Identifying the genetic causes of phenotypically diagnosed Pakistani mucopolysaccharidoses patients by whole genome sequencing

Gul, Rutaba; Firasat, Sabika; Schubert, Mikkel; Ullah, Asmat; Peña, Elionora; Thuesen, Anne C.B.; Hussain, Mulazim; Staeger, Frederik F.; Gjesing, Anette P.; Albrechtsen, Anders; Hansen, Torben

Published in:
Frontiers in Genetics

DOI:
10.3389/fgene.2023.1128850

Publication date:
2023

Document version
Publisher's PDF, also known as Version of record

Document license:
CC BY

Citation for published version (APA):
Identifying the genetic causes of phenotypically diagnosed Pakistani mucopolysaccharidoses patients by whole genome sequencing

Rutaba Gul1,2, Sabika Firasat1*, Mikkel Schubert2, Asmat Ullah2, Elionora Peña3, Anne C. B. Thuesen2, Mulazim Hussain3, Frederik F. Staeger4, Anette P. Gjesing2, Anders Albrechtsen4 and Torben Hansen2*

1Department of Zoology, Faculty of Biological Sciences, Quaid-I-Azam University, Islamabad, Pakistan, 2Novo Nordisk Foundation Center for Basic Metabolic Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark, 3The Children Hospital, Pakistan Institute of Medical Sciences (PIMS), Islamabad, Pakistan, 4Section for Computational and RNA Biology, Department of Biology, University of Copenhagen, Copenhagen, Denmark

Background: Lysosomal storage disorders (LSDs) are a group of inherited metabolic diseases, which encompass more than 50 different subtypes of pathologies. These disorders are caused by defects in lysosomal enzymes, transporters, and other non-lysosomal proteins. Mucopolysaccharidosis (MPS) is the most common subgroup of lysosomal storage disorders in which the body is unable to properly breakdown mucopolysaccharides. The aim of the present study was to identify novel genes and pathogenic variants in families from diverse regions of Pakistan with clinically diagnosed mucopolysaccharidosis type I and mucopolysaccharidosis type II.

Methods: Clinical diagnosis identified 12 with mucopolysaccharidosis I and 2 with mucopolysaccharidosis II in 14 families and whole genome sequencing (WGS) was performed to identify the causative variations in 15 affected individuals. Twenty-two unaffected individuals including parents or normal siblings of patients were also sequenced. Putative causal variants were identified by co-segregation and functional annotation.

Results: Analysis of whole genome sequencing data revealed ten novel and six previously reported variants in lysosomal storage disorders-associated genes (IDUA, GALNS, SGSH, GAA, IDS, ALDOB, TRAPPc4, MASPl, SMARCal, KIAA1109, HERC1, RRAS2) and a novel candidate gene (ABCA5) for lysosomal storage disorder-like phenotypes, which has previously been associated with symptoms strongly related with lysosomal storage disorder in animal models.

Conclusion: Multigenic inheritance was found in several families highlighting the importance of searching for homozygous pathogenic variants in several genes also in families with a high degree of consanguinity.

KEYWORDS
lysosomal storage disorder, mucopolysaccharidosis, Pakistani families, whole genome sequencing, ABCA5
1 Introduction

Lysosomal storage disorders (LSDs) are a diverse group of inherited conditions primarily caused by loss of function of lysosomal proteins leading to the dysregulation of many lysosomes-linked activities including autophagy, lipid homeostasis, membrane repair, cell viability, exocytosis and signaling cascades (Parenti et al., 2015; Marques and Saftig, 2019). Common clinical phenotypes of LSDs include delay in achieving milestones, abdominal distention, hepatosplenomegaly, short stature, coarse facial features, and joint contractures (Gul et al., 2021). LSDs are classified into subtypes based on accumulation of undegraded substrates like glycosaminoglycans (GAGs), glycogens and sphingolipids accumulated as a result of compromised activities of enzymes in lysosomes.

