Computational and Structural Biotechnology Journal 23 (2024) 2615–2622

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

A Copernican revolution of multigenic analysis: A retrospective study on clinical exome sequencing in unclear genetic disorders

M. Chetta [a,*], M. Tarsitano [a], M. Rivieccio [a], M. Oro [a], A.L. Cammarota [b], M. De Marco [b], L. Marzullo [b], A. Rosati [b], N. Bukvic [c]

[a] A.O.R.N. A. Cardarelli Hospital's Laboratory of Medical Genetics and Genomics, Naples, Italy
[b] StressBioLab, Department of Medicine, Surgery and Dentistry "Schola Medica Salernitana," University of Salerno, Baronissi, SA, Italy
[c] U.O.C Genetica Medica, Azienda Ospedaliero – Universitaria Consorziale Policlinico di Bari, Bari, IT, Italy

ARTICLE INFO

Keywords:
Multigenic Analysis
Mendelian diseases
Clinical Exome Sequencing

ABSTRACT

Despite the inevitable shift in medical practice towards a deeper understanding of disease etiology and progression through multigenic analysis, the profound historical impact of Mendelian diseases cannot be overlooked. These diseases, such as cystic fibrosis and thalassemia, are characterized by a single variant in a single gene leading to clinical conditions, and have significantly shaped our medical knowledge and treatments. In this respect, the monogenic approach inevitably results in the underutilization of Next-Generation Sequencing (NGS) data.

Herein, a retrospective study was performed to assess the diagnostic value of the clinical exome in 32 probands with specific phenotypic characteristics (patients with autoinflammation and immunological dysregulation, N = 20; patients diagnosed with Hemolytic uremic syndrome N = 9; and patients with Waldenström macroglobulinemia, N = 3). A gene enrichment analysis was performed using the *. VCF file generated by SOPHiA-DDM-v4. This analysis selected a subset of genes containing pathogenic or likely pathogenic variants with autosomal dominant (AD) inheritance. In addition, all variants of uncertain significance (VUS) were included, filtered by AD inheritance mode, the presence of compound heterozygotes, and a minor allele frequency (MAF) cutoff of 0.05 %.

The aim of the pipeline described here is based on a perspective shift that focuses on analyzing patients' gene assets, offering new light on the complex interplay between genetics and disease presentation. Integrating this approach into clinical practices could significantly enhance the management of patients with rare genetic disorders.

1. Introduction

The latest advances in Next-Generation Sequencing (NGS) technology, in addition to the ongoing refinement of software tools and data processing pipelines, have significantly improved our ability to analyze the genetic complexity of diseases [1,2]. Despite significant advances in understanding the genetic basis of numerous medical conditions, the shift to multigene analysis represents a significant break from the classic Mendelian genetic paradigm. Mendelian diseases, which are characterized by changes in a single gene, have historically dominated genetic research and diagnostic testing. Disorders such as thalassemia and cystic fibrosis are prominent examples of this monogenic framework, and the discovery of pathogenic variants in genes is critical in clinical practice [3].

Nevertheless, accumulating evidence suggests that many complicated diseases are caused by complex interactions between several genes, resulting in a paradigm shift towards multigene analyses [3]. This transition represents the recognition of the complex genetic landscape that supports specific disease problems beyond the simplicity of the Mendelian framework. As our understanding increases, the need to study numerous genes simultaneously becomes clearer. Multigene analysis not only broadens the field of genetic research but also acknowledges the interplay of gene interactions in generating disease clinical manifestations. This change marks a significant advancement in our attempt to unravel the rich genetic tapestry that leads to complex disease conditions [3].

* Corresponding author.
E-mail address: massimiliano.chetta@aocardarelli.it (M. Chetta).

https://doi.org/10.1016/j.csbj.2024.06.011
Received 29 March 2024; Received in revised form 8 June 2024; Accepted 8 June 2024
Available online 15 June 2024
2001-0370/© 2024 Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Due to a variety of circumstances, the NGS diagnosis yield has varied, with rates ranging from 25 % to 50 % [4,5]. These factors include the genetic and allelic diversity of the disease, criteria employed for patient recruitment, clinical manifestations, choice of sequencing platforms, and specific laboratory analytical methods applied [4]. The essence of precision medicine hinges on accurately deciphering the genetic condition of each patient. The complex interplay between these variables can either enhance or limit the efficacy of NGS in clinical diagnostics [6].

Herein, a multigenic enriched analysis approach is proposed as an innovative methodology that requires an extension investigation of data obtained from clinical exome sequencing (CES).

This strategy, similar to the Copernican heliocentric revolution, aims to reorient the analysis by focusing on a unique subset of genes in each individual. The goal was to determine how all these genetic factors contribute to the development of disease or clinical symptoms.

A meticulous evaluation of the entire ensemble of variated genes, leaving no genetic stones unturned, differs from the traditional pipeline that frequently focuses on recognizing a particular gene as a major descriptor of a pathological state.

The importance of not prematurely dismissing genetic alterations with ambiguous implications has been highlighted by considering variants of uncertain significance (VUS) alongside pathogenic and likely pathogenic variants.

This strategy accomplished two main objectives: improving the initial diagnosis and identifying novel characteristics strongly associated with modified genes. Consequently, patients diagnosed with rare genetic diseases will likely take advantage of more effective clinical pathway placements and consequent treatment interventions.

2. Methods

2.1. Patient selection for analysis

All patients provided informed consent, and in the case of pediatric patients, consent was obtained from both parents. Specific attention was given to patients suffering from autoinflammation and immunological dysregulation. These conditions, characterized by significant phenotypic variability, present challenges in clinical interpretation. It is noteworthy that these patients do not have specific entries in OMIM or defined HPO terms, and the use of generic terms could potentially introduce bias into the results. Furthermore, investigations were conducted into two less complex conditions: Hemolytic Uremic Syndrome (HUS), a rare condition characterized by the destruction of red blood cells, resulting in acute kidney failure that primarily affects children and can lead to severe complications if not promptly treated; Waldenström Macroglobulinemia (WM), an uncommon type of non-Hodgkin lymphoma characterized by abnormal white blood cells monoclonal immunoglobulin M (IgM), leading to blood thickening and various systemic symptoms. In these conditions, no results were obtained from the virtual analysis panels. The cohort, in its entirety, consisted of twenty probands diagnosed with immunological dysregulation and autoinflammation, nine diagnosed with HUS, and three with a potential diagnosis of WM.

