathogenicity-association of amino acid substitutions and accept variations in multiple formats. The primer format (protein) is the most responsive, despite presenting a moderately more modest performance than the genome format.

Classic and often used predictors such as SIFT (genome and protein) (Kumar et al., 2009) and PolyPhen2 (HumDiv and HumVar) (Adzhubei et al., 2013) did not perform well in both comparison subsets. For the stricter criteria subset, PolyPhen2 (HDIV), preferred for evaluating rare alleles, had good sensitivity (90%), accuracy (83%), and kappa value (0.372), but specificity lower than 50% (supplementary table 3). CADD score (Combined Annotation Dependent

Depletion) integrates multiple annotations into one metric (Kircher et al., 2014) and presents sensitivity higher than 90% and accuracy higher than 80% for GRCh37/hg19 and GRCh38/hg38. Unfortunately, it possessed one of the smallest specificities and kappa value between evaluated programs. A recently developed program, REVEL, an ensemble method that manages random forest (Ioannidis et al., 2016), displays a compelling performance, despite not being one of the bests, with higher specificity (88%) than sensitivity (75%)

Several predictors use ClinVar and HGMD databases as training datasets. Therefore, some hits in our datasets are reanalysis of training variants and not an accurate interpretation of pathogenicity, but this is not the case for all evaluated variants. Also, it is not likely that this would bias our analysis, even though we worked with variants native to these databases (Figure 2), as the training datasets used for these programs incorporate many more variants in numerous genes.

A recurrent problem in performance evaluation is the disproportionality of training and evaluation sets regarding the number of neutral and deleterious variants, a discrepancy also found in our datasets. We observed a minimal absolute difference between the properties of predicted deleterious and neutral modifications, with the stricter criteria subset having 15.74% of neutral variants while the less stringent criteria subset had 16.25%. This minor difference demonstrates the difficulty of obtaining neutral variants for composing sets, even implementing more comprehensive standards to evaluate these *in silico* predictors. It also reflects the fact that *in silico* software are mostly trained with disease-causing variants, which may cause a bias in the analysis. That was shown by Niroula and Vihinen (2019), who compared ten predictors with a large set of non-pathogenic variants only and found specificity over 80% in just three predictors (PON-P2, VEST, and FATHMM). In our study, despite both subsets presenting various programs with high specificity, the proportion of predicted deleterious and neutral variants does not allow a proper evaluation of specificity or to state which programs would exhibit significant differences in performance in a set with more neutral variants.

This study does not replace the ACMG or Sherloc standards and guidelines. However, it increases confidence in one stage of the classification process (computational predictive programs criterion-PP3, in Sherloc), mainly when used in the absence of additional clinical information, as is the case of variants deposited in public databases. As we do not have access to any clinical information about the 426 variants identified in the public databases, these guidelines could not be applied. Therefore, we used only the classification given by the best five predictors previously selected. A classification for 122 variants (57 predicted deleterious and 65 predicted neutral variants) was obtained with a total consensus of the five programs. The other 304 variants were unclassified by PON-P2 or did not reach an agreement. If PON-P2 was excluded, then 311 variants reached a consensus (predicted deleterious and neutral).

The difference between the number of variants with and without consensus is common and represents a recurrent finding when only information from computational predictive programs is available. This disagreement is probably caused by the metrics used by each predictor and can be a problem when no literature-based validation exists for that particular gene and predictor.

Conclusion

Variants in the *IDUA* gene were evaluated by 33 prediction algorithms and one conservation score for all available training sets. Two subsets were created using stricter and less stringent criteria based on literature information available for each variant. The subsets demonstrated a small difference, with reduced values in the less stringent criteria subset but the same most accurate predictors. The five most accurate predictors were used for evaluating 426 VUS obtained from public databases. Of these, 122 variants showed a total consensus of programs with high confidence in classification, being 57 predicted deleterious and 65 predicted neutral

References

1000 Genomes Project Consortium, Auton A, Brooks LD, et al. A global reference for human genetic variation. *Nature*. 2015;526(7571):68-74. doi:10.1038/nature15393

Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet. 2013;Chapter 7:Unit7.20. doi:10.1002/0471142905.hg0720s76

Alirezaie N, Kernohan KD, Hartley T, Majewski J, Hocking TD. ClinPred: Prediction Tool to Identify Disease-Relevant Nonsynonymous Single-Nucleotide Variants. Am J Hum Genet. 2018;103(4):474-483. doi:10.1016/j.ajhg.2018.08.005

Amr K, Katoury A, Abdel-Hamid M, Bassiouni R, Ibrahim M, Fateen E. Mutational Analysis of the alpha-L-iduronidase gene in three Egyptian families: identification of three novel mutations and five novel polymorphisms. Genet Test Mol Biomarkers. 2009;13(6):761-764. doi:10.1089/gtmb.2009.0057

Aronovich EL, Pan D, Whitley CB. Molecular genetic defect underlying alpha-L-iduronidase pseudodeficiency. Am J Hum Genet. 1996;58(1):75-85.

Atçeken N, Özgül RK, Yücel Yilmaz D, et al. Evaluation and identification of *IDUA* gene mutations in Turkishpatients with mucopolysaccharidosis type I. Turk J Med Sci. 2016;46(2):404-408. Published 2016 Feb 17. doi:10.3906/sag-1411-160

Bach G, Moskowitz SM, Tieu PT, Matynia A, Neufeld EF. Molecular analysis of Hurler syndrome in Druze and Muslim Arab patients in Israel: multiple allelic mutations of the *IDUA* gene in a small geographic area. Am J Hum Genet. 1993;53(2):330-338.

Beesley CE, Meaney CA, Greenland G, et al. Mutational analysis of 85 mucopolysaccharidosis type I families: frequency of known mutations, identification of 17 novel mutations and in vitro expression of missense mutations. Hum Genet. 2001;109(5):503-511. doi:10.1007/s004390100606

Bendl J, Musil M, Štourač J, Zendulka J, Damborský J, Brezovský J. PredictSNP2: A Unified Platform for Accurately Evaluating SNP Effects by Exploiting the Different Characteristics of Variants in Distinct Genomic Regions. PLoS Comput Biol. 2016;12(5):e1004962. Published 2016 May 25. doi:10.1371/journal.pcbi.1004962

Bertola F, Filocamo M, Casati G, et al. *IDUA* mutational profiling of a cohort of 102 European patients with mucopolysaccharidosis type I: identification and characterization of 35 novel α-L-iduronidase (*IDUA*) alleles. Hum Mutat. 2011;32(6):E2189-E2210. doi:10.1002/humu.21479

Bravo H, Neto EC, Schulte J, et al. Investigation of newborns with abnormal results in a newborn screening program for four lysosomal storage diseases in Brazil. Mol Genet Metab Rep. 2017;12:92-97. Published 2017 Jul 4. doi:10.1016/j.ymgmr.2017.06.006

Brooks DA, Fabrega S, Hein LK, et al. Glycosidase active site mutations in human alpha-L-iduronidase. Glycobiology. 2001;11(9):741-750. doi:10.1093/glycob/11.9.741

Bunge S, Clements PR, Byers S, Kleijer WJ, Brooks DA, Hopwood JJ. Genotype-phenotype correlations in mucopolysaccharidosis type I using enzyme kinetics, immunoquantification and in vitro turnover studies. Biochim Biophys Acta. 1998;1407(3):249-256. doi:10.1016/s0925-4439(98)00046-5

Bunge S, Kleijer WJ, Steglich C, Beck M, Schwinger E, Gal A. Mucopolysaccharidosis type I: identification of 13 novel mutations of the alpha-L-iduronidase gene. Hum Mutat. 1995;6(1):91-94. doi:10.1002/humu.1380060119

Bunge S, Kleijer WJ, Steglich C, et al. Mucopolysaccharidosis type I: identification of 8 novel mutations and determination of the frequency of the two common alpha-L-iduronidase mutations (W402X and Q70X) among European patients. Hum Mol Genet. 1994;3(6):861-866. doi:10.1093/hmg/3.6.861

Burton BK, Charrow J, Hoganson GE, et al. Newborn Screening for Lysosomal Storage Disorders in Illinois: The Initial 15-Month Experience. J Pediatr. 2017;190:130-135. doi:10.1016/j.jpeds.2017.06.048

Capriotti E, Calabrese R, Casadio R. Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. Bioinformatics. 2006;22(22):2729-2734. doi:10.1093/bioinformatics/btl423

Capriotti E, Calabrese R, Fariselli P, Martelli PL, Altman RB, Casadio R. WS-SNPs&GO: a web server for predicting the deleterious effect of human protein variants using functional annotation. BMC Genomics. 2013;14 Suppl 3(Suppl 3):S6. doi:10.1186/1471-2164-14-S3-S6

Carter H, Douville C, Stenson PD, Cooper DN, Karchin R. Identifying Mendelian disease genes with the variant effect scoring tool. BMC Genomics. 2013;14 Suppl 3(Suppl 3):S3. doi:10.1186/1471-2164-14-S3-S3

Chistiakov DA, Savost'anov KV, Kuzenkova LM, et al. Molecular characteristics of patients with glycosaminoglycan storage disorders in Russia. Clin Chim Acta. 2014;436:112-120. doi:10.1016/j.cca.2014.05.010

Chkioua L, Khedhiri S, Kassab A, et al. Molecular analysis of mucopolysaccharidosis type I in Tunisia: identification of novel mutation and eight Novel polymorphisms. Diagn Pathol. 2011;6:39. Published 2011 Apr 26. doi:10.1186/1746-1596-6-39

Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the functional effect of amino acid substitutions and indels. PLoS One. 2012;7(10):e46688. doi:10.1371/journal.pone.0046688

Chuang CK, Lin HY, Wang TJ, et al. Status of newborn screening and follow up investigations for Mucopolysaccharidoses I and II in Taiwan. Orphanet J Rare Dis. 2018;13(1):84. Published 2018 May 25. doi:10.1186/s13023-018-0816-4

Chun S, Fay JC. Identification of deleterious mutations within three human genomes. Genome Res. 2009;19(9):1553-1561. doi:10.1101/gr.092619.109

Clarke LA, Giugliani R, Guffon N, et al. Genotype-phenotype relationships in mucopolysaccharidosis type I (MPS I): Insights from the International MPS I Registry. Clin Genet. 2019;96(4):281-289. doi:10.1111/cge.13583

Clarke LA, Nelson PV, Warrington CL, Morris CP, Hopwood JJ, Scott HS. Mutation analysis of 19 North American mucopolysaccharidosis type I patients: identification of two additional frequent mutations. Hum Mutat. 1994;3(3):275-282. doi:10.1002/humu.1380030316

Clarke LA, Scott HS. Two novel mutations causing mucopolysaccharidosis type I detected by single strand conformational analysis of the alpha-L-iduronidase gene. Hum Mol Genet. 1993;2(8):1311-1312. doi:10.1093/hmg/2.8.1311

Cobos PN, Steglich C, Santer R, Lukacs Z, Gal A. Dried blood spots allow targeted screening to diagnose mucopolysaccharidosis and mucolipidosis. JIMD Rep. 2015;15:123-132. doi:10.1007/8904_2014_308

Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput Biol. 2010;6(12):e1001025. Published 2010 Dec 2. doi:10.1371/journal.pcbi.1001025

Delgado Luengo WN, Miranda Contreras LE, Chávez CJ, Solis-Añez E, Cammarata-Scalisi F. Mutation c.1190-1delG/N in intron 8 and c.1708G>C/N in exon 12 not reported in the *IDUA* gene developed a clinical phenotype of Scheie syndrome. Invest Clin. 2014;55(4):365-370.

Dong C, Wei P, Jian X, et al. Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. Hum Mol Genet. 2015;24(8):2125-2137. doi:10.1093/hmg/ddu733

Dou W, Peng C, Zheng JK, Gu XF. Zhonghua Yi Xue Yi Chuan Xue Za Zhi. 2007;24(2):136-139.

Feng BJ. PERCH: A Unified Framework for Disease Gene Prioritization. Hum Mutat. 2017;38(3):243-251. doi:10.1002/humu.23158

Ferrer-Costa C, Gelpí JL, Zamakola L, Parraga I, de la Cruz X, Orozco M. PMUT: a web-based tool for the annotation of pathological mutations on proteins. Bioinformatics. 2005;21(14):3176-3178. doi:10.1093/bioinformatics/bti486

Fokkema IF, Taschner PE, Schaafsma GC, Celli J, Laros JF, den Dunnen JT. LOVD v.2.0: the next generation in gene variant databases. Hum Mutat. 2011;32(5):557-563. doi:10.1002/humu.21438

Fu Y, Liu Z, Lou S, et al. FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. Genome Biol. 2014;15(10):480. doi:10.1186/s13059-014-0480-5

Fuller M, Brooks DA, Evangelista M, Hein LK, Hopwood JJ, Meikle PJ. Prediction of neuropathology in mucopolysaccharidosis I patients. Mol Genet Metab. 2005;84(1):18-24. doi:10.1016/j.ymgme.2004.09.004

Furukawa Y, Hamaguchi A, Nozaki I, et al. Cervical pachymeningeal hypertrophy as the initial and cardinal manifestation of mucopolysaccharidosis type I in monozygotic twins with a novel mutation in the alpha-L-iduronidase gene. J Neurol Sci. 2011;302(1-2):121-125. doi:10.1016/j.jns.2010.11.022

Ghosh A, Mercer J, Mackinnon S, et al. *IDUA* mutational profile and genotype-phenotype relationships in UK patients with Mucopolysaccharidosis Type I. Hum Mutat. 2017;38(11):1555-1568. doi:10.1002/humu.23301

González-Pérez A, López-Bigas N. Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. Am J Hum Genet. 2011;88(4):440-449. doi:10.1016/j.ajhg.2011.03.004

Guffon N, Souillet G, Maire I, Straczek J, Guibaud P. Follow-up of nine patients with Hurler syndrome after bone marrow transplantation. J Pediatr. 1998;133(1):119-125. doi:10.1016/s0022-3476(98)70201-x

Gulko B, Hubisz MJ, Gronau I, Siepel A. A method for calculating probabilities of fitness consequences for point mutations across the human genome. Nat Genet. 2015;47(3):276-283. doi:10.1038/ng.3196

Harris CR, Millman KJ, van der Walt SJ, et al. Array programming with NumPy. Nature. 2020;585(7825):357-362. doi:10.1038/s41586-020-2649-2

Huang YF, Gulko B, Siepel A. Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. Nat Genet. 2017;49(4):618-624. doi:10.1038/ng.3810

Hunter J. D. , Matplotlib: A 2D Graphics Environment, Computing in Science & Engineering, vol. 9, no. 3, pp. 90-95, 2007.