Autosomal recessive disorders are more prevalent in populations like the Pakistani where 60% of all marriages are consanguineous and among them more than 80% are first cousin unions (Hussain and Bittles, 1998). Globally, the prevalence for LSDs ranges from 7.5/100,000 to 23.5/100,000 live births, with sphingolipidoses as the highest prevalent disorder followed by mucopolysaccharidoses (MPSs) (Kingma et al., 2015). Mucopolysaccharidoses (MPS) (OMIM # 252700) are carbohydrate metabolism disorders caused by the deficiency of lysosomal enzymes required for hydrolyzing glycosaminoglycans (GAGs; negatively charged polysaccharides). MPS are characterized by gradual accumulation of various types of GAGs within cells of multiple organs leading to somatic manifestations such as coarse facial features, cardiac issues, skeletal deformities, hepatosplenomegaly, respiratory problems, hematologic, neurologic, and ocular phenotypes. MPS include a spectrum of clinical phenotypes from severe to attenuated forms including all MPS sub-types like MPS I, MPS II, MPS III, MPS IV, MPS VI, MPS VII with overlapping physical features (Parenti and Giugliani, 2022).

In the present study, we recruited 14 families in total (A–N). Twelve of them (A–L) were diagnosed as MPS I with 26 affected individuals (Figure 1) and the other two (M, N) as MPS II with two affected males (Figure 2). In total, 37 individuals including 15 affected and 22 unaffected participated in the present study. Families were recruited from diverse regions of Pakistan (i.e., Punjab, Islamabad capital territory, Khyber Pakhtunkhwa, and Azad Jammu and Kashmir) (Figure 3). Moreover, 77 additional individuals with Pakistani origin that are part of an ongoing LSDs study (unpublished) were also sequenced.

2 Materials and methodology

2.1 Ethical approvals

Ethical approval was granted by the Bioethical Review board, Quaid-i-Azam University, Islamabad, Pakistan (BEC-FBS-QUAU2019-198) and Shaheed Zulfiqar Ali Bhutto Medical University, Islamabad, Pakistan (F-1-1/2105/ERB/SZABMU/179).

Before collection of blood samples, all participants or their guardians signed an informed consent form. The study was conducted in accordance with the Declaration of Helsinki (World Medical Association, 2013).

2.2 Participants

In the present study, we recruited 14 families in total (A–N). Twelve of them (A–L) were diagnosed as MPS I with 26 affected individuals (Figure 1) and the other two (M, N) as MPS II with two affected males (Figure 2). In total, 37 individuals including 15 affected and 22 unaffected participated in the present study.

Peripheral blood samples of 3–5 ml were taken in an EDTA vacutainer (BD vacutainer K2 EDTA 18 mg) and stored as whole blood at −20°C before DNA extraction. DNA extraction was performed using standard Phenol-Chloroform method (Grimberg et al., 1989). DNA was quantified by using the μDrop Plate reader (MultiskanTM, Thermo Fisher Scientific, and Waltham, MA, United States).
2.4 Whole genome sequencing and variant calling

Read mapping and genotype calling was performed using a modified version of the PALEOMIX pipeline (Schubert et al., 2014). The modified pipeline and detailed instructions are available at https://github.com/Hansen-Group/Gul2022.

Briefly, pre-analyses quality assurance was performed using FastQC v0.11.9 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and MultiQC v1.10 (Ewels et al., 2016). Read pairs were subsequently processed using fastp v0.20.1 (Chen et al., 2018) to trim adapter sequences, remove reads with low complexity, and merge and correct overlapping read pairs (with a minimum overlap of 11 bp). Reads were mapped using BWA mem v0.7.17-r1188 (Li, 2013). Against the hg38 human reference genome distributed as part of the GATK resource bundle (https://gatk.broadinstitute.org/hc/en-us/articles/360035890811-Resource-bundle), (McKenna et al., 2010) including alternative and decoy contigs. Mapped reads were post-processed using samtools v1.11 (Danecek et al., 2021),
and the “bwa-postalt.js” script included in the BWA-kit which post-processes alignments mapping to alternative contigs such as HLA variants. Duplicate, unmapped, and orphaned reads, and secondary and supplementary alignments were filtered. The resulting BAMs were recalibrated using GATK ApplyBQSR trained using dbSNP release 151 (Sherry et al., 2001) for known variable sites.

Genotypes (GVCFs) were called for each sample using GATK HaplotypeCaller, merged, and called using GATK GenotypeGVCFs. Variant recalibration was carried out separately for SNPs and indels using GATK ApplyVQSRNode, trained using positive lists from the GATK resource bundle. Features used in the calibration were ExcessHet, DP, MQ, QD, SOR, FS, ReadPosRankSum, MQRankSum, and BaseQRankSum (only for SNPs). We kept only variables within the upper 98.0 tranche for indel and 99.6 tranche for SNPs. The resulting VCFs were annotated using VEP v104 (McLaren et al., 2016) using the Ancestral Allele, ExACpLI, GERP Conservation Scores, and LOFTEE v1.0.3 (https://github.com/konradjk/loftee) (Karczewski et al., 2020) plugins. Additional custom annotation was included as described in the github repository.

