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*.

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*.

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.

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).

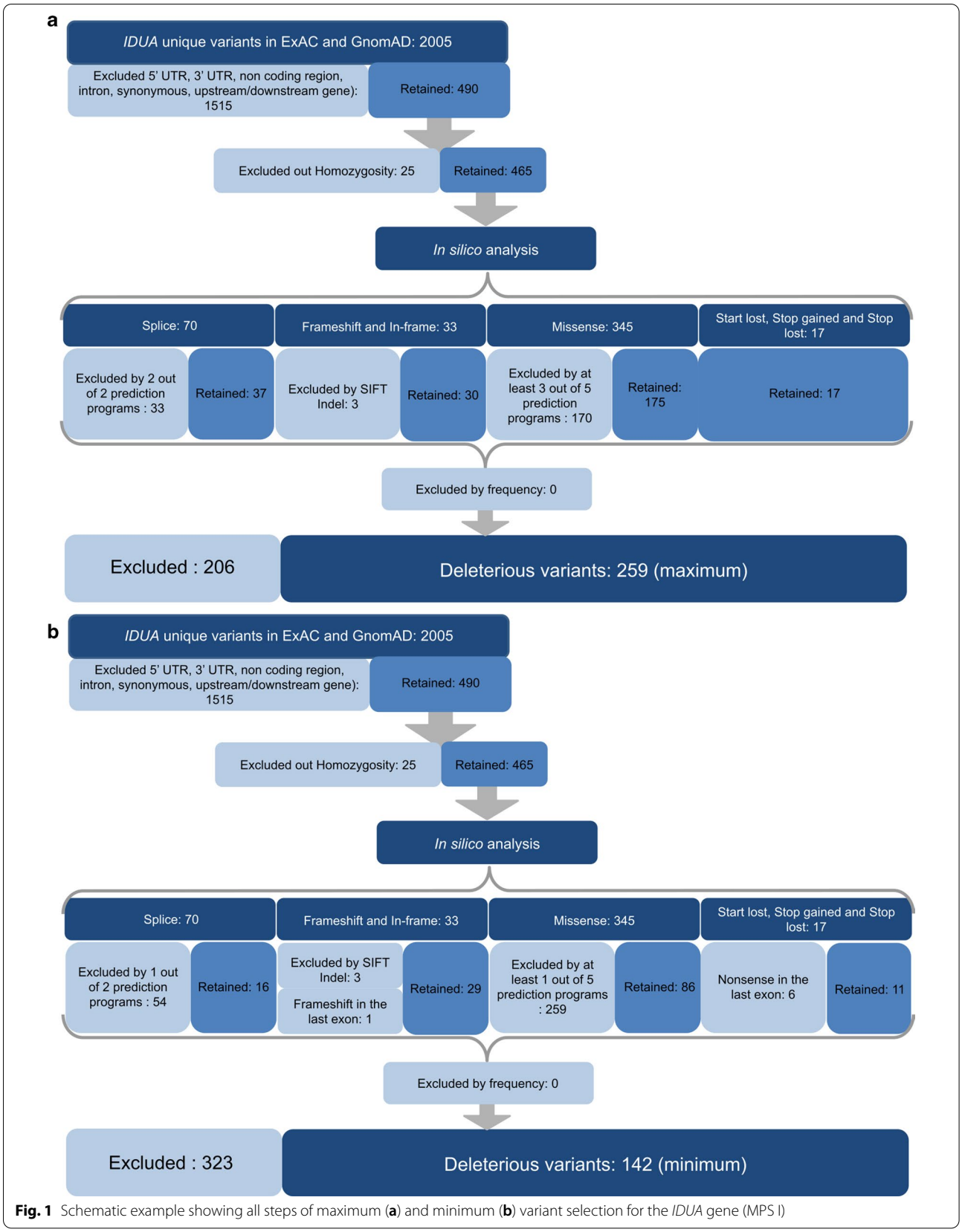

Fig. 1 Schematic example showing all steps of maximum (**a**) and minimum (**b**) variant selection for the *IDUA* gene (MPS I)