Ioannidis NM, Rothstein JH, Pejaver V, et al. REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. Am J Hum Genet. 2016;99(4):877-885. doi:10.1016/j.ajhg.2016.08.016

Jagadeesh KA, Wenger AM, Berger MJ, et al. M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. Nat Genet. 2016;48(12):1581-1586. doi:10.1038/ng.3703

Kamranjam M, Alaei M. Mutation Analysis of the *IDUA* Gene in Iranian Patients with Mucopolysaccharidosis Type 1: Identification of Four Novel Mutations. Genet Test Mol Biomarkers. 2019;23(8):515-522. doi:10.1089/gtmb.2019.0022

Karczewski KJ, Francioli LC, Tiao G, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581(7809):434-443. doi:10.1038/s41586-020-2308-7

Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014;46(3):310-315. doi:10.1038/ng.2892

Kubaski F, de Oliveira Poswar F, Michelin-Tirelli K, et al. Mucopolysaccharidosis Type I. Diagnostics (Basel). 2020;10(3):161. Published 2020 Mar 16. doi:10.3390/diagnostics10030161

Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc. 2009;4(7):1073-1081. doi:10.1038/nprot.2009.86

Kwak MJ, Huh R, Kim J, Park HD, Cho SY, Jin DK. Report of 5 novel mutations of the α-L-iduronidase gene and comparison of Korean mutations in relation with those of Japan or China in patients with mucopolysaccharidosis I. BMC Med Genet. 2016;17(1):58. Published 2016 Aug 12. doi:10.1186/s12881-016-0319-x

Landrum MJ, Lee JM, Riley GR, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. 2014;42(Database issue):D980-D985. doi:10.1093/nar/gkt1113

Langereis EJ, van den Berg IET, Halley DJJ, et al. Considering Fabry, but Diagnosing MPS I: Difficulties in the Diagnostic Process. JIMD Rep. 2013;9:117-120. doi:10.1007/8904_2012_189

Laradi S, Tukel T, Erazo M, et al. Mucopolysaccharidosis I: Alpha-L-Iduronidase mutations in three Tunisian families. J Inherit Metab Dis. 2005;28(6):1019-1026. doi:10.1007/s10545-005-0197-4

Lee IJ, Hwang SH, Jeon BH, et al. Mutational analysis of the alpha-L-iduronidase gene in 10 unrelated Korean type I mucopolysaccharidosis patients: Identification of four novel mutations. Clin Genet. 2004;66(6):575-576. doi:10.1111/j.1399-0004.2004.00374.x

Lee-Chen GJ, Lin SP, Chen IS, Chang JH, Yang CW, Chin YW. Mucopolysaccharidosis type I: Identification and characterization of mutations affecting alpha-L-iduronidase activity. J Formos Med Assoc. 2002;101(6):425-428.

Lee-Chen GJ, Lin SP, Tang YF, Chin YW. Mucopolysaccharidosis type I: characterization of novel mutations affecting alpha-L-iduronidase activity. Clin Genet. 1999;56(1):66-70. doi:10.1034/j.1399-0004.1999.560109.x

Lee-Chen GJ, Wang TR. Mucopolysaccharidosis type I: identification of novel mutations that cause Hurler/Scheie syndrome in Chinese families. J Med Genet. 1997;34(11):939-941. doi:10.1136/jmg.34.11.939

Lehman TJ, Miller N, Norquist B, Underhill L, Keutzer J. Diagnosis of the mucopolysaccharidoses. Rheumatology (Oxford). 2011;50 Suppl 5:v41-v48. doi:10.1093/rheumatology/ker390

Lek M, Karczewski KJ, Minikel EV, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536(7616):285-291. doi:10.1038/nature19057

Li B, Krishnan VG, Mort ME, et al. Automated inference of molecular mechanisms of disease from amino acid substitutions. Bioinformatics. 2009;25(21):2744-2750. doi:10.1093/bioinformatics/btp528

Li P, Wood T, Thompson JN. Diversity of mutations and distribution of single nucleotide polymorphic alleles in the human alpha-L-iduronidase (*IDUA*) gene. Genet Med. 2002;4(6):420-426. doi:10.1097/00125817-200211000-00004

Lin SP, Lin HY, Wang TJ, et al. A pilot newborn screening program for Mucopolysaccharidosis type I in Taiwan. Orphanet J Rare Dis. 2013;8:147. Published 2013 Sep 22. doi:10.1186/1750-1172-8-147

Liu X, Li C, Mou C, Dong Y, Tu Y. dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. Genome Med. 2020;12(1):103. Published 2020 Dec 2. doi:10.1186/s13073-020-00803-9

Malhis N, Jacobson M, Jones SJM, Gsponer J. LIST-S2: taxonomy based sorting of deleterious missense mutations across species. Nucleic Acids Res. 2020;48(W1):W154-W161. doi:10.1093/nar/gkaa288

Matte U, Yogalingam G, Brooks D, et al. Identification and characterization of 13 new mutations in mucopolysaccharidosis type I patients. Mol Genet Metab. 2003;78(1):37-43. doi:10.1016/s1096-7192(02)00200-7

Naslavsky MS, Yamamoto GL, de Almeida TF, et al. Exomic variants of an elderly cohort of Brazilians in the ABraOM database. Hum Mutat. 2017;38(7):751-763. doi:10.1002/humu.23220

Niroula A, Urolagin S, Vihinen M. PON-P2: prediction method for fast and reliable identification of harmful variants. PLoS One. 2015;10(2):e0117380. Published 2015 Feb 3. doi:10.1371/journal.pone.0117380

Niroula A, Vihinen M. How good are pathogenicity predictors in detecting benign variants?. PLoS Comput Biol. 2019;15(2):e1006481. Published 2019 Feb 11. doi:10.1371/journal.pcbi.1006481

Nykamp K, Anderson M, Powers M, et al. Sherloc: a comprehensive refinement of the ACMG-AMP variant classification criteria [published correction appears in Genet Med. 2020 Jan;22(1):240-242]. Genet Med. 2017;19(10):1105-1117. doi:10.1038/gim.2017.37

Oussoren E, Keulemans J, van Diggelen OP, et al. Residual α-L-iduronidase activity in fibroblasts of mild to severe Mucopolysaccharidosis type I patients. Mol Genet Metab. 2013;109(4):377-381. doi:10.1016/j.ymgme.2013.05.016

Pasqualim G, Ribeiro MG, da Fonseca GG, et al. p.L18P: a novel *IDUA* mutation that causes a distinct attenuated phenotype in mucopolysaccharidosis type I patients. Clin Genet. 2015;88(4):376-380. doi:10.1111/cge.12507

Pedregosa et al., Scikit-learn: Machine Learning in Python, JMLR 12, pp. 2825-2830, 2011.

Pineda T, Marie S, Gonzalez J, et al. Genotypic and bioinformatic evaluation of the alpha-l-iduronidase gene and protein in patients with mucopolysaccharidosis type I from Colombia, Ecuador and Peru. Mol Genet Metab Rep. 2014;1:468-473. Published 2014 Oct 30. doi:10.1016/j.ymgmr.2014.10.001

Pollard LM, Jones JR, Wood TC. Molecular characterization of 355 mucopolysaccharidosis patients reveals 104 novel mutations. J Inherit Metab Dis. 2013;36(2):179-187. doi:10.1007/s10545-012-9533-7

Prommajan K, Ausavarat S, Srichomthong C, Puangsricharern V, Suphapeetiporn K, Shotelersuk V. A novel p.E276K *IDUA* mutation decreasing α-L-iduronidase activity causes mucopolysaccharidosis type I. Mol Vis. 2011;17:456-460. Published 2011 Feb 11.

Quang D, Chen Y, Xie X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. Bioinformatics. 2015;31(5):761-763. doi:10.1093/bioinformatics/btu703

Raiman J, D'Aco K. An 8-year-old girl with a history of stiff and painful joints. Pediatr Ann. 2014;43(8):307-309. doi:10.3928/00904481-20140723-05

Reva B, Antipin Y, Sander C. Determinants of protein function revealed by combinatorial entropy optimization. Genome Biol. 2007;8(11):R232. doi:10.1186/gb-2007-8-11-r232

Richards S, Aziz N, Bale 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. Genet Med. 2015;17(5):405-424. doi:10.1038/gim.2015.30

Ritchie GR, Dunham I, Zeggini E, Flicek P. Functional annotation of noncoding sequence variants. Nat Methods. 2014;11(3):294-296. doi:10.1038/nmeth.2832

Schwarz JM, Rödelsperger C, Schuelke M, Seelow D. MutationTaster evaluates disease-causing potential of sequence alterations. Nat Methods. 2010;7(8):575-576. doi:10.1038/nmeth0810-575

Scott CR, Elliott S, Buroker N, et al. Identification of infants at risk for developing Fabry, Pompe, or mucopolysaccharidosis-I from newborn blood spots by tandem mass spectrometry. J Pediatr. 2013;163(2):498-503. doi:10.1016/j.jpeds.2013.01.031

Scott HS, Anson DS, Orsborn AM, et al. Human alpha-L-iduronidase: cDNA isolation and expression. Proc Natl Acad Sci U S A. 1991;88(21):9695-9699. doi:10.1073/pnas.88.21.9695

Scott HS, Bunge S, Gal A, Clarke LA, Morris CP, Hopwood JJ. Molecular genetics of mucopolysaccharidosis type I: diagnostic, clinical, and biological implications. Hum Mutat. 1995;6(4):288-302. doi:10.1002/humu.1380060403

Scott HS, Litjens T, Hopwood JJ, Morris CP. PCR detection of two RFLPs in exon I of the alpha-L-iduronidase (*IDUA*) gene. Hum Genet. 1992;90(3):327. doi:10.1007/BF00220095

Scott HS, Litjens T, Nelson PV, Brooks DA, Hopwood JJ, Morris CP. alpha-L-iduronidase mutations (Q70X and P533R) associate with a severe Hurler phenotype. Hum Mutat. 1992;1(4):333-339. doi:10.1002/humu.1380010412

Scott HS, Litjens T, Nelson PV, et al. Identification of mutations in the alpha-L-iduronidase gene (*IDUA*) that cause Hurler and Scheie syndromes. Am J Hum Genet. 1993;53(5):973-986.

Scott HS, Nelson PV, Litjens T, Hopwood JJ, Morris CP. Multiple polymorphisms within the alpha-L-iduronidase gene (*IDUA*): implications for a role in modification of MPS-I disease phenotype. Hum Mol Genet. 1993;2(9):1471-1473. doi:10.1093/hmg/2.9.1471

Sherry ST, Ward MH, Kholodov M, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001;29(1):308-311. doi:10.1093/nar/29.1.308

Shihab HA, Gough J, Cooper DN, et al. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. Hum Mutat. 2013;34(1):57-65. doi:10.1002/humu.22225

Stenson PD, Mort M, Ball EV, Shaw K, Phillips A, Cooper DN. The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. Hum Genet. 2014;133(1):1-9. doi:10.1007/s00439-013-1358-4

Sun A, Hopwood JJ, Thompson J, Cederbaum SD. Combined Hurler and Sanfilippo syndrome in a sibling pair. Mol Genet Metab. 2011;103(2):135-137. doi:10.1016/j.ymgme.2011.02.011

Sun L, Li C, Song X, Zheng N, Zhang H, Dong G. Three novel α-L-iduronidase mutations in 10 unrelated Chinese mucopolysaccharidosis type I families. Genet Mol Biol. 2011;34(2):195-200. doi:10.1590/s1415-47572011005000006

Sundaram L, Gao H, Padigepati SR, et al. Predicting the clinical impact of human mutation with deep neural networks [published correction appears in Nat Genet. 2019 Feb;51(2):364]. Nat Genet. 2018;50(8):1161-1170. doi:10.1038/s41588-018-0167-z

Taghikhani M, Khatami S, Abdi M, et al. Mutation analysis and clinical characterization of Iranian patients with mucopolysaccharidosis type I. J Clin Lab Anal. 2019;33(8):e22963. doi:10.1002/jcla.22963

Teng YN, Wang TR, Hwu WL, Lin SP, Lee-Chen GJ. Identification and characterization of -3c-g acceptor splice site mutation in human alpha-L-iduronidase associated with mucopolysaccharidosis type IH/S. Clin Genet. 2000;57(2):131-136. doi:10.1034/j.1399-0004.2000.570207.x

The pandas development team, pandas-dev/pandas: Pandas. Zenodo, 2020. doi: 10.5281/zenodo.3509134

Thomas PD, Campbell MJ, Kejariwal A, et al. PANTHER: a library of protein families and subfamilies indexed by function. Genome Res. 2003;13(9):2129-2141. doi:10.1101/gr.772403

Tieu PT, Bach G, Matynia A, Hwang M, Neufeld EF. Four novel mutations underlying mild or intermediate forms of alpha-L-iduronidase deficiency (MPS IS and MPS IH/S). Hum Mutat. 1995;6(1):55-59. doi:10.1002/humu.1380060111

Uttarilli A, Ranganath P, Matta D, et al. Identification and characterization of 20 novel pathogenic variants in 60 unrelated Indian patients with mucopolysaccharidoses type I and type II. Clin Genet. 2016;90(6):496-508. doi:10.1111/cge.12795

Vaser R, Adusumalli S, Leng SN, Sikic M, Ng PC. SIFT missense predictions for genomes. Nat Protoc. 2016;11(1):1-9. doi:10.1038/nprot.2015.123