2.4.1 Data analysis and variant prioritization
All variants were evaluated as per ACMG guidelines (Richards et al., 2020). Variants that were classified to be pathogenic, likely pathogenic or uncertain significance according to ACMG guidelines were prioritized.

Runs of homozygosity (ROH)-based inbreeding coefficient (F) and principal component analysis was calculated using PLINK (v1.90b6) on a dataset including both affected and unaffected individuals as well as 77 additional individuals with Pakistani origin that are part of an ongoing LSDs study (unpublished) for better allele frequency estimates. Results are only shown for the individuals in present study.

For ROH calculation, only variants with minor allele frequency MAF > 5%, and no missingness were used. ROHs were defined as having more than 100 variants, a total length larger than 1 MB, on average at least 1 variant per 50 kB, consecutive variants less than 1 MB apart, and at most 1 heterozygous call per window of 50 variants. The total length of ROHs was then divided by the total length of the genome (3 GB).

Principal component analysis was performed by first merging with the 1000 genomes project (https://www.nature.com/articles/nature15393). Non-overlapping sites and sites with a MAF difference

![Map of Pakistan showing the locations (black triangles) of cities/towns from where families included in this study were collected.](https://example.com/map.png)
<table>
<thead>
<tr>
<th>Family ID</th>
<th>Individual no.</th>
<th>No. of affected individuals</th>
<th>Age/ Sex</th>
<th>Cast/Region/ Consanguinity</th>
<th>Clinical profile</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>V-4</td>
<td>1</td>
<td>9 m/F</td>
<td>Rajput/ICT/Yes</td>
<td>Squint eyes, tip-toe walking, kyphosis, frontal bossing, depressed nasal bridge, umbilical hernia, narrowing of spinal cord, dysostosis, throat problem, weak muscles, more crying episodes, two central incisors widely spaced, joint contractures, coarse facial features</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Pectus carinatum, Hemoglobin = 8.6 g/dL (12–15), HCT = 32 (35–55), MCV = 63.9 fl (75–100), MCH = 17.2 pg (25–35), MCHC = 26.9 g/dL (31–38), Neutrophils = 40.4% (60–70), ESR = 28 mm/Hr (0–15)</td>
</tr>
<tr>
<td>B</td>
<td>V-1</td>
<td>1</td>
<td>1 y/M</td>
<td>Malik/Punjab/Yes</td>
<td>Delay in achieving milestone, speech delay, hernia, short stature, coarse facial features, joint contractures, hepatosplenomegaly, developmental delay, depressed nasal bridge, squinted eyes</td>
</tr>
<tr>
<td>C</td>
<td>V-1</td>
<td>1</td>
<td>4 y/F</td>
<td>Pathan/ICT/Yes</td>
<td>Delay in achieving milestone, abdominal distention, aggressive behavior, speech delay, short stature, respiratory/ear infections, coarse facial features, broad based gait, joint contractures, short trunk, scoliosis/kyphosis, hepatosplenomegaly, low IQ, intellectual disability, low hemoglobin</td>
</tr>
<tr>
<td>D</td>
<td>IV-1</td>
<td>1</td>
<td>1.5 y/F</td>
<td>Sheikh/Punjab/Yes</td>
<td>Delay in achieving milestone, abdominal distention, aggressive behavior, speech delay, gibbus formation, short stature, upper airways obstruction, kyphosis, hazel cornes, coarse facial features, hearing loss, joint contractures, hepatosplenomegaly, Intellectual disability, hypertrichosis</td>
</tr>
<tr>
<td>E</td>
<td>VI-4</td>
<td>6</td>
<td>8 y/F</td>
<td>Rajput/AJK/Yes</td>
<td>Delay in achieving milestone, abdominal distention, aggressive behavior, speech delay, hernia, gibbus formation, short stature, upper airways obstruction, dysostosis, joint contractures, scoliosis, hepatosplenomegaly, corneal clouding</td>
</tr>
<tr>
<td></td>
<td>VI-2</td>
<td></td>
<td>3 y/M</td>
<td></td>
<td>Delay in achieving milestone, abdominal distention, aggressive behavior, speech delay, hernia, gibbus formation, short stature, respiratory/ear infection, coarse facial features, waddling in gait, joint contractures, scoliosis, hepatosplenomegaly, corneal clouding</td>
</tr>
<tr>
<td>F</td>
<td>IV-1</td>
<td>1</td>
<td>16.5 y/F</td>
<td>Narman/AJK/Yes</td>
<td>Delay in achieving milestones, abdominal distention, hernia, short stature, coarse facial features, joint contractures, short trunk, hepatosplenomegaly</td>
</tr>
<tr>
<td>G</td>
<td>IV-4</td>
<td>1</td>
<td>6 y/M</td>
<td>Rajput/Punjab/Yes</td>
<td>Delay in achieving milestones, abdominal distention, aggressive behavior, short stature, respiratory/ear infection, joint contractures, short trunk, speech delay, coarse facial features, hepatosplenomegaly, intellectual disability</td>
</tr>
<tr>
<td>H</td>
<td>IV-2</td>
<td>2</td>
<td>14 y/M</td>
<td>Suddhan/ICT/Yes</td>
<td>Delay in achieving milestones, abdominal distention, coarse facial features, aggressive behavior, short stature, respiratory/ear infections, hip dysplasia, joint stiffness, hepatosplenomegaly, difficulty in walking</td>
</tr>
</tbody>
</table>