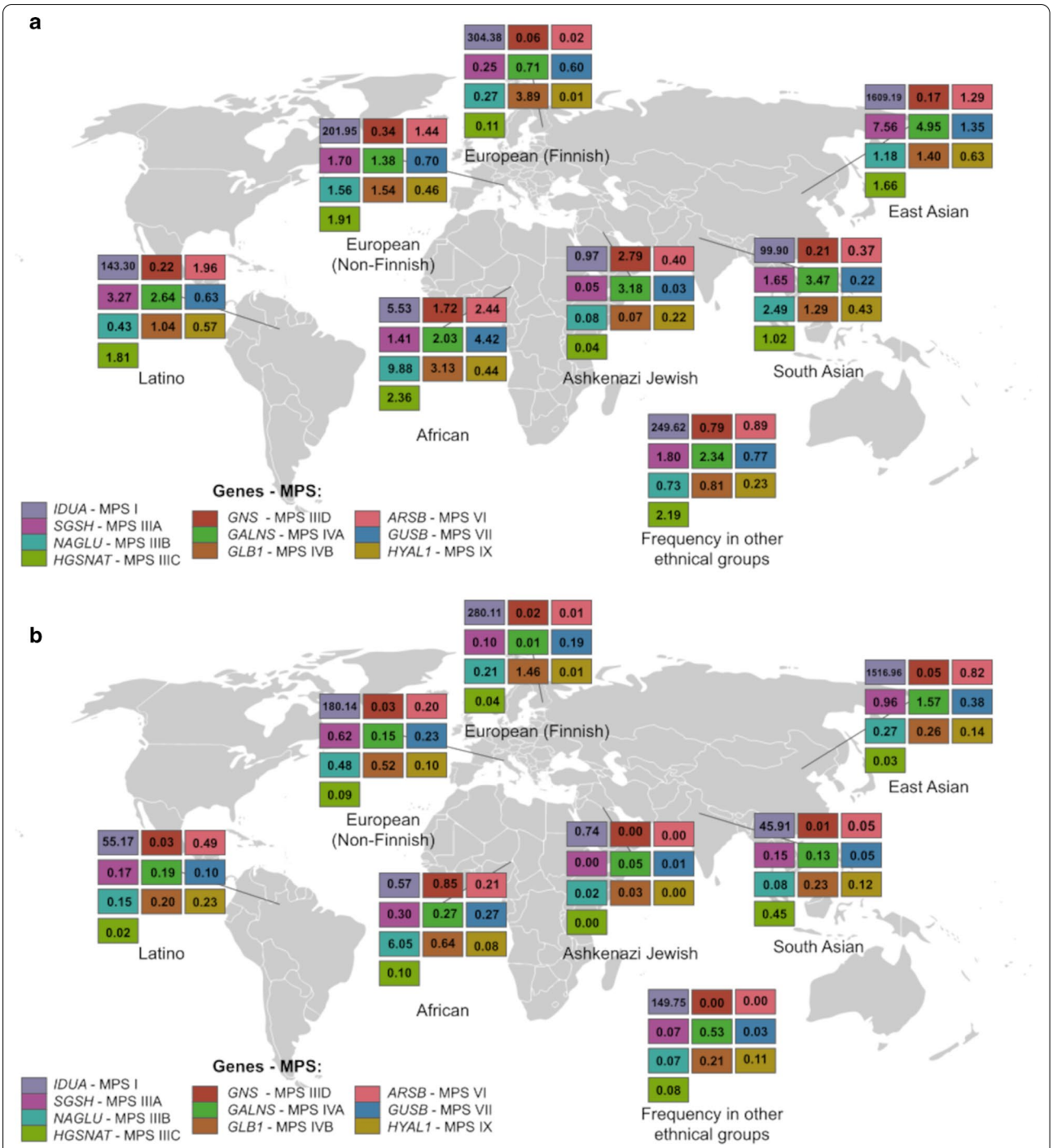

Fig. 2 Estimated maximum (**a**) and minimum (**b**) prevalence of the MPS types per 100,000 individuals in different ethnic groups. Data for MPS II not included (see discussion)

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 ($\leq$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