2.2. Data analysis

Clinical SOPHiA GENETICS' Clinical Exome Solution (CES) sequencing was performed on a NextSeq 550 (Illumina, San Diego, USA) following the manufacturer's instructions. To further investigate the processed clinical exome data, information from the SOPHiA-DDM-v4 platform was rigorously reanalyzed using two separate web tools: Enrichr (https://maayanlab.cloud/Enrichr/) and DisGeNET (https://www.disgenet.org/home/). This dual-tool technique was used for enrichment analysis, which drastically improved our understanding of processed clinical exome data.

The major emphasis of our enrichment process was on a meticulously selected set of genes obtained from *. VCF file created using the SOPHiA-DDM-v4. This collection included all variants categorized as variants of unknown significance (VUS), pathogenic, or likely pathogenic. To improve the precision of our analytical approaches, we filtered the identified genes based on inheritance mode, existence of compound heterozygotes, and a 0.05 % minor allele frequency (MAF) cutoffs.

Enrichr is a powerful and easy-to-use tool for analyzing gene set enrichment. This includes several elements that enhance the analysis and visualization of the enrichment results. It can be easily integrated into existing workflows and is accessible via online browsers, making gene-set enrichment analysis a comprehensive and accurate resource. The software supports the use of several ranking systems for enriched terms, allowing for varied results, prioritization, and analyses. It also has interactive visualization capabilities, most likely powered by JavaScript data-driven documents (D3) [7–9].

DisGeNET is a data-collection tool that can identify links between human genetics and diseases. It is a popular tool that integrates a wide range of research data, prioritizes measures, and provides standard annotations for determining the molecular basis of human diseases and potential drug discovery. Moreover, these data are enhanced by the application of NLP-based text-mining tools, which combine genotype-phenotype interaction data from many databases and include the entire range of human diseases, including Mendelian and unusual traits linked with disease. These statistics were improved using NLP-based text-mining methods for data extraction from the literature. Metrics and annotations were provided to enhance the prioritization process and facilitate data retrieval and analysis [10].

3. Results

In total, 32 patients who visited our Operative Unit between January 2022 and February 2023 were thoroughly examined. These probands either lacked the pathogenic variants required for diagnostic investigation, or their alterations could only partially account for the observed clinical symptoms. The original investigation, which used the SOPHiA-DDM-v4 analytical platform, was unable to provide a satisfactory explanation of the clinical concerns. The objective of this study was to conduct a thorough investigation of probands using databases and integrated text-mining techniques to uncover new information that would improve our understanding of the clinical signs and symptoms of the disease, beginning with the original * . VCF file dataset obtained from the SOPHiA-DDM-v4.

These files provided the basis for subsequent analyses. A subgroup of genes with autosomal dominant (AD) inheritance, containing pathogenic or potentially pathogenic variants, was selected through a gene enrichment analysis. Additionally, all variants of unknown significance (VUS), filtered by compound heterozygotes, AD inheritance mode, and a minor allele frequency (MAF) limit of 0.05 %, were included. The initial stage of filtration included consulting the Human Phenotype Ontology database (HPO, https://hpo.jax.org/app/) to identify terms or conditions using phenotype codes. In parallel, diseases cataloged in the Online Mendelian Inheritance in Man (OMIM, https://www.omim.org/) database were considered. This follows a conventional clinical genetics procedure that uses preexisting datasets to filter and identify potential genes associated with specific phenotypes and disorders [11,12].

The secondary strategy focused on the identification of individual-specific genes through exome sequencing. These identified genes served as the input for the Enrichr and DisGeNET databases. By integrating the results derived from these databases, associations with diseases and clinical traits were not only identified, but novel and previously unexplored correlations between genes and diseases were also unveiled. This methodology thereby substantiates the efficacy of the adopted approach in revealing intricate genetic-disease relationships. Table 1 gives a thorough overview of the discrepancies between the text mining results and the analysis of the HPO and OMIM terms. The first

Table 1
Comparative Analysis of SOPHiA-DDM-v4 Filtering by Text Mining with Enrichr and DisGeNET VS OMIM and HPO Data.