Vazna A, Beesley C, Berna L, et al. Mucopolysaccharidosis type I in 21 Czech and Slovak patients: mutation analysis suggests a functional importance of C-terminus of the *IDUA* protein. Am J Med Genet A. 2009;149A(5):965-974. doi:10.1002/ajmg.a.32812

Venturi N, Rovelli A, Parini R, et al. Molecular analysis of 30 mucopolysaccharidosis type I patients: evaluation of the mutational spectrum in Italian population and identification of 13 novel mutations. Hum Mutat. 2002;20(3):231. doi:10.1002/humu.9051

Viana GM, de Lima NO, Cavaleiro R, et al. Mucopolysaccharidoses in northern Brazil: Targeted mutation screening and urinary glycosaminoglycan excretion in patients undergoing enzyme replacement therapy. Genet Mol Biol. 2011;34(3):410-415. doi:10.1590/S1415-47572011005000025

Vijay S, Wraith JE. Clinical presentation and follow-up of patients with the attenuated phenotype of mucopolysaccharidosis type I. Acta Paediatr. 2005;94(7):872-877. doi:10.1111/j.1651-2227.2005.tb02004.x

Voskoboeva EY, Krasnopolskaya XD, Mirenburg TV, Weber B, Hopwood JJ. Molecular genetics of mucopolysaccharidosis type I: mutation analysis among the patients of the former Soviet Union. Mol Genet Metab. 1998;65(2):174-180. doi:10.1006/mgme.1998.2745

Wang X, Zhang W, Shi H, et al. Mucopolysaccharidosis I mutations in Chinese patients: identification of 27 novel mutations and 6 cases involving prenatal diagnosis [published correction appears in Clin Genet. 2012 May;81(5):501]. Clin Genet. 2012;81(5):443-452. doi:10.1111/j.1399-0004.2011.01680.x

WANG XN, SHI HP, ZHANG WM, et al. Zhonghua Yi Xue Yi Chuan Xue Za Zhi. 2011;28(2):147-151. doi:10.3760/cma.j.issn.1003-9406.2011.02.006

Wasserstein MP, Caggana M, Bailey SM, et al. The New York pilot newborn screening program for lysosomal storage diseases: Report of the First 65,000 Infants. Genet Med. 2019;21(3):631-640. doi:10.1038/s41436-018-0129-y

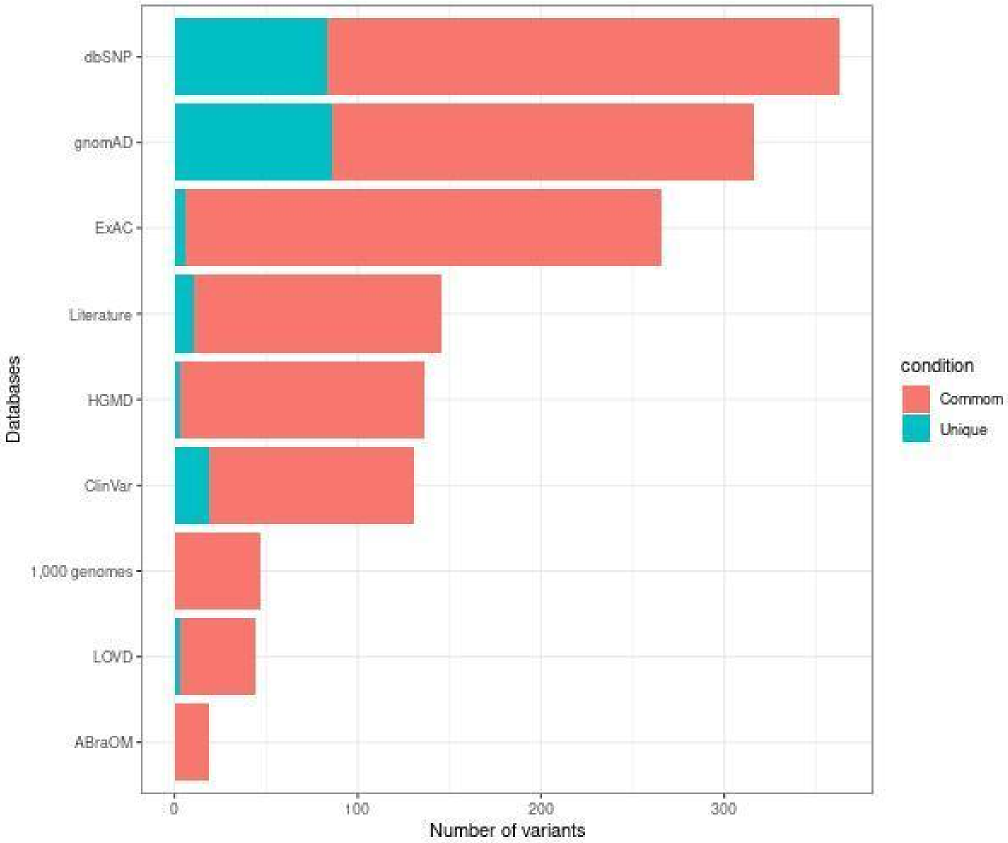
Yassaee VR, Hashemi-Gorji F, Miryounesi M, et al. Clinical, biochemical and molecular features of Iranian families with mucopolysaccharidosis: A case series. Clin Chim Acta. 2017;474:88-95. doi:10.1016/j.cca.2017.08.017

Yates CM, Filippis I, Kelley LA, Sternberg MJ. SuSPect: enhanced prediction of single amino acid variant (SAV) phenotype using network features. J Mol Biol. 2014;426(14):2692-2701. doi:10.1016/j.jmb.2014.04.026

Yogalingam G, Guo XH, Muller VJ, et al. Identification and molecular characterization of alpha-L-iduronidase mutations present in mucopolysaccharidosis type I patients undergoing enzyme replacement therapy. Hum Mutat. 2004;24(3):199-207. doi:10.1002/humu.20081

Zahoor MY, Cheema HA, Ijaz S, Anjum MN, Ramzan K, Bhinder MA. Mapping of *IDUA* gene variants in Pakistani patients with mucopolysaccharidosis type 1. J Pediatr Endocrinol Metab. 2019;32(11):1221-1227. doi:10.1515/jpem-2019-0188

Supplementary figure 1: Contributions in the number of variants from each database.



Supplementary Table 1: Pathogenic variants reported in the literature. The table presents protein consequences, the first reference, and reported level of evidence (expression study, analysis of healthy controls, and/or complete gene sequencing). Variants marked with 'Yes' had the demonstration in the first reference. The other variants feature the reference that presents the evidence. For normal controls, if the number is not indicated, at least 100 controls were analysed for comparison.

Protein Consequence		**1st Reference**	**Expression Study**	**Normal controls**	**Complete gene sequencing**
Met1Thr	M1T	Wang X. (2012)	Yes	Yes	Yes
Met1Ile	M1I	G J Lee-Chen (1997)	-	-	-
Met1Leu	M1L	Atçeken N. (2016)	-	Kamranjam M. (2019)	-
Leu14Arg	L14R	Ghosh A. (2017)	-	-	-
Leu18Pro	L18P	Pasqualim G. (2015)	-	-	-
Pro22Ser	P22S	Cobos PN. (2015)	Yes	-	-
His33Pro	H33P	Wang X. (2012)	Yes	Yes	Yes
Ala36Glu	A36E	Vijay S. (2005)	-	-	-
Gly51Asp	G51D	Bunge S. (1994)	Bunge S. (1998)	Yes	-
Phe52Leu	F52L	Wang X. (2012)	Yes	Yes	Yes
Ala75Pro	A75P	Voskoboeva E.Y. (1998)	-	-	-
Ala75Thr	A75T	Clarke L. (1994)	Tieu P. (1995)	Yes	-
Tyr76Cys	Y76C	Bertola F. (2011)	-	Yes	Yes
Ala79Val	A79V	Lee-Chen GJ (2002)	Yes, Yogalingam G. (2004)	-	-
His82Pro	H82P	Clarke LA (1993)	-	Yes	-
Gly84Ser	G84S	Scott CR. (2013)	Yes	-	Yes
Gly84Ala	G84A	Bravo H. (2017)	-	-	-
Gly84Arg	G84R	Bertola F. (2011)	Taghikhani M. (2019)	Yes	Yes
Arg89Gln	R89Q	Scott H. (1993)	Bunge S. (1998)	-	Matte U (2003)
Arg89Trp	R89W	Bunge S. (1995)	Bunge S. (1998)	48 Controls	-
Thr103Pro	T103P	Bertola F. (2011)	-	Yes	Yes
Tyr109His	Y109H	Taghikhani M. (2019)	Yes	60 Controls	Yes
Asp115Asn	D115N	Chuang CK. (2018)	Yes	-	-
Asp119Tyr	D119Y	Scott CR. (2013)	Yes	-	Yes
Leu121Pro	L121P	Uttarilli A.(2016)	-	-	-
Met133Ile	M133I	Matte U (2003)	Yes	-	Yes
Gly134Val	G134V	Kamranjam M. (2019)	-	Yes	-
Thr141Ser	T141S	Amr K. (2009)	-	-	-
Ser157Pro	S157P	Taghikhani M. (2019)	Yes	Kamranjam M. (2019)	Yes
Ala160Asp	A160D	Venturi N. (2002)	-	-	-
Arg162Ile	R162I	Li P. (2002)	-	Yes	-
Arg162Lys	R162K	Lin SP. (2013)	Yes	-	-
Gly168Val	G168V	Wang X. (2012)	Yes	Yes	Yes
Trp175Arg	W175R	Yassaee VR. (2017)	-	Kamranjam M. (2019)	Yes

Phe177Ser	F177S	Chkioua L. (2011)	-	-	-
Glu178Lys	E178K	Venturi N. (2002)	-	-	-
Thr179Lys	T179K	Ghosh A. (2017)	-	-	-
Thr179Arg	T179R	Wang X. (2012)	Yes	Yes	Yes
Trp180Ser	W180S	Pollard LM. (2013)	-	-	-
Glu182Asp	E182D	Wang X. (2012)	Yes	Yes	Yes
Glu182Ala	E182A	Brooks DA (2001)	Yes	-	-
Glu182Lys	E182K	Brooks DA (2001)	Yes	-	Matte U (2003)
Pro183Arg	P183R	Venturi N. (2002)	-	-	-
Gly197Asp	G197D	Venturi N. (2002)	-	-	-
Gly197Ser	G197S	Resumo:Wang XN (2011)	Wang X. (2012)	Wang X. (2012)	Wang X. (2012)
Asp203Asn	D203N	Dou et al. (2007)	-	50 Controls	-
Cys205Tyr	C205Y	Beesley C. (2001)	Yes	-	-
Ser206Leu	S206L	Chuang CK. (2018)	Yes	-	-
Gly208Asp	G208D	Li P. (2002)	-	Yes	Matte U (2003)
Gly208Val	G208V	Beesley C. (2001)	Yes	-	-
Leu209Arg	L209R	Guffon N. (1998)	Yes	-	-
Leu218Pro	L218P	Bunge S. (1994)	Oussoren E. (2013)	Yes	-
Gly219Glu	G219E	Bertola F. (2011)	-	Yes	Yes
Gly220Asp	G220D	Ghosh A. (2017)	-	-	-
Pro228Gln	P228Q	Lee IJ (2004)	-	-	Yes
Leu237Arg	L237R	Wang X. (2012)	Yes	Yes	Yes
Leu238Arg	L238R	Wang X. (2012)	Yes	Yes	Yes
Leu238Gln	L238Q	Yogalingam G. (2004)	Yes	-	-
His240Arg	H240R	Beesley C. (2001)	Yes	-	-
His240Asp	H240D	Chistiakov DA. (2014)	Yes	-	Yes
Gly253Cys	G253C	Uttarilli A.(2016)	-	-	-
Asp257His	D257H	Taghikhani M. (2019)	Yes	60 Controls	Yes
Ser260Phe	S260F	Matte U (2003)	Yes	-	Yes
Gly265Asp	G265D	Ghosh A. (2017)	-	-	-
Gly265Arg	G265R	Yogalingam G. (2004)	Yes	-	-
Ile270Ser	I270S	Laradi S. (2005)	-	-	-
Glu276Lys	E276K	Prommajan K., (2011)	Oussoren E. (2013)	Yes	-
Glu299Ala	E299A	Brooks DA (2001)	Yes	-	-
Asp301Glu	D301E	Taghikhani M. (2019)	Yes	Kamranjam M. (2019)	Yes
Leu303Pro	L303P	Zahoor MY. (2019)	-	Yes	-
Trp306Leu	W306L	Bertola F. (2011)	-	Yes	Yes
Leu308Pro	L308P	Sun A. (2011)	Yes	-	Yes
Asp315Tyr	D315Y	Scott H, (1995)	Vazna A. (2009)	-	-
Ala319Val	A319V	Beesley C. (2001)	Yes	-	-