(Continued on following page)
between LSD-samples and the 1000 genomes South Asian super population of more than 25% were removed. The PCA was calculated using PLINK – pca function on remaining overlapping sites with overall MAF > 0.05, no missingness, and LD-pruned (window size of 1 MB, step size of 1, R^2-threshold of 0.8).

Non-synonymous variants including missense, non-sense, frameshift and indels in coding and splice sites changes in non-coding regions with a MAF of less than 0.01 in large, outbred populations (gnomAD exomes of South Asian population) were considered as putative causal candidates. Further prioritization of these rare variants was based on segregation with disease phenotypes within respective family. As a first phase for WGS data analysis, already reported genes causing MPS were screened for pathogenic variants and during second phase, unsolved families were checked for variants in other LSDs-related genes. In 12 families (A–L) diagnosed with MPS I, variants homozygous in affected individuals and heterozygous in their parents were prioritized based on recessive inheritance pattern of the disease and consanguinity among parents revealed by their pedigrees. The lists of heterozygous variants in each family were searched for compound heterozygous or digenic inheritance. In two families (M, N) diagnosed with MPS II, pathogenic variants on X-chromosome segregated from the mother were extracted. In families where genotypes of parents were available, exomes of each family were analyzed for dual molecular diagnosis (Supplementary Figure S1). As a reference population, more than 100 healthy control samples from the general Pakistani population were analyzed for the prevalence of potentially causal variant as per ACMG guidelines.

The identified variants were subjected to in silico analysis using the following tools: Have (y) Our Protein Explained (HOPE) (https://www3.cmbi.umcn.nl/hope/), Predictor of human deleterious single nucleotide polymorphisms (PhD-SNP) (https://snps.biofold.org/phd-snp/phd-snp.html), Protein analysis through evolutionary relationships (PANTHER) (http://www.pantherdb.org/tools/csnpScoreForm.jsp), I-Mutant (https://folding.biofold.org).