ID	OMIN and HPO results	Disgenet and Enrich results	Protein	VAF%	gnomAD
Ces_01	CUBN: c.8593G>A LPL c.953A>G	CUBN: c.8593G>A,	p.Val2865Met	48%	0.25
		LPL c.953A>G,	p.Asn318Ser	56.9%	0.25
		CD74: c.364G>C	p.Ala122Pro	60%	0.25
Ces_02	RAG1: c.416G>T	RAG1: c.416G>T,	p.Gly139Val,	44.8%	0.0
		CNTNAP2: c.3328G>A,	p.Gly1110Arg,	47.3%	N/A
		DICER1: c.1124C>G,	p.Pro375Arg	47%	4.0E-4
		MEFV: c.2177T>C,	p.Val726Ala	45.7%	0.001
Ces_03	TCF3: c.607G>A	TCF3: c.607G>A,	p.Ala203Thr,	45.3%	1.0E-4
		GLI3: c.3239T>G,	p.Met1080Arg,	46.3%	N/A
		TGFBR2: c.931T>C,	p.Ser311Pro	48.5%	N/A
Ces_04	CASP10: c.703_704dup TNFRSF13B: c.260T>A	CASP10: c.703_704dup,	p.Leu235Phefs*2,	46.1%	0.0
		TNFRSF13B: c.260T>A,	p.Ile87Asn,	52%	5.0E-4
		DCC: c.3059T>C,	p.Phe1020Ser,	46.1%	N/A
		FEZF2: c.697C>A,	p.Pro233Thr,	52.3%	0.0
		POMC: c.157G>A	p.Asp53Asn	50.4%	0.0
Ces_05	WT	ATN1: c.530G>A,	p.Arg177Gln,	53%	1.0E-4
		CABIN1:c.5024G>A,	p.Gly1675Glu,	36.5%	N/A
		IRF8:c.11G>A,	p.Arg4Gln,	29.6%	0.0
		KLF1:c.887T>C,	p.Leu296Pro,	52.2%	N/A
		SIRT1:c.25C>T,	p.Leu9Phe	38.7%	N/A
Ces_06	CD40: c.647-1G>A IL6ST: c.736A>G	CD40: c.647- 1G>A,	p.?,	20.4%	N/A
		IL6ST: c.736A>G,	p.Asn216Asp,	44.3%	0.0
		CCR3: c.92C>T,	p.Thr31Ile,	57.6%	0.0
		COL4A3: c.4678C>A,	p.Val1560Ile,	43.2%	1.0E-4
		CREBBP: c.3109A>G,	p.Thr1037Ala,	40%	0.0
		IL17F: c.371A>C,	p.Gln124Pro,	54.9%	N/A
		MYH9: c.32A>G,	**p.Tyr11Cys,**	**48.8%**	**1.0E-4**
		MYH9: c.1102A>G,	**p.Asn368Asp**	**48.5%**	**0.0**
Ces_07	WT	SLC4A7:c.317C>T,	p.Gln373*	39.7%	1.0E-4
Ces_08	WT	GC: c.553A>G,	p.Met185Val,	43.4%	1.0E-4
		ICAM1:c.1432C>T,	p.Arg478Trp,	56.1%	0.0041
		PDGFRA:c.1516C>T,	p.Leu506Phe,	100%	0.0
		RORA: c.242_244del	p.Phe81del	49.2%	0.0
Ces_09	TGc.1567T>C	TG: c.1567T>C,	p.Ser523Pro,	56.5%	0.0018
		CACNA1:c.6019C>T,	p.Pro2007Ser,	54.7%	0.0
		CNTNAP2: c.2611G>T,	**p.Val871Leu,**	**42.5%**	**0.0**
		CNTNAP2:c.3328G>A,	**p.Gly1110Arg,**	**48.2%**	**N/A**
		KITLG: c.53T>C,	p.Leu18Pro,	43.3%	N/A
		KMT2D:c.1168G>A,	p.Val390Ile,	51.3%	0.0
		SELP:c.2180G>A,	p.Gly727Glu,	47.4%	0.0012
		TLX1:c.412A>G,	p.Arg138Gly,	50.9%	N/A
		TTN:c.6127A>G	p.Lys2043Glu	51.6%	N/A
Ces_10	**TCF3: c.920A>G TCF3:c.931G>G,**	**TCF3: c.920A>G,**	**p.His307Arg,**	**46.4%**	**0.0**
		TCF3:c.931G>G,	**p.Val311Leu,**	**53.1%**	**0.0**
		ADAM10: c.409C>A,	p.Val137Ile,	51.5%	0.0
		AXL: c.299G>T,	p.Arg100Leu,	50%	N/A
		THBD: c.1502C>T,	p.Pro501Leu,	40.3%	0.0018
		TIRAP: c.427C>T	p.Arg143Trp	57%	1.0E-4
Ces_11	SIAE: c.467A>G	SIAE: c.467A>G,	p.Tyr166Cys,	50.6%	N/A
	UNC13D:c.73A G	**UNC13D:c.73A>G,**	**p.Arg25Gly,**	**54.5%**	**N/A**
	UNC13D:c.-130C>T	**UNC13D:c.-130C>T,**	**p.?,**	**50.0%**	**N/A**
	UNC13D:c.2341G>A,	**UNC13D: c.2341G>A,**	**p.Val781Ile**	**52.4%**	**0.0013**
Ces_12	NLRP3: c.2113C>A SIAE:c.835C>T,	NLRP3: c.2113C>A,	p.Gln705Lys,	47.7%	0.04
		SIAE:c.835C>T,	p.Arg279Cys,	46.9%	0
		BCOR: c.3331C>T,	p.Pro1111Ser,	45%	0
		CCR5: c.187A>T,	p.Ser63Cys,	55.60%	0.0008
		EXT1:c.1814G>A,	p.Arg605Gln,	45.50%	N/A
		MMP1: c.152+1G>A,	**p.?,**	**42.30%**	**0.0001**
		MMP1: c.375del,	**p.Ile125Metfs*45,**	**43.90%**	**0.012**
		MYH9: c.3630- 4C>T,	p.?,	43.60%	0
		NLRP2:c.1765G>T,	p.Asp589Tyr,	44%	0.0004
		TRPM2: c.3122T>G,	p.Leu1041Arg,	50%	N/A
Ces_13	TCN2: c.508C>T UNC13D: c.5C>T	TCN2: c.508C>T,	p.Arg170Trp,	43.80%	0
		UNC13D: c.5C>T,	p.Ala2Val,	49.10%	0
		LRP5: c.1738G>A,	p.Val580Ile,	46.30%	0.0005
		MAPK1: c.928C>A,	p.His310Asn,	32%	N/A
		NOD2: c.649G>A,	p.Glu217Lys	52.50%	N/A
Ces_14	**WT**	APOE: c.21_26del,	p.Arg7_Lys8del	53%	N/A
		C3: c.4740G>C,	p.Gln1580His	48.60%	N/A
		CCR3: c.32A>G,	p.His121Arg	48.70%	N/A
		ENG: c.1099G>A,	p.Ala367Thr	43.50%	0
		FOXO1: c.- 2C>T,	p.?	43.20%	N/A
		IFIH1: c.505A>T,	p.Lys169*	42.90%	N/A
		NFATC2: c.2284C>T,	p.Arg762Cys	50.30%	0

(continued on next page)

Table 1 (*continued*)