Lys324Arg	K324R	Uttarilli A.(2016)	-	-	-
Ala327Pro	A327P	Bunge S. (1994)	Bunge S. (1998)	Yes	Matte U (2003)
Leu346Arg	L346R	Teng YN (2000)	Yes	-	-
Ser347Pro	S347R	Raiman J. (2014)	Yes	-	-
Asn348Lys	N348K	Bertola F. (2011)	Chistiakov DA (2014)	Yes	Yes
Asp349Asn	D349N	Scott H, (1995)	-	-	-
Asp349Tyr	D349Y	Venturi N. (2002)	Matte U (2003)	-	Matte U (2003)
Asn350Ile	N350I	Matte U (2003)	Yes	-	Yes
Arg363His	R363H	Sun L. (2011)	Yes	-	-
Arg363Cys	R363C	Yogalingam G. (2004)	Yes	-	-
Thr364Met	T364M	G J Lee-Chen (1997)	1997 e 1999	-	-
Thr366Pro	T366P	Bach G.(1993)	Yes	-	-
Asn372Ser	N372S	Uttarilli A.(2016)	-	-	-
Thr374Asn	T374N	Furukawa Y. (2011)	-	-	-
Gln380Arg	Q380R	Scott H, (1995)	-	-	Matte U (2003)
Arg383His	R383H	Bunge S. (1995)	Bunge S. (1998)	48 Controls	Matte U (2003)
Pro385Leu	P385L	Pineda T. (2014)	-	-	-
Pro385Arg	P385R	Bertola F. (2011)	-	Yes	Yes
Leu387Pro	L387P	Uttarilli A.(2016)	-	-	-
Thr388Lys	T388K	Kwak MJ. (2016)	-	-	-
Leu396Pro	L396P	Bertola F. (2011)	-	Yes	Yes
Leu396Arg	L396R	Ghosh A. (2017)	-	-	-
Leu421Pro	L421P	Wang X. (2012)	Yes	Yes	Yes
Ser423Arg	S423R	Yogalingam G. (2004)	Yes	-	-
Ala436Pro	A436P	Bertola F. (2011)	-	Yes	Yes
Arg489Pro	R489P	Bunge S. (1994)	Bunge S. (1998)	-	-
Leu490Pro	L490P	Tieu P. (1995)	Bunge S. (1998)	98 controls-Bunge S. (1995)	-
Arg492Pro	R492P	Tieu P. (1995)	Yes	-	-
Pro496Leu	P496L	Tieu P. (1995)	Yes	-	-
Pro496Arg	P496R	Beesley C. (2001)	Yes	-	-
Met504Arg	M504R	Uttarilli A.(2016)	-	-	-
Met504Thr	M504T	Bunge S. (1995)	Bunge S. (1998)	48 Controls	-
Leu526Pro	L526P	Pollard LM. (2013)	-	-	-
Pro533Arg	P533R	Scott H. (1992)	Bunge S. (1998)	-	-
Pro533Leu	P533L	Voskoboeva E.Y. (1998)	-	80 Controls	-
Leu535Phe	L535F	Bertola F. (2011)	-	Yes	Yes
Asp570His	D570H	Luengo WN. (2014)	Yes	-	-
Cys577Tyr	C577Y	Kwak MJ. (2016)	-	-	-
Leu578Gln	L578Q	Chkioua L. (2011)	-	-	-
Phe602Ile	F602I	Yogalingam G. (2004)	Yes	-	-

Arg619Gly	R619G	Lee-Chen G. (1999)	Yes	-	-
Val620Phe	V620F	Vazna A. (2009)	Yes	-	-
Arg621Leu	R621L	Pineda T. (2014)	-	-	-
Leu623Pro	L623P	Cobos PN. (2015)	Yes	-	-
Asp624Val	D624V	Ghosh A. (2017)	-	-	-
Tyr625Cys	Y625C	Pineda T. (2014)	-	-	-
Tyr625Ser	Y625S	Chuang CK. (2018)	Yes	-	-
Trp626Arg	W626R	Bunge S. (1995)	Bunge S. (1998)	48 Controls	-
Arg628Pro	R628P	Matte U (2003)	Yes	-	Yes
Ser633Leu	S633L	Beesley C. (2001)	Yes	-	-
Ser633Trp	S633W	Cobos PN. (2015)	Yes	-	-

Supplementary Table 2: Benign variants reported in the literature. The table presents protein consequences, variant classification, the first reference, and reported level of evidence (expression study, analysis of normal controls, and/or complete gene sequencing). Variants marked with 'Yes' possessed the evidence in the first reference. The other options feature the reference that presents the evidence.

Protein Consequence		Classification	1st Reference	Expression Study	Normal controls	Complete gene sequencing
Leu4Pro	L4P	Polymorphism	Venturi N. (2002)	-	-	-
Ala24Asp	A24D	Polymorphism	Lee IJ (2004)	-	-	Yes
His33Gln	H33Q	Polymorphism	Scott H. (1992)	-	Bertola F. (2011)	Bertola F. (2011)
Gln63Pro	Q63P	Polymorphism	Scott H. (1991)	Yes	-	-
Asn73His	N73H	Polymorphism	Amr K. (2009)	-	-	-
Ala79Thr	A79T	Pseudodeficiency	Wasserstein MP. (2019)	Yes	-	-
His82Gln	H82Q	Pseudodeficiency	Yogalingam G. (2004)	Yes	-	-
Thr99Ile	T99I	Pseudodeficiency	Wasserstein MP. (2019)	Yes	-	-
Arg105Gln	R105Q	Polymorphism	Scott H. (1991)	Yes	Bertola F. (2011)	Bertola F. (2011)
Gly116Arg	G116R	Polymorphism	Bunge S. (1994)	Wasserstein MP. (2019)	-	-
Asp223Asn	D223N	Pseudodeficiency	Wasserstein MP. (2019)	Yes	-	-
Arg263Trp	R263W	Pseudodeficiency	Langereis EJ. (2013)	Yes	-	-
Ala300Thr	A300T	Pseudodeficiency	Aronovich EL. (1996)	Yes	-	-
Val322Glu	V322E	Pseudodeficiency	Burton BK. (2017)	Yes	-	-
Ala361Thr	A361T	Polymorphism	Scott H. (1993)	-	Bertola F. (2011)	Bertola F. (2011)
Arg363Ser	R363S	Polymorphism	Amr K. (2009)	-	-	-
Gly409Arg	G409R	Polymorphism	Bach G. (1993)	Yes	-	-
His449Asn	H449N	Polymorphism	Bertola F. (2011)	-	Yes	Yes
Val454Ile	V454I	Polymorphism	Bunge S. (1994)	-	Bertola F. (2011)	Bertola F. (2011)
Ala591Thr	A591T	Polymorphism	Beesley C. (2001)	Yes	Bertola F. (2011)	Bertola F. (2011)

Supplementary Table 3: Values of sensitivity, specificity, accuracy, positive predictive value (PPV), negative predictive value (NPV), false positive rate (FPR) and kappa value for the stricter criteria subset.

Programs	Sensitivity	Specificity	Accuracy	Positive predictive value (PPV)	Negative predictive value (NPV)	False positive rate (FPR)	Kappa Value
SIFT_p	0.790	0.760	0.790	0.950	0.410	0.240	0.409
SIFT_g	0.710	0.820	0.730	0.960	0.350	0.180	0.347
SIFT4G	0.820	0.760	0.810	0.950	0.450	0.240	0.458
Polyphen2_HDIV	0.900	0.470	0.830	0.900	0.470	0.530	0.372
Polyphen2_HVAR	0.890	0.530	0.830	0.910	0.470	0.470	0.400
LRT	0.740	0.590	0.710	0.910	0.290	0.410	0.231
MutationTaster	0.910	0.530	0.850	0.910	0.530	0.470	0.441
MutationAssessor	0.910	0.650	0.870	0.930	0.580	0.350	0.532
PROVEAN_p	0.860	0.650	0.820	0.930	0.460	0.350	0.432
PROVEAN_g	0.760	0.760	0.760	0.950	0.370	0.240	0.366
VEST4	0.950	0.590	0.890	0.920	0.670	0.410	0.560
MetaSVM	0.800	0.710	0.790	0.940	0.400	0.290	0.388
MetaLR	0.810	0.530	0.770	0.900	0.350	0.470	0.282
MCAP	1.000	0.000	0.890	0.890	-	1.000	-,
REVEL	0.750	0.880	0.770	0.970	0.390	0.120	0.419
MutPred_db	0.600	0.750	0.610	0.970	0.130	0.250	0.110
MutPred_site	0.840	0.710	0.810	0.930	0.480	0.290	0.458
PrimateAI	0.180	0.760	0.270	0.800	0.150	0.240	-0.021
DEOGEN2	0.960	0.470	0.880	0.910	0.670	0.530	0.485
BayesDel_addAF	0.980	0.820	0.950	0.970	0.880	0.180	0.821
BayesDel_noAF	0.960	0.710	0.920	0.950	0.750	0.290	0.678
ClinPred	0.920	0.880	0.920	0.980	0.680	0.120	0.719
LISTS2	0.730	0.710	0.720	0.930	0.320	0.290	0.292
CADD_raw	0.980	0.000	0.820	0.840	0.000	1.000	-0.034
CADD	0.950	0.350	0.850	0.890	0.550	0.650	0.348
CADD_raw_hg19	0.980	0.060	0.830	0.850	0.330	0.940	0.055
CADD_hg19	0.930	0.240	0.820	0.870	0.400	0.760	0.203
FATHMMMKL	0.840	0.530	0.790	0.900	0.380	0.470	0.312
FATHMMXF	0.740	0.760	0.740	0.940	0.350	0.240	0.339
FATHMM_CV_W	0.930	0.060	0.800	0.840	0.140	0.940	-0.009
FATHMM_C V_U	0.660	0.880	0.690	0.970	0.330	0.120	0.320
integrated_fitCons	1.000	0.000	0.840	0.840	-	1.000	-
GM12878_fitCons	1.000	0.000	0.840	0.840	-	1.000	-
H1hESC_fitCons	0.930	0.000	0.790	0.830	0.000	1.000	-0.089

HUVEC_fitCons	0.890	0.120	0.770	0.840	0.170	0.880	0.009
GERP++	0.910	0.290	0.810	0.870	0.380	0.710	0.228
PhDSNP	0.760	0.710	0.750	0.930	0.350	0.290	0.330
PANTHER	0.600	0.820	0.640	0.950	0.290	0.180	0.242
SNPs&GO	0.480	0.940	0.560	0.980	0.250	0.060	0.202
PSNPE	0.600	0.820	0.640	0.950	0.280	0.180	0.239
CADD 1.2	0.520	0.880	0.570	0.960	0.250	0.120	0.199
DANN	0.690	0.760	0.700	0.940	0.320	0.240	0.290
FATHMMMKL	0.550	0.820	0.590	0.940	0.250	0.180	0.195
FunSeq	0.000	1.000	0.160	-	0.160	0.000	-
GWAVAE	0.860	0.130	0.720	0.810	0.180	0.870	-0.008
SuSPect	0.260	0.880	0.360	0.920	0.180	0.120	0.057
PMut	0.660	0.880	0.690	0.970	0.330	0.120	0.320
CONDEL	0.960	0.410	0.870	0.900	0.640	0.590	0.429
PONP2 _I	0.910	0.880	0.900	0.980	0.580	0.130	0.646
PONP2_ P	0.910	0.880	0.910	0.980	0.580	0.130	0.647
PONP2_G	0.930	0.880	0.920	0.980	0.640	0.130	0.692

Supplementary Table 4: Values of sensitivity, specificity, accuracy, positive predictive value (PPV), negative predictive value (NPV), false positive rate (FPR) and kappa value for the less stringent criteria subset.

Programs	Sensitivity	Specificity	Accuracy	Positive predictive value (PPV)	Negative predictive value (NPV)	False positive rate (FPR)	Kappa Value
SIFT_p	0.79	0.62	0.76	0.91	0.36	0.38	0.318
SIFT_gen	0.74	0.62	0.72	0.91	0.31	0.38	0.255
SIFT4G	0.85	0.62	0.81	0.92	0.44	0.38	0.404
Polyphen2_HDIV	0.90	0.42	0.83	0.89	0.46	0.58	0.336
Polyphen2_HVAR	0.86	0.50	0.80	0.90	0.41	0.50	0.328
LRT	0.72	0.52	0.69	0.89	0.26	0.48	0.173
MutationTaster	0.92	0.50	0.85	0.90	0.54	0.50	0.431
MutationAssessor	0.92	0.58	0.86	0.92	0.58	0.42	0.493
PROVEAN_p	0.84	0.62	0.80	0.92	0.42	0.38	0.380
PROVEAN_gen	0.76	0.65	0.74	0.92	0.35	0.35	0.306
VEST4	0.92	0.54	0.86	0.91	0.56	0.46	0.464
MetaSVM	0.78	0.65	0.76	0.92	0.37	0.35	0.334
MetaLR	0.80	0.50	0.75	0.89	0.33	0.50	0.245
M-CAP	1.00	0.00	0.88	0.88	-	1.00	-
REVEL	0.74	0.81	0.75	0.95	0.38	0.19	0.373
MutPred	0.58	0.60	0.58	0.95	0.09	0.40	0.049
MutPred_site	0.79	0.65	0.76	0.91	0.43	0.35	0.367
PrimateAI	0.22	0.85	0.32	0.88	0.18	0.15	0.027
DEOGEN2	0.93	0.42	0.85	0.89	0.55	0.58	0.392
BayesDel_addAF	0.96	0.77	0.93	0.96	0.80	0.23	0.743
BayesDel_noAF	0.93	0.62	0.88	0.93	0.62	0.38	0.541
ClinPred	0.93	0.81	0.91	0.96	0.68	0.19	0.680
LIST-S2	0.72	0.65	0.71	0.92	0.31	0.35	0.263
CADD_raw_p	0.99	0.08	0.84	0.85	0.50	0.92	0.094
CADD	0.95	0.23	0.83	0.86	0.46	0.77	0.224
CADD_raw_hg19_p	0.99	0.08	0.84	0.85	0.50	0.92	0.094
CADD_hg19	0.93	0.15	0.81	0.85	0.31	0.85	0.109
fathmm-MKL_coding	0.86	0.46	0.79	0.89	0.39	0.54	0.297
fathmm-XF_coding	0.76	0.73	0.76	0.94	0.37	0.27	0.355
FATHMM (weighted)	0.90	0.08	0.76	0.83	0.13	0.92	-0.033
FATHMM (unweighted)	0.69	0.85	0.71	0.96	0.34	0.15	0.335
integrated_fitCons	1.00	0.00	0.84	0.84	-	1.00	-
GM12878_fitCons	1.00	0.00	0.84	0.84	-	1.00	-
H1-hESC_fitCons	0.93	0.00	0.78	0.83	0.00	1.00	-0.099

HUVEC_fitCons	0.87	0.12	0.75	0.84	0.15	0.88	-0.013
GERP++	0.93	0.23	0.82	0.86	0.40	0.77	0.197
PhD-SNP	0.76	0.65	0.74	0.92	0.35	0.35	0.306
PANTHER	0.62	0.77	0.64	0.93	0.29	0.23	0.229
SNPs&GO	0.49	0.85	0.55	0.94	0.24	0.15	0.170
PSNPE	0.62	0.73	0.64	0.92	0.27	0.27	0.208
CADD 1.2	0.46	0.85	0.53	0.94	0.23	0.15	0.150
DANN	0.68	0.65	0.68	0.91	0.28	0.35	0.218
FATHMM-MKL	0.58	0.77	0.61	0.93	0.26	0.23	0.198
FunSeq2	0.00	1.00	0.16	-	0.16	0.00	-
GWAVAE 1.0	0.88	0.15	0.75	0.83	0.21	0.85	0.036
SuSPect	0.28	0.88	0.38	0.93	0.19	0.12	0.065
PMut	0.64	0.88	0.68	0.97	0.32	0.12	0.309
CONDEL	0.97	0.38	0.88	0.89	0.71	0.62	0.436
PON-P2_I	0.89	0.77	0.88	0.96	0.53	0.23	0.553
PON-P2_P	0.91	0.77	0.89	0.96	0.56	0.23	0.581
PON-P2_G	0.94	0.77	0.92	0.96	0.67	0.23	0.666

Capítulo 3

Esse capítulo foca na implementação de um script de NLP para extrair as informações da literatura sobre as variantes, bem como utilizar as avaliações realizadas no capítulo anterior para comparar o resultado de diferentes proteínas. O capítulo apresenta as etapas preliminares da extração de informação, bem como os passos subsequentes.