**TABLE 1 (Continued) Showing the details of each patient affected with Mucopolysaccharidosis type I and II.**

<table>
<thead>
<tr>
<th>Family ID</th>
<th>Individual no.</th>
<th>No. of affected individuals</th>
<th>Age/ Sex</th>
<th>Cast/Region/ Consanguinity</th>
<th>Clinical profile</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>V-1</td>
<td>4</td>
<td>12 y/F</td>
<td>Mughal/Punjab/Yes</td>
<td>Joint contractures, corneal clouding, very less eyesight, abdominal distention, coarse facial features, short stature, underweight, umbilical hernia, cannot walk, developmental delay, cardiac issues, hearing issues, respiratory problems, loud snoring</td>
</tr>
<tr>
<td>J</td>
<td>V-2</td>
<td>2</td>
<td>2.5 y/M</td>
<td>Mughal/KPK/Yes</td>
<td>Delay in achieving milestones, aggressive behavior with multiple crying episodes, respiratory/ear infection, no walk, intellectual disability</td>
</tr>
<tr>
<td>K</td>
<td>IV-4</td>
<td>4</td>
<td>8.5 y/M</td>
<td>Abbasi/Punjab/Yes</td>
<td>Delay in achieving milestones, abdominal distention, aggressive behavior, speech delay, umbilical hernia, gibbus formation, short stature, constipation, coarse facial features, cannot walk (Wheel chair bound), joint contractures, scoliosis and kyphosis, hepatosplenomegaly, fish mouth vertebrae, oar shaped ribs, low IQ, widely spaced eyes, narrowing of eye opening (right), drooping eyelids, intellectual disability, developmental delay</td>
</tr>
<tr>
<td>L</td>
<td>IV-5</td>
<td>2</td>
<td>8 y/M</td>
<td>Yousafzai/KPK/Yes</td>
<td>Delay in achieving milestones, abdominal distention, aggressive behavior, speech delay, hernia, short stature, coarse facial features, joint contractures, hepatosplenomegaly, low IQ, intellectual disability, dysostosis multiplex</td>
</tr>
<tr>
<td>M</td>
<td>VI-1</td>
<td>1</td>
<td>4.5 y/M</td>
<td>Raipur Bharti, Khosa Baloch/ Punjab/No</td>
<td>Delay in achieving milestones, abdominal distention, aggressive behavior, speech delay, hernia, short stature, constipation, coarse facial features, joint contractures, hepatosplenomegaly, intellectual disability, hernia</td>
</tr>
<tr>
<td>N</td>
<td>II-1</td>
<td>1</td>
<td>5 y/M</td>
<td>Aryn/Punjab/No</td>
<td>Hernia, short stature, coarse facial features</td>
</tr>
</tbody>
</table>

AIK, Azad Jammu and Kashmir; ESR, erythrocytes sedimentation rate; F, female; HCT, hematocrit; ICT, Islamabad capital territory; IQ, Intelligent quotient; KPK, Khyber Pakhtunkhwa; M, male; MCV, mean corpuscle volume; MCH, mean corpuscle hemoglobin; MCHC, mean corpuscle hemoglobin concentration; m, months; RBG, random blood glucose; y, years.
<table>
<thead>
<tr>
<th>Pedigree/Patient ID</th>
<th>Gene variant</th>
<th>Patients’ genotype</th>
<th>dbsNP</th>
<th>in-silico tools with deleterious output</th>
<th>MAF gnomAD</th>
<th>Classification ACMG</th>
<th>VarSome</th>
<th>HGMD accession number</th>
</tr>
</thead>
<tbody>
<tr>
<td>A/V-4</td>
<td>GALNS</td>
<td>Homozygous</td>
<td>rs118204448</td>
<td>4/7</td>
<td>0.00001972</td>
<td>Pathogenic</td>
<td>PP3, PP4, PP5, PM1, PM2</td>
<td>Pathogenic</td>
</tr>
<tr>
<td>B/V-1</td>
<td>IDUA</td>
<td>Homozygous</td>
<td>rs1560548356</td>
<td>2/4</td>
<td>0.0000007</td>
<td>Likely pathogenic</td>
<td>PP4, PM2, PM4</td>
<td>Likely pathogenic</td>
</tr>
<tr>
<td>C/V-1</td>
<td>SGSH</td>
<td>Homozygous</td>
<td>rs104894636</td>
<td>17/17</td>
<td>0.00002270</td>
<td>Pathogenic</td>
<td>PP5, PS1, PM1, PM2, PM5, PP3</td>
<td>Pathogenic</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>rs752735168</td>
<td>16/18</td>
<td>0.00003306</td>
<td>Likely pathogenic</td>
<td>PM1, PM2, PP3, PP4</td>
<td>Pathogenic</td>
</tr>
<tr>
<td>D/IV-1</td>
<td>ABCA5</td>
<td>Homozygous</td>
<td>rs747667004</td>
<td>7/7</td>
<td>0.0001090</td>
<td>Likely Pathogenic</td>
<td>PM1, PM2, PP3, PP4</td>
<td>Uncertain significance</td>
</tr>
<tr>
<td>E/IV-2-IV-4</td>
<td>IDUA</td>
<td>Homozygous</td>
<td>rs121965027</td>
<td>7/7</td>
<td>0.00004681</td>
<td>Pathogenic</td>
<td>PS3, PM1, PM2, PP4, PP5, PS3</td>
<td>Pathogenic</td>
</tr>
<tr>
<td>F/IV-1</td>
<td>IDUA</td>
<td>Homozygous</td>
<td>rs121965027</td>
<td>7/7</td>
<td>0.00004681</td>
<td>Pathogenic</td>
<td>PS3, PM1, PM2, PP4, PP5, PS3</td>
<td>Pathogenic</td>
</tr>
<tr>
<td>G/IV-4</td>
<td>IDS</td>
<td>Hemizygous</td>
<td>—</td>
<td>4/4</td>
<td>—</td>
<td>Pathogenic</td>
<td>PS1, PM2, PP3, PP4</td>
<td>Pathogenic</td>
</tr>
<tr>
<td>H/IV-2</td>
<td>ALDOB</td>
<td>Homozygous</td>
<td>rs2000585150</td>
<td>11/18</td>
<td>0.0004219</td>
<td>VUS</td>
<td>PM1, PM2, PP3, BP1</td>
<td>Uncertain Significance</td>
</tr>
<tr>
<td>I/IV-1</td>
<td>IDUA</td>
<td>Homozygous</td>
<td>rs121965027</td>
<td>7/7</td>
<td>0.00004681</td>
<td>Pathogenic</td>
<td>PS3, PM1, PM2, PP4, PP5</td>
<td>Pathogenic</td>
</tr>
<tr>
<td>J/IV-2</td>
<td>TRAPPC4</td>
<td>Homozygous</td>
<td>rs375776811</td>
<td>N/A</td>
<td>0.0002403</td>
<td>Pathogenic</td>
<td>PP5, PM2, PP3</td>
<td>Likely pathogenic</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>—</td>
<td>4/7</td>
<td>—</td>
<td>VUS</td>
<td>PM1, PP2, PP3</td>
<td>Uncertain significance</td>
</tr>
<tr>
<td>K/IV-4</td>
<td>IDUA</td>
<td>Homozygous</td>
<td>rs121965027</td>
<td>7/7</td>
<td>0.00004681</td>
<td>Likely Pathogenic</td>
<td>PM1, PM2, PP3, PP4</td>
<td>Pathogenic</td>
</tr>
</tbody>
</table>