ID	OMIN and HPO results	Disgenet and Enrich results	Protein	VAF%	gnomAD
		SLC22A1: c.188G>T,	p.Gly63Val	42.50%	N/A
Ces_15	**ATM:c.6860G>A**	**ATM:c.1236-2A>T,**	**p.Gly2287Glu**	**44.70%**	**0**
	ATM:c.1236-2A>T	**ATM:c.6860G>A,**	**p.?**	**38.90%**	**0**
	ITCH: c.1117C>T	ITCH: c.1117C>T,	p.Arg373Cys	55%	N/A
	KRAS: c.407G>A,	KRAS: c.407G>A,	p.Ser136Asn	55.90%	0
		GABBR1: c.97C>T,	p.Pro33Ser	41.90%	N/A
		HAVCR1: c.329G>A,	p.Arg110His	46.70%	0.0001
		PON3: c.931T>A,	p.Ser311Thr	43.90%	0
Ces_16	JAK3c.362G>A	JAK3: c.362G>A,	p.Arg121His	43.50%	0.0002
		AXIN1: c.644C>T,	p.Ser215Leu	45.50%	0.0004
		CYP21A2: c.518T>A,	**p.Ile173Asn**	**26.90%**	**N/A**
		CYP21A2: c.844G>T,	**p.Val282Leu**	**24.10%**	**N/A**
		CYP4F3: c.753del,	p.Asp252Metfs*53	47.90%	N/A
		MAP2K2: c.238G>A	p.Ala80Thr	47%	0
Ces_17	**WT**	ADAM10: c.1931G>A,	p.Arg644Gln	44.60%	0
		CBLB: c.227A>G,	p.Lys76Arg	46.20%	N/A
		CIC: c.112G>A,	p.Asp38Asn	44.10%	N/A
		EGFR: c.3200A>G,	p.Asn1067Ser	54.30%	0
		EOMES: c.491C>T,	p.Pro164Leu	56.60%	N/A
		P2RX7: c.1456C>A,,	p.Gln486Lys	50.40%	N/A
		PKM: c.1178G>T,	p.Arg393Leu	42.30%	N/A
		SH3BP2: c.688G>T,	p.Gly230Cys	48.30%	0
		SLC11A1: c.335G>A,	p.Arg112His	42.10%	0
		TPM3: c.*87_*104dup	p.?	42.30%	N/A
Ces_18	NLRP1: c.-155T>G NLRP12: c.779C>T SLC46A1: c.946C>G ZFAT: c.3385G>A	NLRP1: c.- 155T>G,	p.?	57.10%	N/A
		NLRP12: c.779C>T,	p.Thr260Met	41.40%	0.0007
		SLC46A1: c.946C>G,	p.Leu316Val	49.30%	N/A
		ZFAT: c.3385G>A,	p.Glu1129Lys	43.80%	0
		A4GALT: c.973C>T,	p.Arg325Trp	52%	0
		ABI3BP: c.1522+2T>A,	p.?	56.20%	0
		CILP: c.1547G>A,	p.Arg516His	50.70%	0.0001
		LPA: c.4282C>T,	**p.Arg1771Cys**	**47.50%**	**0.0019**
		LPA: c.5311C>T,	**p.Pro1428Ser**	**46.30%**	**N/A**
		LRRK2: c.7532T>C,	p(Ile2511Thr	44.10%	N/A
		NLRX1: c.830G>A,	p(Arg277His	52.50%	0.0002
		TNN: c.70294G>C,	**p.Arg23535His**	**48.40%**	**0.0001**
		TNN: c.70604G>A,	**p.Glu23432Gln**	**47.40%**	**N/A**
Ces_19	**WT**	PSTPIP1: c.*21C>T,	p.?	39.6%	3.0E-4
		REN: c.145C>T	p.Arg49*	35.7%	1.0E-4
Ces_20	UNC13D: c.1609G>T	UNC13D: c.1609G>T,	p.Val537Leu	39.90%	0
		ALDH2: c.913G>A,	p.Glu305Lys	67%	N/A
		GPC3: c.301A>C,	p.Lys101Gln	43%	0
		HDAC9:1621T>C,	p.Trp541Arg	50%	N/A
		ITGAM: c.1237C>	p.Arg413Trp	49.60%	0
		T, SPP1: c.98T>C,	p.Leu33Pro	39.70%	N/A
		TTN: c.47925G>T,	**p.Leu29065Arg**	**49.30%**	**0**
		TTN: c.87194T>G,	**p.Trp15975Cys**	**46.30%**	**0**
Ces_21	**WT**	NUMA1 c.5636G>C,	p.Ser1879Thr	50.4%	N/A
		SPP1: c.680A>G	p.Lys227Arg	54.6%	4.0E-4
Ces_22	CD46 c.1148C>T	CD46 c.1148C>T,	p.Thr383Ile	46.5%	6.0E-4
		MYH9: c.2180A>G	p.Asn727Ser	48.2%	N/A
Ces_23	**WT**	KMT2D: c.*48T>A,	p.?	26.7%	N/A
		MAP2K1: c.1138G>A,	p.Gly380Ser	45.1%	0.0
		NUMA1: c.6295-148_*124del	p.?	23.5%	N/A
Ces_24	**WT**	COG4: c.340A>G,	p.Ser114Gly	47.20%	0
		FBN1: c.3509G>A,	p.Arg1170His	36%	0.0012
		RYR1 c.3301G>A,	p.Val1101Met	46.70%	0
		SALL4 c.1287T>G,	p.Phe429Leu	37.90%	0.001
Ces_25	**WT**	F8: c.1318A>G,	p.Arg440Gly,	46.2%	N/A
		PML: c.1754C>T,	p.Ala585Val	49.7%	6.0E-4
Ces_26	**WT**	ALG8 c. 1516G>A,	p.Ala506Thr	48.20%	0.0005
		COG4 c.2039A>G,	p.Asn680Ser	57%	0
		FLT1 c.2174C>A,	p.Ser725*	46.90%	0
		ITGA2B c.457G>A,	p.Ala153Thr	47.20%	0.0002
		RUNX1 c.109G>A,	p.Gln1026Arg	39.60%	0
		NOS2 c.3077A>G	p.Val1326Met	50.80%	0
		PTPRJ: c.3976G>A	p.Gly37Ser	52.10%	N/A
Ces_27	**WT**	NLRP1:c.3589C>A	p.Leu1197Ile	51.1%	0.0
Ces_28	**WT**	ANKRD11: c.5614G>A,	p.Val1872Ile,	46%	N/A
		PTPRC: c.2822C>T,	p.Pro941Leu,	37.5%	N/A
		VWF: c.6112A>G	p.Met2038Val	30%	0.0
Ces_29	SLC7A7:c.1381_1384dup	SLC7A7:c.1381_1384dup,	p.Arg462Asnfs*7	46.80%	0
		SLC22A1 c.784C>T,	p.Arg262Cys	52.40%	0
		THPO	p.Ala209_*354del	33.10%	N/A
		TUBB1 c.925C>T,	**p.Gln43Pro**	**44.60%**	**N/A**

(*continued on next page*)

Table 1 (*continued*)

ID	OMIN and HPO results	Disgenet and Enrich results	Protein	VAF%	gnomAD
		TUBB1 c.c.128_129delinsCC	***p.Arg309Cys***	***45.40%***	***0***
Ces_30	***WT***	*NOD2 c.1965G>T*	*p.Leu655Phe*	*44.4%*	*5.0E-4*
Ces_31	***WT***	*LRRC8A c.1634G>A*	*p. Arg545His*	*56.6%*	*0.0*
Ces_32	***WT***	*ID3 c.*4G>A,*	*UTR3*	*35.3%*	*2.0E-4*
		VWF c.1653C>A	*'p.Asn551Lys*	*50.8%*	*N/A*

group of 20 probands in the table came with a diagnosis of autoinflammation and immunological dysregulation, while the other nine had suspicions of hemolytic uremic anemia (HUS), but no pathogenic variants were discovered by the virtual panel comprising the primary known genes connected with the disease (*ADAMTS13, C3, CD46, CFB, CFH, CFHR1, CFHR3, CFHR, CFI, COG1, DGKE, MMACHC, MTRR, PRDX1, THBD, TMEM165, ZNFX1*). The remaining three patients may have WM, although no identified pathogenic variants were found in *MYD88*.