A mineração de texto é o processo de derivar informações significativas de texto em linguagem natural e envolve ao menos cinco etapas: obtenção dos dados, limpeza, análise, visualização e extração do conhecimento (Liang et al., 2017). A obtenção dos dados depende do objetivo do trabalho. No caso de uma busca na literatura científica, os artigos são as fontes primárias de informação. A limpeza é uma etapa geral que filtra as palavras mais comuns dos textos, como preposições, conjunções, artigos, pronomes e reduz o máximo possível de palavras ao seu radical a fim de realizar uma comparação mais eficaz e direta. A etapa de análise é a que possui maior gama de métodos para ser realizada, dependendo da linguagem de programação escolhida, das bibliotecas disponíveis e do objetivo do trabalho. Python é uma linguagem que apresenta muitos recursos para esse tipo de análise, com diversas bibliotecas que trabalham desde a extração e conversão do texto total do arquivo PDF para o TXT, como a pyPDF2, até a análise de todos os dados, como Natural Language Toolkit (NLTK), Gensim, spaCy. As etapas de visualização e a extração do conhecimento são processos finais onde pode-se trabalhar com as informações contidas nos dados originais.

Os artigos curados manualmente no capítulo dois foram transformados em texto utilizando a biblioteca pyPDF2. O processo de conversão foi automatizado e testado. A língua inglesa, comum à maioria dos artigos científicos analisados, é considerada uma linguagem de altos recursos, o que facilita o seu processamento (Hirschberg and Manning, 2015).

Para o passo de análise, é necessário construir uma rede neural artificial. Isso está sendo feito utilizando a biblioteca spaCy. O resultado desejado nessa etapa é uma lista que relacione cada variantes com seu efeito clínico de acordo com o artigo analisado. Ainda não foi possível implementar a rede neural em sua totalidade em função da necessidade da adaptação do código fonte.

A predição das variantes será buscada nos dados do dbNSFP4 v.4 (Liu et al., 2020). Este banco de dados foi desenvolvido para predição funcional e anotação de todas potenciais variantes não sinônimas no genoma humano e compila escores de 37 preditores. Os pontos de corte utilizados são os padrões dos programas. A sensibilidade, especificidade, acurácia, valor kappa e o teste de fisher será realizado com as bibliotecas numpy pandas, openpyxl, scipy.stats, sklearn.metrics and matplotlib.pyplot. Esse processo também já está automatizado e validado, tendo sido utilizado no trabalho descrito no capítulo 2.

Sabe-se que a paridade no número de variantes patogênicas e benignas é importante para que se possa calcular corretamente a taxa de falsos positivos e negativos e realizar as análises estatísticas (Jørgensen et al., 2018; Vadillo et al., 2016). Porém, a maioria das variantes relatadas são encontradas em estudos que descrevem pacientes e focam na busca de variantes causais, o que gera uma uma desproporção de variantes patogênicas reportadas na literatura. A utilização de variantes encontradas em bancos de dados populacionais pode auxiliar no incremento de variantes não patogênicas. Apesar de difundida, essa estratégia deve ser tratada com atenção, pois normalmente depende do estabelecimento de um ponto de corte de frequência, que pode ser variável de acordo com a doença (Kopanos et al., 2019).

Além disso, os resultados dos testes de acurácia de preditores variam de acordo com certos parâmetros usados na validação. Por exemplo, os grupos de treinamento e teste precisam ter tamanhos substanciais e, para obter esse volume de dados, os programas juntam informações de todas as variantes disponíveis na literatura, de diferentes proteínas (Niroula and Vihinen, 2019; Adzhubei et al., 2010; Ng et al., 2001). Porém, é razoável supor que um preditor possa avaliar bem um certo grupo proteico mas seus resultados sejam menos confiáveis para outro. No entanto, como o resultado é avaliado pela combinação dos diferentes grupos proteicos, essa especificidade se perde, devido à heterogeneidade do grupo de treinamento. Por isso, uma vez terminado o processo de automatização e integração de todas as etapas descritas acima, a predição para diferentes grupos de proteínas será comparada. Para análise, foram escolhidas proteínas dos seguintes grupos: enzimas, proteínas transmembrana e proteínas desestruturadas. Em cada grupo serão buscadas proteínas associadas a doenças genéticas e com o maior número de variantes com evidências na literatura. Até o momento, foram feitas as análises do subgrupo de enzimas. Foram selecionados 100 genes, pertencentes aos 7 subgrupos enzimáticos. O número final de genes utilizados na pesquisa será confirmado após a automatização da revisão da literatura.

Espera-se que, com esse *pipeline*, os pesquisadores possam, de maneira simplificada, testar quais os melhores preditores para os seus genes de interesse usando como referência dados curados da literatura. É preciso ressaltar que, vários aspectos poderão influenciar o desempenho deste teste, como frequência e penetrância da doença, as quais impactam diretamente o uso de bases de dados populacionais.

Considerações Finais

A escolha do preditor para análise de variantes impacta o resultado. Os programas apresentam diferentes estratégias para avaliação, entretanto a literatura e a análise crítica sobre cada preditor não respondem diretamente qual preditor devemos utilizar na avaliação como visto no capítulo um. Os preditores mais citados e menos citados podem utilizar os mesmos princípios de avaliação e terem as mesmas características. Não existe uma definição clara do porquê alguns preditores são mais ou menos utilizados, a não ser o ano de surgimento do preditor. Os primeiros preditores aparentam serem mais lembrados e citados que os atuais.

A avaliação de desempenho de um determinado preditor para um conjunto de variantes geralmente é realizada pela comparação com outros preditores ou com dados curados da literatura. A curadoria manual desses dados é trabalhosa, pois os artigos, apesar de serem fontes confiáveis, relatam diferentes tipos de informação. Assim, é necessário o estabelecimento de critérios para essas informações e questionar se todos os significados clínicos relacionados às variantes e relatadas na literatura são aceitáveis. Por exemplo, antigamente, reportar uma variante encontrada em algum dos éxons analisados era o suficiente, sem necessariamente realizar estudos de expressão ou comparar com um número suficiente de controles. Após a difusão e barateamento das técnicas de biologia molecular, passou a ser comum a realização do sequenciamento completo do gene do paciente, a comparação com um número significativo de controles normais, além dos estudos de expressão. No entanto, conforme demonstrado no capítulo dois, a acurácia dos preditores não parece ser necessariamente influenciada pela qualidade da informação contida na literatura. Por outro lado, nota-se que os preditores mais recentes apresentam um desempenho melhor do ponto de vista estatístico que alguns dos mais comumente utilizados.

Porém, os dados obtidos para o gene *IDUA* não podem ser automaticamente transpostos para outros genes, pois é possível que o desempenho dos preditores varie de acordo com o tipo de proteína. Portanto, essa avaliação manual de uma grande quantidade de preditores deveria ser realizada para cada gene ou pelo menos para cada família de proteínas, que possuem características similares. No entanto, essa análise é extremamente trabalhosa e não ocorre de forma automatizada. Assim, o capítulo três apresenta uma estratégia para realizar essa avaliação com os preditores disponíveis a partir de uma análise utilizando processamento de linguagem natural. Como a interpretação dos dados da literatura contém dificuldades inerentes, a implementação de um algoritmo automatizado apresenta diversos desafios e ainda não foi finalizado. Uma vez implementado, esse algoritmo poderá ser usado para comparar preditores entre diferentes grupos proteicos.

Referências Bibliográficas

Adzhubei IA, Schmidt S, Peshkin L, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010;7(4):248-249. doi:10.1038/nmeth0410-248

Castiglia D, Zambruno G. Mutation mechanisms. Dermatol Clin. 2010;28(1):17-22. doi:10.1016/j.det.2009.10.002

Choe H, Deirmengian CA, Hickok NJ, Morrison TN, Tuan RS. Molecular diagnostics. J Am Acad Orthop Surg. 2015;23 Suppl(0):S26-S31. doi:10.5435/JAAOS-D-14-00409

Dyle MC, Kolakada D, Cortazar MA, Jagannathan S. How to get away with nonsense: Mechanisms and consequences of escape from nonsense-mediated RNA decay. Wiley Interdiscip Rev RNA. 2020;11(1):e1560. doi:10.1002/wrna.1560

Esteva A, Robicquet A, Ramsundar B, et al. A guide to deep learning in healthcare. Nat Med. 2019;25(1):24-29. doi:10.1038/s41591-018-0316-z

Hirschberg J, Manning CD. Advances in natural language processing. Science. 2015;349(6245):261-266. doi:10.1126/science.aaa8685

Jian X, Boerwinkle E, Liu X. In silico tools for splicing defect prediction: a survey from the viewpoint of end users. Genet Med. 2014;16(7):497-503. doi:10.1038/gim.2013.176

Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, et al. Predicting Splicing from Primary Sequence with Deep Learning. Cell. 2019;176(3):535-548.e24. doi:10.1016/j.cell.2018.12.015

Jørgensen M, Konge L, Subhi Y. Contrasting groups' standard setting for consequences analysis in validity studies: reporting considerations. Adv Simul (Lond). 2018;3:5. Published 2018 Mar 9. doi:10.1186/s41077-018-0064-7

Kopanos C, Tsiolkas V, Kouris A, et al. VarSome: the human genomic variant search engine. Bioinformatics. 2019;35(11):1978-1980. doi:10.1093/bioinformatics/bty897

Lindeboom RG, Supek F, Lehner B. The rules and impact of nonsense-mediated mRNA decay in human cancers. Nat Genet. 2016;48(10):1112-1118. doi:10.1038/ng.3664

Ng PC, Henikoff S. Predicting deleterious amino acid substitutions. Genome Res. 2001;11(5):863-874. doi:10.1101/gr.176601

Niroula A, Vihinen M. How good are pathogenicity predictors in detecting benign variants?. PLoS Comput Biol. 2019;15(2):e1006481. Published 2019 Feb 11. doi:10.1371/journal.pcbi.1006481

Nykamp K, Anderson M, Powers M, et al. Sherloc: a comprehensive refinement of the ACMG-AMP variant classification criteria (published correction appears in Genet Med. 2020 Jan;22(1):240-242]. Genet Med. 2017;19(10):1105-1117. doi:10.1038/gim.2017.37

Richards S, Aziz N, Bale 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. Genet Med. 2015;17(5):405-424. doi:10.1038/gim.2015.30

Tang H, Thomas PD. Tools for Predicting the Functional Impact of Nonsynonymous Genetic Variation. Genetics. 2016;203(2):635-647. doi:10.1534/genetics.116.190033

Uçar MK, Nour M, Sindi H and Polat K. The Effect of Training and Testing Process on Machine Learning in Biomedical Datasets. Mathematical Problems in Engineering 2020:1-17. doi:10.1155/2020/2836236

Vadillo MA, Konstantinidis E, Shanks DR. Underpowered samples, false negatives, and unconscious learning. Psychon Bull Rev. 2016;23(1):87-102. doi:10.3758/s13423-015-0892-6

Wong YKE, Lam KW, Ho KY, et al. The applications of big data in molecular diagnostics. Expert Rev Mol Diagn. 2019;19(10):905-917. doi:10.1080/14737159.2019.1657834

Zeng Z, Bromberg Y. Predicting Functional Effects of Synonymous Variants: A Systematic Review and Perspectives. Front Genet. 2019;10:914. Published 2019 Oct 7. doi:10.3389/fgene.2019.00914

Anexo

Neste item consta um artigo publicado durante o período de mestrado em tema relacionado ao da dissertação.

Borges *et al. Orphanet J Rare Dis* (2020) 15:324
https://doi.org/10.1186/s13023-020-01608-0

Orphanet Journal of
Rare Diseases

RESEARCH

Open Access

Estimated prevalence of mucopolysaccharidoses from population-based exomes and genomes

Pâmella Borges[1,2,3], Gabriela Pasqualim[4], Roberto Giugliani[3,5,6], Filippo Vairo[7,8*] and Ursula Matte[1,2,3,5]

Abstract

Background: In this study, the prevalence of different types of mucopolysaccharidoses (MPS) was estimated based on data from the exome aggregation consortium (ExAC) and the genome aggregation database (gnomAD). The population-based allele frequencies were used to identify potential disease-causing variants on each gene related to MPS I to IX (except MPS II).

Methods: We evaluated the canonical transcripts and excluded homozygous, intronic, 3′, and 5′ UTR variants. Frameshift and in-frame insertions and deletions were evaluated using the SIFT Indel tool. Splice variants were evaluated using SpliceAI and Human Splice Finder 3.0 (HSF). Loss-of-function single nucleotide variants in coding regions were classified as potentially pathogenic, while synonymous variants outside the exon–intron boundaries were deemed non-pathogenic. Missense variants were evaluated by five in silico prediction tools, and only those predicted to be damaging by at least three different algorithms were considered disease-causing.