(Continued on following page)

3 Results

3.1 Clinical characterization

In total, fourteen families (A-N) of Pakistani origin with phenotypes associated with MPS I and MPS II were included in the present study. Patients showed a wide variety of characteristic symptoms from both MPS I and MPS II including coarse facial features, joint contractures, abdominal distention, hepatosplenomegaly, hernia, and mental abnormalities among others (Table 1). Twelve of fourteen (A-L) families showed consanguinity and clinical phenotypes consistent with MPS I and were consequently labeled as suspected cases of MPS I by the clinicians (Table 1; Figure 1). The other two families (M and N) did not show consanguinity and phenotypes matched clinically with MPS II (Table 1; Figure 2).

3.2 Genetic characterization

Principal component analysis of the LSDs families merged with the 1000 genomes projects showed the LSDs families clustering with the South Asian (SAS) superpopulation (Figure 4A). The SAS population includes a Punjabi population (PJL) verifying the genetic ancestry of the LSDs families. A high rate of consanguinity in MPS I families was also reflected in the genetic data (Figure 4B) and close to 0 for the non-consanguineous MPS II affected individuals.

Analysis of WGS data of available affected and unaffected individuals in each family revealed ten novel and six previously reported homozygous or hemizygous variants in the genes involved in MPS I, MPS II or other lysosomal storage disorders. Five families (C, I, K, J, L) showed multigenic inheritance (Table 2). All the variants were found in ROH in homozygous/hemizygous state in affected individuals while heterozygous or wild type in their parents. In family D, no homozygous or compound heterozygous variants were found in MPS-related genes. Therefore, screening for other LSD-associated or novel candidate genes was performed. The analysis revealed a homozygous missense variant (c.26569G>A; R857C) in ABCA5 located in the ROH. Based on the gene expression, function and pathways, ABCA5 was predicted to be the strongest candidate. Phenotypically normal siblings were either wild type or heterozygous carrier of the respective variants in each family (Figures 1, 2).