Text mining analysis provides valuable insights into the variability of the affected organs and their clinical characteristics, offering important information on the complex landscape of autoimmune and autoinflammatory conditions. Numerous interactions between immune-related disorders and their mechanisms have been revealed by a substantial amount of knowledge on these disorders, highlighting their complexity. These conditions involve distinct modifications related to innate and adaptive immunity, which involve a series of events that lead from an initial inflammatory state to the damage of specific organs. The analysis revealed several genes with distinct relationships, as well as possible etiological and phenotypic alterations, providing new knowledge into the mechanisms that control the beginning and progression of these conditions, which are still mostly unknown.

In the CES_05 study, where traditional filtering with OMIM and HPO terms yielded no results, we identified five genes associated with the phenotype of immunological dysregulation: *ATN1*, *CABIN1, IRF8, KLF1*, and *SIRT1*. Specifically, the protein encoded by *ATN1* (atrophin1) interacts with CABIN1 (Calcineurin Binding Protein 1) via HDAC7 (Histone Deacetylase 7), and with SIRT1 (Sirtuin1), a post-translational regulator that modulates inflammation. CABIN1 is known to play a role in T cell activation as a negative regulator of calcineurin [13].

Similarly, KLF1 (KLF Transcription Factor 1) interacts with IRF8 (Interferon Regulatory Factor 8) through the protein BCL11A (Transcription factor B-cell lymphoma/leukemia 11A) [14,15]. IRF8 is essential for the development and maturation of myeloid cells (dendritic cells, monocytes, macrophages), and for the expression of intrinsic anti-microbial functions such as antigen capture, processing, and presentation to lymphoid cells, and for the activation of these cells in response to cytokines and pro-inflammatory stimuli. Further examination of the potential correlations between the identified proteins reveals a complex network of regulatory mechanisms involving gene expression modulation, T cell activation, and inflammatory response regulation. This highlights the complexity of immune dysregulation and the potential for targeted therapeutic interventions [16].

The analysis was also extended to patients initially diagnosed with HUS but without a pathogenic variation identified using the virtual gene panel. Text mining revealed a wider array of genetic variants in this patient sample, suggesting a probable lack of precision in the diagnosis and genetic heterogeneity within the condition.

Interestingly, the diagnosis was refined for the Ces_27 sample, confirming the presence of the *NLRP1* variant c .3589C>A. *NLRP1* has been associated with an increased risk of Addison's disease, a rare chronic condition caused by adrenal gland failure, which can lead to renal microangiopathy and, eventually, renal failure [17].

In the third group of three patients with WM, different genetic variants, including *NOD2, LRRC8A,* and *ID3* with *VWF*, were detected. Primary analysis using conventional OMIM and HPO keywords did not identify related genes. However, the genes discovered are related to Schnitzler's syndrome (SchS), an autoinflammatory disease, congenital agammaglobulinemia, and primary Sjögren's syndrome (pSS). These findings enable the reassessment of patients and the potential use of these data for differential studies to enhance classification [18–21].

Remarkably, the text-mining approach revealed a much larger pool of gene matches, successfully concluding the diagnosis in all cases, compared to only 16 cases partially diagnosed using the conventional OMIM and HPO approaches. Additionally, the text-mining method ensures that no information is lost from the conventional OMIM and HPO filtering, as it includes the resulting genes in the larger list. This demonstrates the power and potential of text mining in genetic research and diagnosis.

4. Discussion

The accelerated evolution of NGS technologies, complemented by the continuous refinement of software tools and data analysis pipelines, has revolutionized our ability to explore the genetic foundations of diseases. This extends our understanding of the genetic contributions to various medical conditions [22].

However, the NGS revolution is characterized by variable diagnosis rates that do not have defined values but rather span a wide range (25–50 %); thus, they are influenced by a variety of factors that affect the accuracy and efficacy of this technology [4,5].

Diagnostic success becomes more complicated when dealing with diseases characterized by extensive genetic and allelic diversity. In such scenarios, multiple genes may contribute to the condition, each with varying levels of influence, requiring additional analytical depth and precision.

The application of basic filters within the currently used tools, which include the utilization of specific HPO terms or diseases documented in the OMIM database, plays a pivotal role in curating genetic information. These filters, while seemingly straightforward, possess the remarkable ability to significantly restrict the vast wealth of genetic data accessible from an entire range of genes [23,24].

HPO terms offer a structured vocabulary to describe the phenotypic abnormalities associated with a particular genetic condition. OMIM serves as a comprehensive resource cataloging a myriad of inherited genetic disorders and the genes implicated in their etiology. By harnessing the power of these two resources, the software can narrow the search for pertinent genetic information [24].

One of the primary functions of these filters is to identify genes that are directly relevant to a specific disease or set of clinical characteristics. This targeted approach is instrumental in the diagnosis and management of genetic conditions. This allows healthcare professionals and researchers to efficiently shift through genomic data and focus their attention on genes most likely to be associated with the observed clinical manifestations [25] this precision comes at a cost and a significant portion of the genomic data is excluded. When these filters were applied, several genes that did not directly align with the specified HPO terms or OMIM-listed diseases were screened. Although this is advantageous in terms of efficiency and focus, potentially valuable genomic information residing outside the predefined scope remains unexplored [25]. This highlights the existing dichotomy between the "analog" data provided by clinical characteristics, which often overlap or are attributable to completely different pathologies, and the "digital" data provided by variants that define genes implicated in syndromic forms, even those

different from the initially suspected ones, as seen in reverse genomics.