Results: The combined frequencies of selected variants (ranged from 127 in *GNS* to 259 in *IDUA*) were used to calculate prevalence based on Hardy–Weinberg's equilibrium. The maximum estimated prevalence ranged from 0.46 per 100,000 for MPSIIID to 7.1 per 100,000 for MPS I. Overall, the estimated prevalence of all types of MPS was higher than what has been published in the literature. This difference may be due to misdiagnoses and/or underdiagnoses, especially of the attenuated forms of MPS. However, overestimation of the number of disease-causing variants by in silico predictors cannot be ruled out. Even so, the disease prevalences are similar to those reported in diagnosis-based prevalence studies.

Conclusion: We report on an approach to estimate the prevalence of different types of MPS based on publicly available population-based genomic data, which may help health systems to be better prepared to deal with these conditions and provide support to initiatives on diagnosis and management of MPS.

Keywords: Mucopolysaccharidoses (MPS), Estimated prevalence, Exome aggregation consortium (ExAC), Genome aggregation database (gnomAD), In silico analysis

Introduction

The mucopolysaccharidoses (MPS) are a group of lysosomal diseases characterized by the deficiency of one of eleven enzymes involved in the breakdown of glycosaminoglycans (GAGs) which are constituents of the extracellular matrix. When there is a disturbance in their activities this leads to downstream consequences at the cellular level affecting multiple organs and systems. The MPS may be divided into different types according to the enzyme deficiency and the accumulated substrate (type I, II, IIIA, IIIB, IIIC, IIID, IVA, IVB, VI, VII, and IX). GAGs are constituents of the extracellular matrix,

*Correspondence: vairo.filippo@mayo.edu
[7] Center for Individualized Medicine, Mayo Clinic, Rochester, MN, USA
Full list of author information is available at the end of the article

where impaired activities can lead to a spate of negative consequences both at the cellular and the physiological levels. Affected individuals usually have coarse facial features, cardiac and pulmonary problems, and, depending on the MPS type, bone dysplasia (dysostosis multiplex) and/or neurological impairment such as behavioural problems and developmental delay [1–3]. The severity of the diseases is variable, and individuals with MPS I, II, IVA, VI, and VII may benefit from market-approved enzyme replacement therapy, while there are novel therapies such as fusion proteins, gene therapy, and genome editing under investigation for several MPS [4].

Incidence and prevalence data are important to back up health system decisions and are necessary to calculate the cost–benefit of new therapies and treatment. Despite extensive molecular characterization having been done for the genes that encode the enzymes involved in these diseases with over 2,109 pathogenic variants reported in the Human Gene Disease Database (HGMD®) [5], there is still lack of specific epidemiology data on MPS. Newborn screening programs that include lysosomal diseases have arisen worldwide and may bring valuable information. However, such programs are still largely restricted to very few countries and most types of MPS are not included in the list of screened diseases [6, 7]. Population-based genomic data can help narrow the information gap, since now it is possible to rely on carrier frequency instead of the incidence of a disease among live births. However, care must be taken when using in silico predictors to classify genetic variants in order to have the most reliable data possible.

Herein, we used the frequency of potential disease-causing variants present in population-based genomic databases such as the Exome Aggregation Consortium (ExAC) [8] and the Genome Aggregation Database (gnomAD) [9], to estimate the prevalence of the different types of MPS after applying Hardy–Weinberg principles [10].

Results

Table 1 shows the number of variants present in each database and after the merger, which ranged from 961 (*IDS*) to 2988 (*GALNS*). After subsequent filtering steps, these numbers were reduced, ranging from 31 (*IDS*) to 259 (*IDUA*) (Table 2). A detailed description of the excluded variants can be found in Additonal file 1: Table S1.

The number of variants excluded due to homozygosis ranged between 3 in *GNS* and *GUSB* to 113 in *IDS* (in homozygosis or hemizygosis); none of them were stop gain, stop loss, or start loss. The overall number of heterozygous canonical and non-canonical splice site variants considering all genes was 452, with 224 being considered deleterious by the in silico algorithms. One splice site variant could not be analysed by HSF nor SpliceAI (Additonal file 3: Table S3). In addition, 213 out of 218 frameshift and 188 in-frame insertions and deletions were considered deleterious. Variants that could not be analysed by SIFT Indel were excluded from further analysis. All variants considered deleterious by only one splice program as well as frameshift and nonsense variants in the last exon or located < 50 nucleotides upstream of the 3' most splice-generated exon-exon junction were excluded from the calculations of minimum frequency. The number of variants considered deleterious in each category is shown in Table 2.

Table 1 Number of variants in each gene present in ExAC and gnomAD

MPS type	Gene	ExAC variants	gnomAD variants	Common	Retained variants**
MPS I	*IDUA*	1246	1439	680	2005
MPS II	*IDS*	300	920	259	961
MPS IIIA	*SGSH*	1188	1400	545	2043
MPS IIIB	*NAGLU*	640	805	397	1048
MPS IIIC	*HGSNAT*	598	1456	521	1533
MPS IIID	*GNS*	429	1116	404	1141
MPS IVA	*GALNS*	1390	2254	656	2988
MPS IVB	*GLB1**	871	1322	564	1629
MPS VI	*ARSB*	407	1122	370	1159
MPS VII	*GUSB*	593	1067	519	1141
MPS IX	*HYAL1*	669	700	287	1082

*Variants may be associated with GM1 Gangliosidosis or with MPS IVB

**Retained variants represent unique variants after merging both databases

All 3,111 missense variants were analysed by five different in silico tools. A consensus on pathogenicity was reached for 588 variants, while 548 variants were classified as pathogenic by four tools and 382 variants by three.

The allele frequencies of each variant for a given gene were added together and considered as the minimum and maximum frequency of the deleterious recessive allele. This number was then used to calculate minimum and maximum prevalence of disease based on the Hardy–Weinberg equilibrium (Table 3). As the number of variants retained for *IDS* was very low (31 variants), the estimated frequency of MPS II must be viewed with caution. It is worth noticing that variants on *GLB1* can be associated either with MPS IVB or GM1 gangliosidosis.

Only two of the 2,061 retained variants have frequencies over 0.001—p.(His356Pro) in *NAGLU* with 0.007993 and p.(Asp152Asn) in *GUSB* with 0.001153. After all five tier variant selections, maximum and minimum estimated disease prevalence was calculated based on global allele frequency (Table 3).

Table 2 Number of variants considered deleterious per category for each gene

	Frameshift**	In-frame insertion/ deletion	Splice site**	Start loss	Stop gain**	Stop loss**	Missense**	Total**
IDUA	17–18	12	16–37	1	10–15	0–1	86–175	142–259
IDS	0	1	1–2	0	0	0	4–28	6–31
SGSH	8–14	7	5–7	0	4–14	0	73–194	97–236
NAGLU	11–20	2	6–10	1	8–16	0	87–176	115–225
HGSNAT	11	4	22–37	0	8–9	0	18–98	63–159
GNS	5	3	14–23	0	4	0–1	29–91	55–127
GALNS	11	7	14–26	1	10–11	0–1	57–187	100–244
*GLB1**	12–13	3	18–34	1	11–13	0	67–161	112–225
ARSB	9–12	5	10–18	0	8–12	0	48–141	80–188
GUSB	11–13	6	17–27	2	13–14	0–2	62–160	111–224
HYAL1	12–13	8	1–3	1	8–9	0	57–107	87–141
All genes	107–130	58	124–224	7	84–117	0–5	588–1515	968—2059

*Variants may be associated with GM1 Gangliosidosis or to MPS IVB

**Numbers represent minimum and maximum frequencies. In the case of frameshift, stop gain or stop loss minimum frequency excludes variants in the last exon or located < 50 nucleotides upstream of the 3′ most splice-generated exon-exon junction. For splice site and missense variants, minimum frequency considers only variants deemed pathogenic by a consensus of all software packages

Table 3 Estimated disease prevalence based on allele frequencies of potentially disease-causing variants for each gene

Gene	Disease-causing variants	CI in 100,000 (max)	CI in 100,000 (min)
IDUA	259	7.103–7.096	2.479–2.476
IDS	29	0.0108–0.0107	0.00014–0.00013
SGSH	236	2.365–2.363	0.4116–0.4112
NAGLU	225	1.532–1.530	0.366–0.365
HGSNAT	159	1.566–1.565	0.107–0.106
GNS	127	0.459–0.458	0.0549–0.0548
GALNS	224	2.363–2.361	0.25–0.25
*GLB1**	225	1.677–1.676	0.456–0.455
ARSB	188	1.119–1.117	0.1761–0.1758
GUSB	224	1.144–1.141	0.2081–0.2078
HYAL1	141	0.4393–0.4388	0.1081–0.1079

*Variants may be associated to GM1 gangliosidosis or to MPS IVB.
CI = Confidence interval

In addition to estimated overall disease prevalence, the prevalence of MPS in specific populations was calculated for eight ethnic groups present in the databases (Figs. 1, 2 and Additonal file 4: Table S4).

Discussion

In this study, we used public data from WES and WGS to estimate the prevalence of different types of MPS. As MPS symptoms usually show up in the first decade of life, it is unlikely that severely affected individuals would be part of such databases. However, the possibility of undiagnosed individuals with milder phenotypes being included in those cannot be ruled out. Importantly, individuals homozygous for rare variants present in any MPS gene (Additonal file 2: Table S2), which could represent individuals with attenuated forms of the disease were filtered out in the second-tier variant selection.

The estimated global frequency for all types of MPS except for type VI found in this study was either above or at the upper limit in comparison to frequencies of MPS in different countries based on the number of diagnosed cases in reference centres [20] (Table 4). Worthy of note is the fact that the maximum prevalence as reported by Khan et al., 2017 is for a limited number of countries, whereas our data was calculated collectively for the different ethnic backgrounds present in the databases. This means that we may have overestimated the prevalence of diseases in the general population. A recent study estimated the prevalence of MPS in Brazil based on 600 affected individuals with all types of MPS included in a national network database [21]. The researchers found discrepancy when comparing the estimated prevalence based on diagnosis (0.24/100,000) to the estimated prevalence based on genetic screening for the most common pathogenic variant in *IDUA* among healthy volunteers (0.95/100,000), for example. Furthermore, the estimated prevalence of MPS VI in Brazil was the second highest in the world, with prevalence similar to that found in the present study (1.02/100,000 compared with 1.12/100,000).

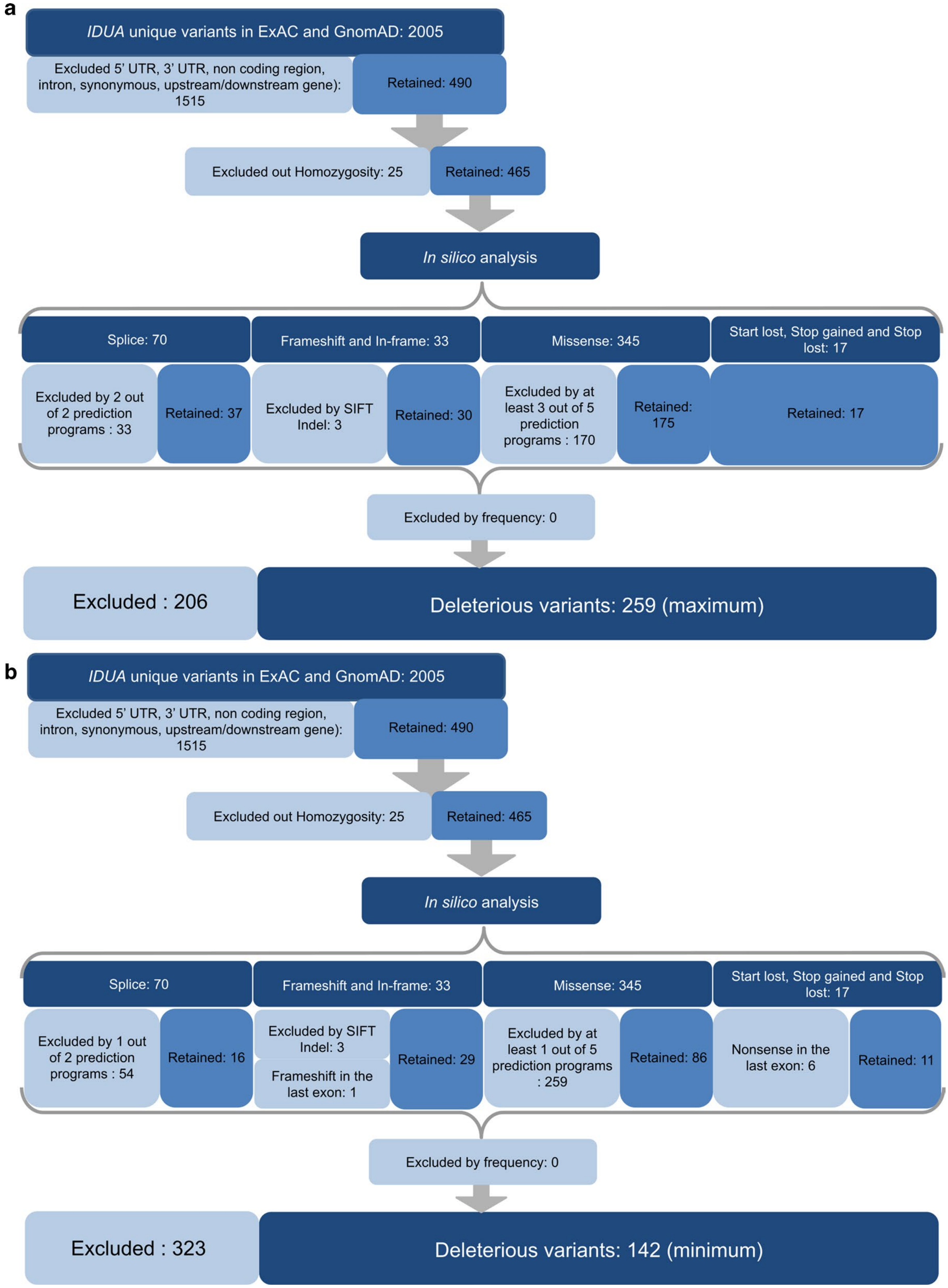

a
IDUA unique variants in ExAC and GnomAD: 2005
Excluded 5' UTR, 3' UTR, non coding region, intron, synonymous, upstream/downstream gene): 1515
Retained: 490
Excluded out Homozygosity: 25
Retained: 465
In silico analysis
Splice: 70
Frameshift and In-frame: 33
Missense: 345
Start lost, Stop gained and Stop lost: 17
Excluded by 2 out of 2 prediction programs : 33
Retained: 37
Excluded by SIFT Indel: 3
Retained: 30
Excluded by at least 3 out of 5 prediction programs : 170
Retained: 175
Retained: 17
Excluded by frequency: 0
Excluded : 206
Deleterious variants: 259 (maximum)
b
IDUA unique variants in ExAC and GnomAD: 2005
Excluded 5' UTR, 3' UTR, non coding region, intron, synonymous, upstream/downstream gene): 1515
Retained: 490
Excluded out Homozygosity: 25
Retained: 465
In silico analysis
Splice: 70
Frameshift and In-frame: 33
Missense: 345
Start lost, Stop gained and Stop lost: 17
Excluded by 1 out of 2 prediction programs : 54
Retained: 16
Excluded by SIFT Indel: 3
Frameshift in the last exon: 1
Retained: 29
Excluded by at least 1 out of 5 prediction programs : 259
Retained: 86
Nonsense in the last exon: 6
Retained: 11
Excluded by frequency: 0
Excluded : 323
Deleterious variants: 142 (minimum)