These variants were predicted pathogenic, likely pathogenic or variant of uncertain significance according to ACMG classification (Table 2). None of the identified variants was found in homozygous state in any human genome variations databases including

<table>
<thead>
<tr>
<th>Pedigree/Patient ID</th>
<th>Gene variant</th>
<th>Patients' genotype</th>
<th>dbsNP</th>
<th>In-silico tools with deleterious output</th>
<th>MAF gnomAD</th>
<th>Classification ACMG</th>
<th>VarSome</th>
<th>HGMD accession number</th>
</tr>
</thead>
<tbody>
<tr>
<td>L/IV-5</td>
<td>NM_002023.5:c.14697T&gt;C (p.Leu499Pro)</td>
<td>Homozygous</td>
<td>rs69042881</td>
<td>N/A</td>
<td>0.0001591</td>
<td>P3, P1, P2, P4, P5</td>
<td>Uncertain significance</td>
<td>Current study</td>
</tr>
<tr>
<td></td>
<td>MASP1</td>
<td></td>
<td></td>
<td>-</td>
<td></td>
<td>Pathogenic</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>NM_139125.4:c.1767C&gt;T (p.Arg588Ter)</td>
<td>Homozygous</td>
<td></td>
<td>-</td>
<td></td>
<td>Likely pathogenic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M/VI-1</td>
<td>SMARCAL1</td>
<td></td>
<td></td>
<td>-</td>
<td></td>
<td>Pathogenic</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>NM_014140.4:c.1810C&gt;T (p.Gln604Ter)</td>
<td>Homozygous</td>
<td>6/7</td>
<td>-</td>
<td></td>
<td>Pathogenic</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>RRAS2</td>
<td></td>
<td></td>
<td>16/18</td>
<td></td>
<td>Likely pathogenic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M/VI-1</td>
<td>IDS</td>
<td></td>
<td></td>
<td>17/23</td>
<td></td>
<td>Likely pathogenic</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>NM_000202.8:c.1264C&gt;G (p.Cys422Ser)</td>
<td>Hemiogenous</td>
<td></td>
<td>-</td>
<td></td>
<td>Pathogenic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N/III-1</td>
<td>IDS</td>
<td></td>
<td></td>
<td>11/11</td>
<td></td>
<td>Pathogenic</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

N/A, not applicable; VUS, variant of uncertain significance; -, not found.

TABLE 2 (Continued) List of variants identified in the current study.
gnomAD, ExAC, 1000 Genome Project and ethnically matched healthy controls.

4 Discussion

In the present study, fourteen families (A-N) initially diagnosed with MPS underwent WGS to identify potentially pathogenic variants. Among these, twelve families (A-L) were clinically characterized as MPS I while two (M-N) were clinically diagnosed as MPS II. The diagnosis was based on clinical phenotypes including facial features, stature, laboratory tests and inheritance of the phenotypes in pedigree. WGS revealed homozygous variants in previously MPS-associated genes (IDUA, GALNS, SGSH, GAA and IDS) and a novel LSDs-associated candidate gene (ABCA5) in affected individuals. In addition to these genes, variants were found in seven genes (ALDOB, MASP1, KIAA1109, TRAPPC4, SMARCAL1, HERC1, RRAS2) that were previously not reported in MPS but associated with other metabolic disorders.

Our findings in family A (i.e., MPS I), reported a homozygous sequence variant (c.871G>A; A291T) in GALNS. The gene encodes N-acetylgalactosamine-6-sulfatase which is a lysosomal exohydrolase required for the degradation of the glycosaminoglycans, keratan sulfate, and chondroitin 6-sulfate. Previously, the identified variant (A291T) has been reported in compound heterozygous and homozygous forms in two different families of Pakistani origin segregating MPS IV (Tomatsu et al., 1995; Ullah et al., 2017). In-silico analysis revealed that missense mutation (p.A291S) resulted in the formation of additional intramolecular interactions, most likely affecting the conformation of the active site (Asp288 and Asn289) and its substrate binding ability. Therefore, variation at amino acid position 291 of GALNS would affect structure and enzymatic function of the protein.