This may have resulted in the oversight of genes with previously undiscovered associations with the conditions under investigation. These "hidden" genes may provide crucial clues regarding disease mechanisms, potential treatment targets, or genetic modifiers that could enhance our understanding and management of this condition. Stringent adherence to these filters can inadvertently limit the ability to reveal novel genetic insights.

However, the application of HPO and OMIM filters is not a one-size-fits-all approach. Its utility depends on the specific objectives of the genetic analysis. In situations where a clinician is seeking a rapid diagnosis of a patient with well-defined clinical features, these filters are invaluable. They expedite the diagnostic process by identifying genes that are most likely to be causative. However, for researchers exploring the intricacies of complex genetic disorders or less-characterized conditions, a broader, less-filtered approach may be necessary to cast a wider net and explore the full genomic spectrum.

This multigenic approach emphasizes the significance of a thorough genetic perspective that illustrates the complex network of hereditary factors influencing health and disease. Understanding the complexity of many diseases, the role played by different genes, and how a person's unique genetic profile is defined.

It may also be combined with information offered by the statistical method of Polygenic Risk Scores (PRS), which predict a particular phenotype by integrating the cumulative impact of several genetic variants that only marginally increase the overall risk.

The potential integration of the multigenic approach with PRS is fascinating because it will help us comprehend the genetic landscape of complicated diseases. The potential benefit of combining multigenic analysis with PRS collective genetic risk could provide a more complete picture of genetic susceptibility, especially in the absence of a diagnosis. This is evident in diseases where multigenic research reveals a variety of genetic variants but is unable to identify the key contributors.

In these scenarios, the ability to distinguish between hereditary and non-genetic disorders with significant environmental components is extremely useful for disease diagnosis. This technique advances precision medicine by adding PRS and multigenic analysis to the diagnostic process, allowing medical professionals to provide patients and their families with personalized treatment.

5. Conclusion

In conclusion, our retrospective study highlights the limitations of traditional methods for obtaining information on uncommon genetic diseases. The need to improve diagnostic accuracy also becomes clear in the case of autosomal recessive and autosomal dominant single-gene conditions.

NGS has ushered in a new age by enabling the use of multigenic approaches and the availability of a plethora of genomic data. This comprehensive multigenic approach allows for the customization of treatment plans according to the individual genetic profile of each patient. Conditions such as PKU and thalassemia are excellent examples of the complex features of autosomal recessive and autosomal dominant illnesses, which highlight the complex interactions between genes and disease symptoms. When a multigenic approach is applied, genetic testing is used to detect variations accurately, offering important information regarding the likelihood of the disease [26,27].

Furthermore, the shift in our understanding from monogenic to multigenic disorders reflects the changing complexity of genetic contributions to health and disease. As we move away from the period of "orphan" illnesses, which are frequently overlooked owing to their rarity, towards a more comprehensive understanding of multigenic factors, the potential for therapeutic advances becomes clearer [28]. The transition to multigenic illnesses, which includes the investigation of modifier genes, reveals the dynamic characteristics of genetic research and its implications for precision medicine [28].

To summarize, the trajectory of genetic research, together with advances in technology such as NGS, establishes multigenic analysis as the foundation of precision medicine. This radical change promises to improve diagnostic accuracy while also opening the door for personalized and targeted treatment approaches, eventually improving patient outcomes and influencing medical care approaches.

The table presents a detailed comparison of two distinct methodologies: the application of SOPHiA-DDM-v4 filters for OMIM and HPO analysis, and a text mining strategy involving the Enrichr and DisGeNET databases.

The first 20 samples (Ces_01 to Ces_20) are from patients with autoinflammation and immunological dysregulation. The second column lists genes identified using specific DDM platform filters, including OMIM Inflammation, Autoinflammatory disease, and HPO terms (HP:0000285, HP:0004313, HP:0010976, HP:0001433, HP:0004315, HP:0001875). The third column displays genes identified through DisGeNET and Enrichr analysis using terms such as Autoimmune Autoinflammatory Disease, Neutropenia Lymphocytopenia, and Primary Immunodeficiencies.

Samples Ces_21 to Ces_29 are related to suspected Hemolytic Uremic Syndrome (HUS) cases. These were analyzed using the OMIM term: HUS and HPO terms such as HP:0001919, HP:0001937, HP:0004431, HP:0005575. The resulting genes are displayed in the second column. The third column shows genes identified through DisGeNET and Enrichr analysis using terms such as Hemolytic and Uremic Syndrome, Thrombocytopenia, Kidney Injury, and Complement Deficiency.

Samples Ces_30 to Ces_32 are suspected cases of Waldenström's Macroglobulinemia (WM). Genes identified using the OMIM term: Waldenström's Macroglobulinemia and HPO term: HP:0005508 are displayed in the second column. The third column shows genes identified through DisGeNET and Enrichr analysis using terms such as Waldenström's Macroglobulinemia, Hypogammaglobulinemia, Macroglobulinemia, and Cryoglobulinemia.

It's noteworthy that all genes identified through the OMIM and HPO filters were consistently present in the results of the text mining analysis. Genes and variants highlighted in red are of particular interest. Compound heterozygotes are shown in bold font. The final three columns report the corresponding amino acid changes, the variant allele frequency (VAF), and the frequency in the gnomAD (Genome Aggregation Database). (Fig. 1).

Ethics approval and consent to participate

Written informed consent was obtained from patients and their parents to perform genetic testing and use the data for scientific purposes.

Author agreement

All authors have read, revised, and approved the final version of the manuscript. The authors warrant that the article is the original work of the authors, has not been published previously, and is not under consideration for publication elsewhere.

Competing interests

The authors declare that they have no competing interests.