Fig. 1 Schematic example showing all steps of maximum (**a**) and minimum (**b**) variant selection for the *IDUA* gene (MPS I)

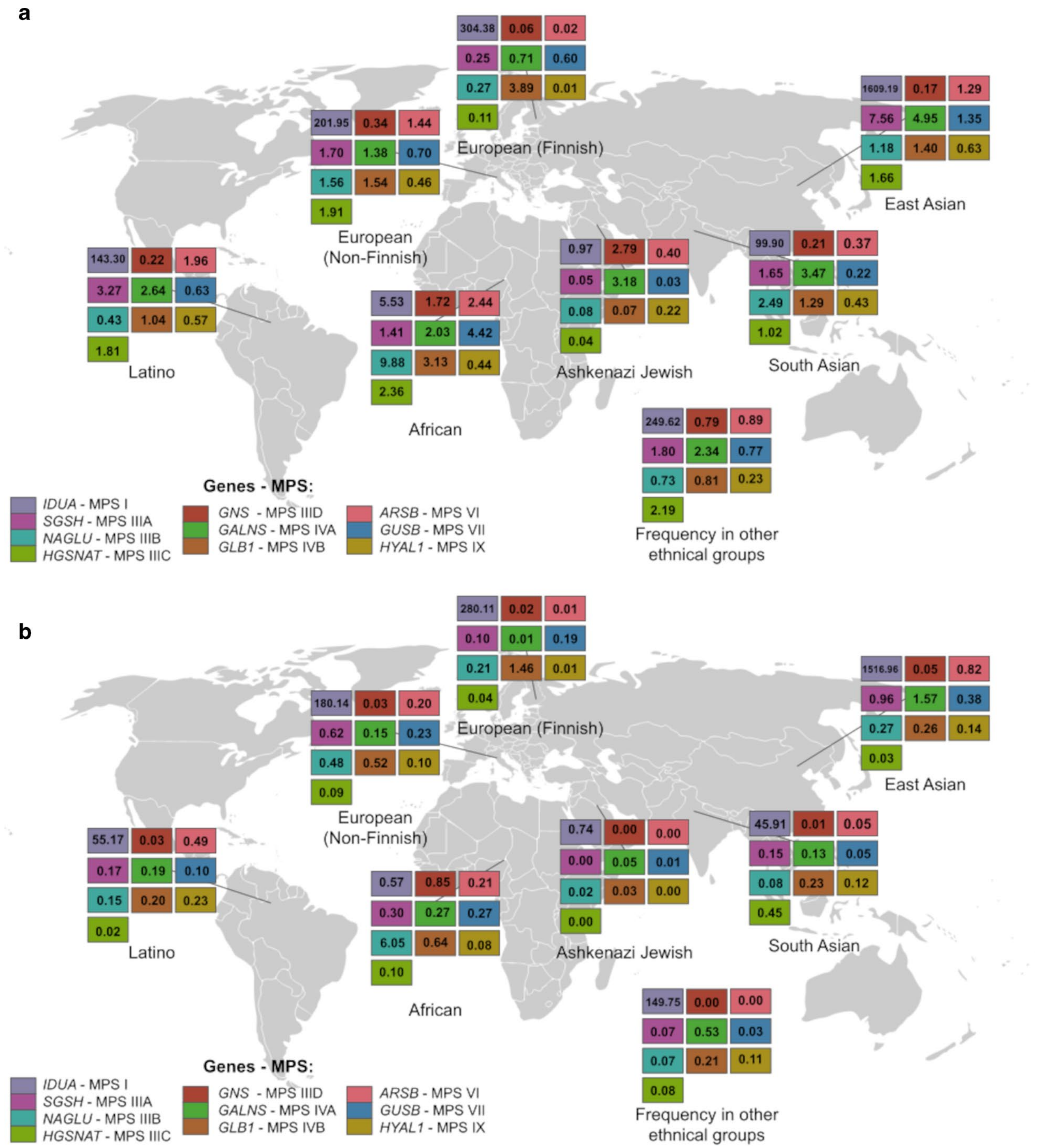


Fig. 2 Estimated maximum (**a**) and minimum (**b**) prevalence of the MPS types per 100,000 individuals in different ethnic groups. Data for MPS II not included (see discussion)

Several measures were taken to reduce the chance of prevalence overestimation. For example, variants were filtered in sequential steps, in order to obtain the most specific data possible. Also, both homozygotes and variants with frequency higher than 0.001 were excluded. Additional filtering based on functional predictions was also performed in order to include only variants more likely to affect protein function. After that, all variants remaining for analysis had allele frequencies below 0.001 and most of them have not been previously reported as

Table 4 Estimated prevalence in the present study compared to the incidence (in 100,000) as reported by Khan et al., 2017 for each MPS type

MPS type	Gene	This study (max.–min.)	Khan et al. 2017 (max.–min.)
MPS I	*IDUA*	7.10–2.48	3.62–0.11
MPS II	*IDS*	0.0108–0.00013	2.16–0.1
MPS IIIA	*SGSH*	2.36–0.41	1.62–0.08
MPS IIIB	*NAGLU*	1.53–0.37	0.72–0.02
MPS IIIC	*HGSNAT*	1.57–0.11	0.42–0.03
MPS IIID	*GNS*	0.46–0.05	0.10–0.09
MPS IVA	*GALNS*	2.36–0.25	1.30–0.15
MPS IVB	*GLB1*	1.68–0.46*	0.14–0.01
MPS VI	*ARSB*	1.12–0.18	7.85–0.02
MPS VII	*GUSB*	1.14–0.21	0.29–0.02
MPS IX	*HYAL1*	0.44–0.11	NA

*Combined frequency of GM1 Gangliosidosis and MPS IVB

disease-causing. This was expected since variants classified as of uncertain significance (VUS) based on the standards and guidelines of the American College of Medical Genetics/Association of Molecular Pathology (ACMG/AMP) [10] are known to account for a substantial part of disease-causing variants for MPS and have a significant impact on incidence estimates. For example, Clark et al. [22] showed that 25% of VUS analysed in MPS IIIB were potentially disease-causing and cause reduced enzyme activity.

It is worthy of note that sequential filtering steps and use of consensus scores do not guarantee that only pathogenic variants are selected or that only non-pathogenic variants are discarded. However, the estimation error is not directly measurable. Furthermore, the high frequency filter is necessary to exclude variants with frequencies incompatible with MPS disease. Although this may lead the possibility of underascertainment, frequencies like 0.007993 and 0.001153 for variant c.1067A > C; p.(His356Pro) in *NAGLU* and the c.454G > A; p.(Asp152Asn) in *GUSB* are not found in clinical practice. These were the only two variants excluded because of high frequency. We considered using curated variants reported either on ClinVar or Human Genome Mutation Database (HGMD), however, this would significantly reduce the number of retained variants (for instance, from 259 to 47 for *IDUA*, data not shown). Different in silico tools were used to estimate the likelihood of a variant being disease-causing. However, as no data on the sensitivity and specificity of such softwares are available for MPS genes, it is impossible to estimate the number of false-positive results. For instance, several well characterized pathogenic variants reported in HGMD had low deleteriousness scores as evaluated by the Combined Annotation-Dependent Depletion (CADD) [23] that has an overall higher performance than other predictors (data not shown).

The existence of compound heterozygotes cannot be ruled out. In fact, most individuals with MPS who are not a result of from consanguineous marriage are indeed compound heterozygotes. However, due to the structure of both databases used in this study, it is impossible to set up conditions where the occurrence of variants in *cis* cannot be ruled out, which would contribute to the overestimation of disease prevalence.

Despite these limitations, a similar approach has been used by Appadurai et al., 2015 to estimate the prevalence of cerebrotendinous xanthomatosis (CTX). As in the present study, the authors suggested an apparent underdiagnosis of CTX based on the allele frequency of potentially disease-causing variants present in ExAC. Interestingly, the discrepancy between genomic data and the diagnosis-based incidence is more pronounced for the rarest MPS diseases, such as MPS IIIC, IIID, IVB, VII, and IX. For some forms of MPS I, II, VI, and IX, it is possible that variants leading to deficient enzyme activity are not clinically recognized due to attenuated phenotypes [24–26]. On the other hand, severe cases of MPS VII may lead to premature death before the diagnosis is reached or even sought [27].

Notably, data emerging from large datasets of WES and WGS are disclosing novel phenotypes for well-known diseases, especially intermediate phenotypes [28–30]. This may also be the case for MPS and could help explain the higher prevalence predicted by our work, with patients not being recognized clinically due to an unusual presentation.

In the case of MPS IVB, there is an additional complexity since the same gene is involved in another lysosomal disorder with different accumulated substrate and clinical features, called GM1 gangliosidosis [31]. In this study, variants of *GLB1* were considered disease-causing regardless of the associated phenotype. Therefore, the overall frequency of alleles was used to estimate the prevalence of MPS IVB, whereas in fact only about 13.3% of curated disease-causing variants in this gene are associated with MPS IVB, the rest leading to the three types of GM1 gangliosidosis [32].

After the filtering steps, *IDS* had a limited number of retained disease-causing variants (29 variants), and therefore the estimated prevalence for MPS II was lower than what has been previously reported [20]. The higher prevalence observed in studies based on reference centres and diagnostic laboratories may be related to the proportion of patients having de novo variants. Pollard et al. [33] show that this happens in 22.5% of MPS

II cases. In addition, recombination events between *IDS* and its pseudogene *IDS2* are a common cause of the disease, with structural variants such as gross rearrangements and complete or partial deletions seen in between 10 and 28% of affected individuals [34–40]. Those types of variants could not be taken into account in our estimates because of the structure of the populational databases used. As a result, the estimated prevalence of MPS II is not as reliable as it is for the other types of MPS. It is worth mentioning that the other study that uses a similar method for two X-linked diseases (Menkes disease and *ATP7A*-related disorders) [41] also found a very low number of variants, which could suggest that this strategy is not the best approach for X-linked disorders.

Conclusions

In summary, we report on an approach to estimate the prevalence of the different types of MPS based on publicly available population-based genomic data that may help to better tailor screening and diagnostic programs for these diseases, to prepare the health systems to deal with a more precise estimated number of patients, and may serve as a starting point for other rare-disease initiatives.

Methods

Database

Genetic variants (GRCh37/hg19) from ExAC V0.3.1 and gnomAD v2.0.2 [8, 9] *were used to estimate the prevalence of different types of MPS. These public data aggregated information from 125,748 WES and 15,708 WGS collected from unrelated individuals and 1,756 parent–offspring trios with no known rare disease. The genetic data were collected from case–control studies of adult-onset common diseases, spanning six global and eight sub-continental ancestries, determined by ancestry-informative markers* [9]. *Although related individuals can have an influence upon the frequency of variants, the size of the database which has a total of 141,456 individuals makes the influence of 1,756 trios irrelevant.*

The data was retrieved separately for each gene, and then merged to create one single unified database. When variants were common to both databases, the allele frequencies from gnomAD were used for further analysis, as it includes ExAC data.

First-tier variant selection

Variants of the gene located in 5′ and 3′ UTR, upstream and downstream, as well as intronic and non-coding transcript exons, were excluded assuming that no disease-causing variant has been described in such positions for any MPS. In addition, synonymous variants outside the exon–intron boundaries were also excluded, as well as variants in non-canonical transcripts.

Second-tier variant selection

In second-tier analysis, missense, nonsense, stop gain and stop-loss, frameshift, and splice site variants present in homozygosis (and hemizygosis for *IDS*) were excluded based on the assumption that neither ExAC and gnomAD include MPS-affected individuals as they exclude samples from patients with severe pediatric diseases and their relatives [8]. Therefore, any homozygous variant should not be pathogenic. Heterozygous loss-of-function variants such as stop gain, stop loss, and start loss were considered as potentially disease-causing, considering the impact on protein function and strong evidence of pathogenicity as per the ACMG/AMP guidelines [10].

Third-tier variant selection

Heterozygous alterations in canonical or non-canonical splice site were analysed using Human Splice Finder [11] and SpliceAI [12]. In-frame insertions, deletions and frameshift variants outside the last exon were analysed using SIFT Indel [13]. Variants were classified based on the default algorithms parameters for deleteriousness.