Informed consent statement

All procedures performed in studies involving human participants were in accordance with the ethical standards of the WMA Declaration of Helsinki - Ethical Principles for Medical Research Involving Human Subjects" (https://www.wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-human-subjects/ (accessed on June 5, 2023)). This article does not contain any

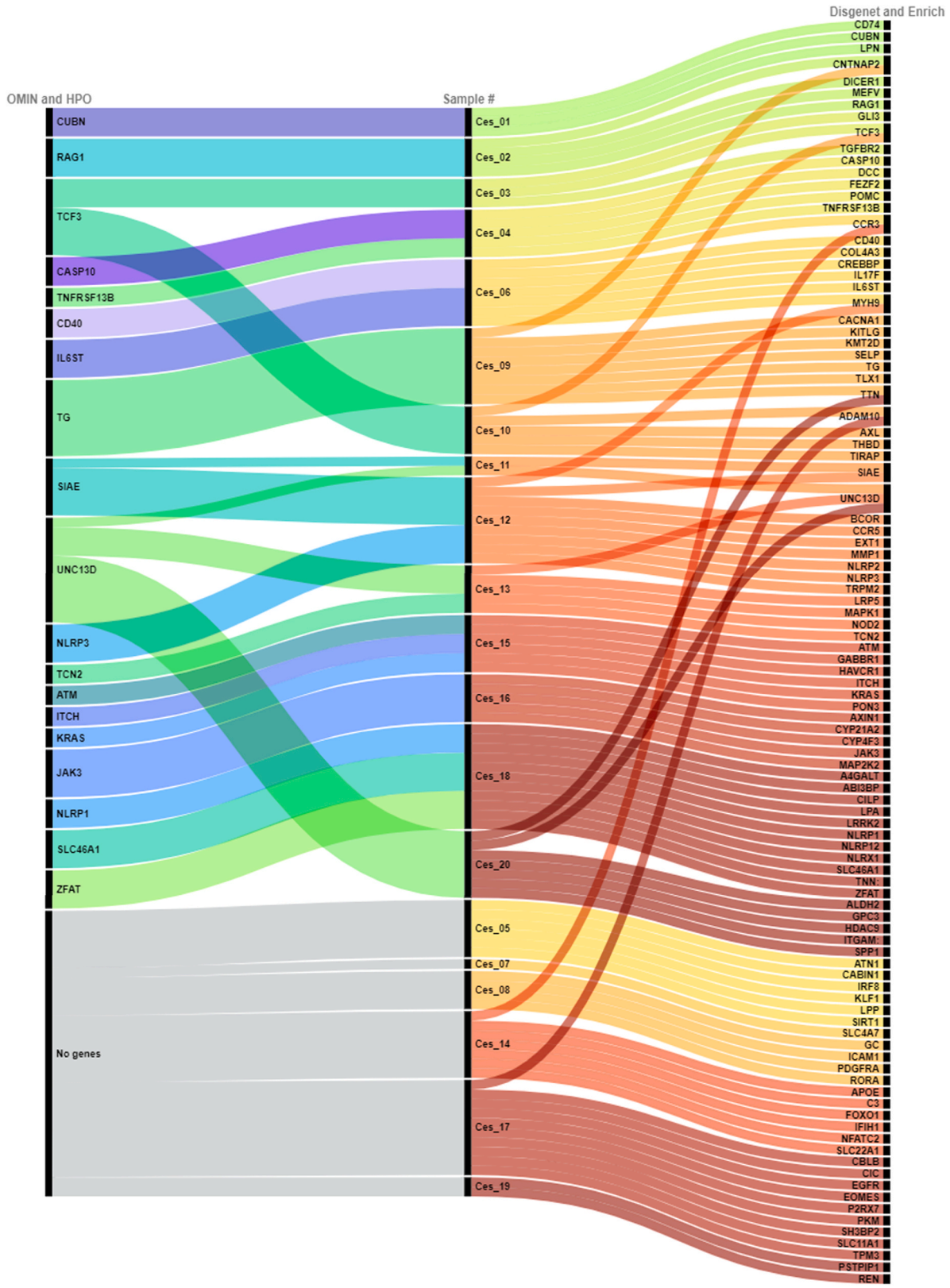

Fig. 1. In the diagram, a comprehensive comparison is presented between two approaches: one involving the application of SOPHiA-DDM-v4 filters in the analysis of OMIM and HPO approaches, and the other employing a text mining strategy utilizing Enrichr and DisGeneNET databases. The results reveal the identification of multiple causative genes for each patient, underscoring the significant heterogeneity of the autoinflammatory syndrome condition. Additionally, potential causative genes were identified in samples Ces_05, Ces_07, Ces_08, Ces_14, Ces_17, and Ces_19 where a conventional filtering method using OMIM and HPO had failed to identify any causative gene.

studies with animals performed by any of the authors. Written informed consent for genetic examination and publication was obtained from all patients or their legal representatives.

Declaration of Competing Interest

We have no financial interests or connections, direct or indirect, that might raise questions about the objectivity or impartiality of our work. We have no personal relationships with individuals or organizations that could influence our work inappropriately. We have not been involved in any activities or relationships that could be perceived as a conflict of interest. We all recognize that in order to maintain the openness and legitimacy of scientific research, it is crucial to disclose any possible conflicts of interest. We affirm that the information provided in this declaration is accurate and complete to the best of our knowledge.

Data availability

We can provide this information or conduct a reanalysis upon request. Raw data cannot be posted because of patient confidentiality issues.

References

[1] van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C. Ten years of next-generation sequencing technology. Trends Genet 2014;30(9):418–26. https://doi.org/10.1016/j.tig.2014.07.001. Epub 2014 Aug 6. PMID: 25108476.

[2] Pereira R, Oliveira J, Sousa M. Bioinformatics and computational tools for next-generation sequencing analysis in clinical genetics. J Clin Med 2020;9(1):132. https://doi.org/10.3390/jcm9010132. PMID: 31947757; PMCID: PMC7019349.

[3] Marchant G, Barnes M, Evans JP, LeRoy B, Wolf SM. From genetics to genomics: facing the liability implications in clinical care. J Law, Med Ethics 2020;48(1):11–43.

[4] Zhong Y, Xu F, Wu J, Schubert J, Li MM. Application of next generation sequencing in laboratory medicine. Ann Lab Med 2021;41(1):25–43. https://doi.org/10.3343/alm.2021.41.1.25. Epub 2020 Aug 25. PMID: 32829577; PMCID: PMC7443516.

[5] Sullivan Jennifer A, Schoch Kelly, Spillmann Rebecca C, Shashi Vandana. Exome/genome sequencing in undiagnosed syndromes. Annu Rev Med 2023;74(1):489–502.

[6] Strianese O, Rizzo F, Ciccarelli M, Galasso G, D'Agostino Y, Salvati A, et al. Precision and personalized medicine: how genomic approach improves the management of cardiovascular and neurodegenerative disease. Genes 2020;11(7):747. https://doi.org/10.3390/genes11070747. PMID: 32640513; PMCID: PMC7397223.

[7] Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinforma 2013;14:128.

[8] Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res 2016:gkw377.

[9] Xie Z, Bailey A, Kuleshov MV, Clarke DJB, Evangelista JE, Jenkins SL, et al. Gene set knowledge discovery with Enrichr. Curr Protoc 2021;1:e90. https://doi.org/10.1002/cpz1.90.

[10] Piñero Janet, Ramírez-Anguita Juan Manuel, Saüch-Pitarch Josep, Ronzano Francesco, Centeno Emilio, Sanz Ferran, et al. The DisGeNET knowledge platform for disease genomics: 2019 update. Nucleic Acids Res 2020;48(D1):D845–55.

[11] Valencia CA, Mathur A, Denton J, Wei C, Wang X, Husami A, et al. CCEPAS: the creation and validation of a fast and sensitive clinical whole exome analysis pipeline based on gene and variant ranking. J Transl Genet Genom 2018;2:1. https://doi.org/10.20517/jtgg.2017.05.

[12] Köhler S, Øien NC, Buske OJ, Groza T, Jacobsen JOB, McNamara C, et al. Encoding clinical data with the human phenotype ontology for computational differential diagnostics. Curr Protoc Hum Genet 2019;103(1):e92. https://doi.org/10.1002/cphg.92. PMID: 31479590; PMCID: PMC6814016.

[13] Sun L, Youn HD, Loh C, Stolow M, He W, Liu JO. Cabin 1, a negative regulator for calcineurin signaling in T lymphocytes. Immunity 1998;8(6):703–11. https://doi.org/10.1016/s1074-7613(00)80575-0. PMID: 9655484.

[14] Zhou D, Liu K, Sun CW, Pawlik KM, Townes TM. KLF1 regulates BCL11A expression and gamma- to beta-globin gene switching. Nat Genet 2010;42(9):742–4. https://doi.org/10.1038/ng.637. Epub 2010 Aug 1. PMID: 20676097.

[15] Kurotaki D, Nakabayashi J, Nishiyama A, Sasaki H, Kawase W, Kaneko N, et al. Transcription factor IRF8 governs enhancer landscape dynamics in mononuclear phagocyte progenitors. Cell Rep 2018;22(10):2628–41. https://doi.org/10.1016/j.celrep.2018.02.048. PMID: 29514092.

[16] Moorman HR, Reategui Y, Poschel DB, Liu K. IRF8: mechanism of action and health implications. Cells 2022;11(17):2630. https://doi.org/10.3390/cells11172630. PMID: 36078039; PMCID: PMC9454819.

[17] Levandowski CB, Mailloux CM, Ferrara TM, Gowan K, Ben S, Jin Y, et al. NLRP1 haplotypes associated with vitiligo and autoimmunity increase interleukin-1β processing via the NLRP1 inflammasome. Proc Natl Acad Sci USA 2013;110(8):2952–6. https://doi.org/10.1073/pnas.1222808110. Epub 2013 Feb 4. PMID: 23382179; PMCID: PMC3581876.

[18] Navetta-Modrov B, Yao Q. Macroglobulinemia and autoinflammatory disease. Rheuma Immunol Res 2021 Dec 31;2(4):227–32. https://doi.org/10.2478/rir-2021-0031. PMID: 36467983; PMCID: PMC9524799.

[19] Sawada A, Takihara Y, Kim JY, Matsuda-Hashii Y, Tokimasa S, Fujisaki H, et al. A congenital mutation of the novel gene LRRC8 causes agammaglobulinemia in humans. J Clin Invest 2003;112(11):1707–13. https://doi.org/10.1172/JCI18937. PMID: 14660746; PMCID: PMC281644.

[20] Hayakawa I, Tedder TF, Zhuang Y. B-lymphocyte depletion ameliorates Sjögren's syndrome in Id3 knockout mice. Immunology 2007;122(1):73–9. https://doi.org/10.1111/j.1365-2567.2007.02614.x. Epub 2007 Apr 30. PMID: 17472721; PMCID: PMC2265983.

[21] Owari M, Harada-Shirado K, Togawa R, Fukatsu M, Sato Y, Fukuchi K, et al. Acquired von Willebrand syndrome in a patient with multiple comorbidities, including MALT lymphoma with IgA monoclonal gammopathy and hyperviscosity syndrome. Intern Med 2023;62(4):605–11. https://doi.org/10.2169/internalmedicine.9815-22. Epub 2022 Jul 22. PMID: 35871597; PMCID: PMC10017253.

[22] Satam H, Joshi K, Mangrolia U, Waghoo S, Zaidi G, Rawool S, et al. Next-generation sequencing technology: current trends and advancements. Biology 2023;12:997. https://doi.org/10.3390/biology12070997.

[23] Bone WP, Washington NL, Buske OJ, Adams DR, Davis J, Draper D, et al. Computational evaluation of exome sequence data using human and model organism phenotypes improves diagnostic efficiency. Genet Med 2016 Jun;18(6):608–17. https://doi.org/10.1038/gim.2015.137. Epub 2015 Nov 12. PMID: 26562225; PMCID: PMC4916229.

[24] Austin-Tse Chrissy, Jobanputra Vaidehi, Perry Denise, Bick David, Taft Ryan, Venner Eric, et al. Best practices for the interpretation and reporting of clinical genome sequencing (Supplement) Genet Med 2022;Volume 24(Issue 3):S365–6.

[25] Groza Tudor, Köhler Sebastian, Moldenhauer Dawid, Vasilevsky Nicole, Baynam Gareth, Zemojtel Tomasz, et al. The Human phenotype ontology: semantic unification of common and rare disease. Am J Hum Genet 2015;97(1):111–24.

[26] Genetic Alliance; District of Columbia Department of Health. Understanding Genetics: A District of Columbia Guide for Patients and Health Professionals. Washington (DC): Genetic Alliance; 2010 Feb 17. Appendix B, Classic Mendelian Genetics (Patterns of Inheritance) Available from: https://www.ncbi.nlm.nih.gov/books/NBK132145/.

[27] Antonarakis S, Beckmann J. Mendelian disorders deserve more attention. Nat Rev Genet 2006;7:277–82. https://doi.org/10.1038/nrg1826.

[28] Badano J, Katsanis N. Beyond Mendel: an evolving view of human genetic disease transmission. Nat Rev Genet 2002;3:779–89. https://doi.org/10.1038/nrg910.