Fourth-tier variant selection

The analysis of missense variants was made using five in silico algorithms: MutPred [14], PolyPhen2 [15], PROVEAN [16], SIFT [17], and REVEL [18]. Since Polyphen2 provides more than two categories, results were transformed into binary data considering "possibly pathogenic" and “probably pathogenic” as deleterious. For REVEL, an ensemble algorithm, a rank score over 0.75 was considered deleterious. To calculate the maximum prevalence of the disease, a variant was considered deleterious when at least three software packages agreed on pathogenicity. For the minimum prevalence, we included missense variants for which all in silico tools agreed on pathogenicity.

Fifth-tier variant selection

The remaining variants were analysed to make sure that only rare alleles were retained. Therefore any variant with a frequency greater than 0.001 was excluded, as no variants associated with low enzymatic activity (≤ 15% wild type) were found with higher allele frequencies [19].

Calculation of disease prevalence using Hardy–Weinberg principles

The frequency of a given variant retained as being disease-causing was calculated by dividing the number of chromosomes bearing the genetic change by the total number of chromosomes subjected to analysis in this position. Then the sum of all variant frequencies for each gene was used as the frequency of the recessive allele (q). The prevalence was then calculated as q^2, from the Hardy–Weinberg formula $p^2+2pq+q^2$. The incidence for each specific population was calculated using the population-specific frequencies.

Calculation of confidence Interval

A script in R was used to estimate the confidence interval. The variances in the frequency of variants and in the prevalence estimate were calculated equally as exhibit eqautions 5 and 13 from Clark et al. [22]. The confidence intervals were adapted to consider the sum of allele frequencies instead of probability, as suggested by Clark et al. [22].

Supplementary information

is available for this paper at https://doi.org/10.1186/s13023-020-01608-0.

Additonal file 1. The number of variants excluded at each category for each MPS gene at the calculated maximums frequency. Bold numbers identify retained variants.

Additional file 2. The total number of variants excluded for homozygosis for each MPS gene and the number of homozygosis variants with frequency less than 0.001.

Additional file 3. The number of variants excluded from the analysis for each MPS gene.

Additional file 4. The number of variants excluded from the analysis for each MPS gene.

Abbreviations

MPS: Mucopolysaccharidoses; GAGs: Glycosaminoglycans; HGMD: Human gene disease database; ExAC: Exome aggregation consortium; gnomAD: Genome aggregation database; VUS: Variants classified as of uncertain significance; CADD: Combined Annotation-Dependent Depletion; CTX: Cerebrotendinous xanthomatosis; WES: Whole exome sequencing; WGS: Whole genome sequencing.

Acknowledgements

The authors would like to thank the Research Incentive Fund of the Clinicas Hospital in Porto Alegre (*Fundo de Incentivo à Pesquisa do Hospital de Clínicas de Porto Alegre*—- FIPE/HCPA).

Authors' contributions

UM conceived the study, PB and GP collected the data; PB and FV carried out the analysis and interpretation of data; PB, UM, and FV wrote the manuscript; UM, RG, FV and GP revised the manuscript. All authors read and approved the submitted version of the manuscript.

Funding

This work was supported by the Brazilian National Council for Technological and Scientific Development (CNPq) and the Research Incentive Fund of the Clinicas Hospital in Porto Alegre (FIPE/HCPA).

Availability of data and materials

The authors confirm that the data supporting the findings of this study are available within the article [and/or] its supplementary materials.

Ethics approval and informed consent to participate

No ethical approval was required.

Consent for publication

Not applicable.

Competing interests

The authors declare no conflict of interests.

Author details

[1] Cell, Tissue and Gene Laboratory, Clinicas Hospital of Porto Alegre, Rio Grande do Sul, Brazil. [2] Experimental Research Centre, Bioinformatics Core, Clinicas Hospital of Porto Alegre, Rio Grande do Sul, Brazil. [3] Graduate Programme in Genetics and Molecular Biology, Federal University of Rio Grande Do Sul (UFRGS), Rio Grande do Sul, Brazil. [4] Genetics Laboratory, Biological Sciences Institute, Federal University of Rio Grande (FURG), Rio Grande do Sul, Brazil. [5] Department of Genetics, UFRGS, Porto Alegre, Brazil. [6] Medical Genetics Service, HCPA, Porto Alegre, Brazil. [7] Center for Individualized Medicine, Mayo Clinic, Rochester, MN, USA. [8] Department of Clinical Genomics, Mayo Clinic, Rochester, MN, USA.

Received: 4 August 2020 Accepted: 9 November 2020
Published online: 18 November 2020

References

1. Muenzer J. Overview of the mucopolysaccharidoses. Rheumatology (Oxford). 2011;50(5):v4–12. https://doi.org/10.1093/rheumatology/ker394.
2. Giugliani R. Mucopolysacccharidoses: From understanding to treatment, a century of discoveries. Genet Mol Biol. 2012;35(Suppl 4):924–31. https://doi.org/10.1590/s1415-47572012000600006.
3. Sun A. Lysosomal storage disease overview. Ann Transl Med. 2018;6(24):476. https://doi.org/10.21037/atm.2018.11.39.
4. Giugliani R, Federhen A, Vairo F, et al. Emerging drugs for the treatment of mucopolysaccharidoses. Expert Opin Emerg Drugs. 2016;21(1):9–26. https://doi.org/10.1517/14728214.2016.1123690.
5. Stenson PD, Mort M, Ball EV, Shaw K, Phillips A, Cooper DN. The human gene mutation database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. Hum Genet. 2014;133(1):1–9. https://doi.org/10.1007/s00439-013-1358-4.
6. Robinson BH, Gelb MH. The importance of assay imprecision near the screen cutoff for newborn screening of lysosomal storage diseases. Int J Neonatal Screen. 2019;5(2):17. https://doi.org/10.3390/ijns5020017.
7. Schielen PCJI, Kemper EA, Gelb MH. Newborn screening for lysosomal storage diseases: a concise review of the literature on screening methods, therapeutic possibilities and regional programs. Int J Neonatal Screen. 2017;3(2):6. https://doi.org/10.3390/ijns3020006.
8. Lek M, Karczewski KJ, Minikel EV, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536(7616):285–91. https://doi.org/10.1038/nature19057.
9. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. bioRxiv. 2019;531210. Available from: https://www.biorxiv.org/content/https://doi.org/10.1101/531210v2
10. Appadurai V, DeBarber A, Chiang PW, et al. Apparent underdiagnosis of cerebrotendinous xanthomatosis revealed by analysis of ~60,000 human exomes. Mol Genet Metab. 2015;116(4):298–304. https://doi.org/10.1016/j.ymgme.2015.10.010.
11. Desmet FO, Hamroun D, Lalande M, Collod-Béroud G, Claustres M, Béroud C. Human splicing finder: an online bioinformatics tool to predict splicing signals. Nucleic Acids Res. 2009;37(9):e67. https://doi.org/10.1093/nar/gkp215.
12. Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, et al. Predicting splicing from primary sequence with deep learning. Cell. 2019;176(3):535-548.e24. https://doi.org/10.1016/j.cell.2018.12.015.
13. Hu J, Ng PC. SIFT Indel: predictions for the functional effects of amino acid insertions/deletions in proteins. PLoS One. 2013;8(10):e77940. Published 2013 Oct 23; doi:https://doi.org/10.1371/journal.pone.0077940

14. Li B, Krishnan VG, Mort ME, et al. Automated inference of molecular mechanisms of disease from amino acid substitutions. Bioinformatics. 2009;25(21):2744–50. https://doi.org/10.1093/bioinformatics/btp528.
15. Adzhubei IA, Schmidt S, Peshkin L, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010;7(4):248–9. https://doi.org/10.1038/nmeth0410-248.
16. Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the functional effect of amino acid substitutions and indels. PLoS ONE. 2012;7(10):e46688. https://doi.org/10.1371/journal.pone.0046688.
17. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc. 2009;4(7):1073–81. https://doi.org/10.1038/nprot.2009.86.
18. Ioannidis NM, Rothstein JH, Pejaver V, et al. REVEL: an Ensemble method for predicting the pathogenicity of rare missense variants. Am J Hum Genet. 2016;99(4):877–85. https://doi.org/10.1016/j.ajhg.2016.08.016.
19. Clarke LA, Giugliani R, Guffon N, et al. Genotype-phenotype relationships in mucopolysaccharidosis type I (MPS I): Insights from the International MPS I registry. Clin Genet. 2019;96(4):281–9. https://doi.org/10.1111/cge.13583.
20. Khan SA, Peracha H, Ballhausen D, et al. Epidemiology of mucopolysaccharidoses. Mol Genet Metab. 2017;121(3):227–40. https://doi.org/10.1016/j.ymgme.2017.05.016.
21. Federhen A, Pasqualim G, de Freitas TF, et al. Estimated birth prevalence of mucopolysaccharidoses in Brazil. Am J Med Genet A. 2020;182(3):469–83. https://doi.org/10.1002/ajmg.a.61456.
22. Clark WT, Yu GK, Aoyagi-Scharber M, LeBowitz JH. Utilizing ExAC to assess the hidden contribution of variants of unknown significance to Sanfilippo Type B incidence. PLoS One. 2018;13(7):e0200008. https://doi.org/10.1371/journal.pone.0200008.
23. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. 2019;47(D1):D886–94. https://doi.org/10.1093/nar/gky1016.
24. Kiykim E, Barut K, Cansever MS, et al. Screening mucopolysaccharidosis Type IX in patients with juvenile idiopathic arthritis. JIMD Rep. 2016;25:21–4. https://doi.org/10.1007/8904_2015_467.
25. Pinto E, Vairo F, Conboy E, de Souza CFM, et al. Diagnosis of attenuated mucopolysaccharidosis VI: clinical, biochemical, and genetic pitfalls. Pediatrics. 2018;142(6):e20180658. https://doi.org/10.1542/peds.2018-0658.
26. Rigoldi M, Verrecchia E, Manna R, Mascia MT. Clinical hints to diagnosis of attenuated forms of Mucopolysaccharidoses. Ital J Pediatr. 2018;44(Suppl 2):132. https://doi.org/10.1186/s13052-018-0551-4.
27. Sands MS. Mucopolysaccharidosis type VII: a powerful experimental system and therapeutic challenge. Pediatr Endocrinol Rev. 2014;12(Suppl 1):159–65.
28. Bonafé L, Kariminejad A, Li J, et al. Brief report: peripheral osteolysis in adults linked to ASAH1 (Acid Ceramidase) mutations: a new presentation of farber's disease. Arthritis Rheumatol. 2016;68(9):2323–7. https://doi.org/10.1002/art.39659.
29. Kim SY, Choi SA, Lee S, et al. Atypical presentation of infantile-onset farber disease with novel ASAH1 mutations. Am J Med Genet A. 2016;170(11):3023–7. https://doi.org/10.1002/ajmg.a.37846.
30. Yu FPS, Amintas S, Levade T, Medin JA. Acid ceramidase deficiency: farber disease and SMA-PME. Orphanet J Rare Dis. 2018;13(1):121. https://doi.org/10.1186/s13023-018-0845-z.
31. Lee JS, Choi JM, Lee M, et al. Diagnostic challenge for the rare lysosomal storage disease: late infantile GM1 gangliosidosis. Brain Dev. 2018;40(5):383–90. https://doi.org/10.1016/j.braindev.2018.01.009.
32. Caciotti A, Garman SC, Rivera-Colón Y, et al. GM1 gangliosidosis and Morquio B disease: an update on genetic alterations and clinical findings. Biochim Biophys Acta. 2011;1812(7):782–90. https://doi.org/10.1016/j.bbadis.2011.03.018.
33. Pollard LM, Jones JR, Wood TC. Molecular characterization of 355 mucopolysaccharidosis patients reveals 104 novel mutations. J Inherit Metab Dis. 2013;36(2):179–87. https://doi.org/10.1007/s10545-012-9533-7.
34. Bunge S, Rathmann M, Steglich C, et al. Homologous nonallelic recombinations between the iduronate-sulfatase gene and pseudogene cause various intragenic deletions and inversions in patients with mucopolysaccharidosis type II. Eur J Hum Genet. 1998;6(5):492–500. https://doi.org/10.1038/sj.ejhg.5200213.
35. Brusius-Facchin AC, Schwartz IV, Zimmer C, et al. Mucopolysaccharidosis type II: identification of 30 novel mutations among Latin American patients. Mol Genet Metab. 2014;111(2):133–8. https://doi.org/10.1016/j.ymgme.2013.08.011.
36. Kosuga M, Mashima R, Hirakiyama A, et al. Molecular diagnosis of 65 families with mucopolysaccharidosis type II (Hunter syndrome) characterized by 16 novel mutations in the IDS gene: Genetic, pathological, and structural studies on iduronate-2-sulfatase. Mol Genet Metab. 2016;118(3):190–7. https://doi.org/10.1016/j.ymgme.2016.05.003.
37. Chiong MA, Canson DM, Abacan MA, Baluyot MM, Cordero CP, Silao CL. Clinical, biochemical and molecular characteristics of Filipino patients with mucopolysaccharidosis type II - Hunter syndrome. Orphanet J Rare Dis. 2017;12(1):7. https://doi.org/10.1186/s13023-016-0558-0.
38. Dvorakova L, Vlaskova H, Sarajlija A, et al. Genotype-phenotype correlation in 44 Czech, Slovak, Croatian and Serbian patients with mucopolysaccharidosis type II. Clin Genet. 2017;91(5):787–96. https://doi.org/10.1111/cge.12927.
39. Zanetti A, D'Avanzo F, Rigon L, et al. Molecular diagnosis of patients affected by mucopolysaccharidosis: a multicenter study. Eur J Pediatr. 2019;178(5):739–53. https://doi.org/10.1007/s00431-019-03341-8.
40. Zhang W, Xie T, Sheng H, et al. Genetic analysis of 63 Chinese patients with mucopolysaccharidosis type II: Functional characterization of seven novel IDS variants. Clin Chim Acta. 2019;491:114–20. https://doi.org/10.1016/j.cca.2019.01.009.
41. Kaler SG, Ferreira CR, Yam LS. Estimated birth prevalence of Menkes disease and ATP7A-related disorders based on the Genome Aggregation Database (gnomAD). Mol Genet Metab Rep. 2020;5(24):100602. https://doi.org/10.1016/j.ymgmr.2020.100602.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.