Multiple mutations in IDUA gene have been identified and associated with MPSI (Tieu et al., 1995). The IDUA gene is located at 4p16.3 and comprises 14 exons. IDUA codes for the enzyme α-L-iduronidase, which is required for the degradation of the GAGs dermatan and heparan sulfate. Deficiency of the enzyme leads to the accumulation of these GAGs in tissues throughout the body, playing a central role in the pathogenesis of MPS I. A novel homozygous insertion mutation (c.1172-1173insGCTGCTGGC; G391insLLA) in IDUA was identified in an affected individual in family B. Glycine at position 391 is involved in the formation of torsion angles due to its high degree of rotational ability which maybe lost after the insertion of LLA leading to abnormal function of the enzyme. Furthermore, affected individuals in families E, F, I and K also revealed a previously reported variant (c.1469T>C; L490P) in IDUA (Gul et al., 2020). In these cases, a high interfamilial variability in the phenotypes of patients segregating same sequence variant (IDUA; c.1469T>C; L490P) was observed in our families (E, F, I, K). Deep analysis of the genome of individuals revealed additional homozygous pathogenic variants in other genes that may act as modifiers or dual molecular diagnosis. Previously sequence variants in KIAA1109 (4q27) encoding a protein playing role in endosomal trafficking and endosome recycling of lipids have been reported to cause Alkuraya-Kucinskas syndrome characterized by brain abnormalities associated with cerebral parenchymal underdevelopment, arthrogryposis, clubfoot, and global developmental delay. Patients in family I showed phenotypes of MPS as well as Alkuraya syndrome. Similarly, sequence variants in MASP1 (3q27.3) encoding a serine protease having an essential role in the innate and adaptive immune response have been associated with 3MC syndrome 1 characterized by widely spaced eyes (hypertelorism), a narrowing of the eye opening (blepharophimosis), droopy eyelids (ptosis) and highly arched eyebrows. Patients in the present family K showed phenotypes of 3MC syndrome in addition to MPS.

FIGURE 4

(A) Principal component analysis of the LSDs families merged with the 1000 Genomes project colored according to super population: African ancestry (AFR), East Asian ancestry (EAS), European ancestry (EUR), and South Asian ancestry (SAS). (B) Distribution of runs of homozygosity-based F-coefficient of the LSDs families.
**IDUA**, **KIAA1109** and **MAP51** are located at different loci of human genome. Due to higher rate of consanguinity, the number of homozgyous regions increases leading to a high risk of occurrence of multiple pathogenic variants. This might lead to complex phenotypes due to co-segregation of two syndromes in the same patient. Therefore, analyzing genomes of populations with a high rate of consanguinity, dual molecular diagnosis should be considered.

In family C, we can appreciate a multi-genic inheritance, segregating two damaging variants in known LSDs enzymes **SGSH** (c.220G>A; **R74C**) and **GAA** (c.1930G>T; **A644S**), in homozygous state, both of which lie within the mapped interval at 17q25.3. **SGSH** encodes a lysosomal enzyme sulfamidase involved in the lysosomal degradation of heparan sulfate. Pathogenic sequence variants in the gene lead to **MPS II** (Weber et al., 1997). Arginine at amino acid position 74 is likely to be involved in the formation of the active site of sulfamidase. Transition of arginine at the position to a non-conserved Cysteine may grossly affect enzyme activity. **GAA** encodes a lysosomal enzyme alpha-glucosidase. The lysosomal alpha-glucosidase is essential for the degradation of glycogen to glucose in lysosomes. Defects in the enzyme lead to glycogen storage disease. In addition to **MPS** phenotypes associated with variants in **SGSH**, the patient showed low hemoglobin concentration which is found in glycogen storage disease caused by pathogenic variant in **GAA**. These findings, suggest that the disease phenotypes in the present patient could be due to the contribution of variants in both genes, **SGSH** and **GAA**.

In family D, we have identified a missense homozygous variant (c.26569G>A; **R857C**) in **ABCA5**. The **ABCA5** protein is a member of the superfamily of ATP binding cassette (ABC) transporters. **ABC** proteins play an important role in transporting various molecules across extra- and intracellular membranes. Loss of function variants in human **ABCA5** have been associated with hair overgrowth in human (DeStefano et al., 2014; Raza et al., 2020). On the other side, **abca5**−/− knockout mice developed lysosomal disease-like symptoms (Kubo et al., 2005). Interestingly, the affected individual in family D carrying **ABCA5** missense variant showed phenotypes overlaying with **MPS** associated with hair overgrowth. To the best of our knowledge, this is the first human case associating a variation in **ABCA5** with **MPS** disease with hair overgrowth.

**ABCA5** has 1642 amino acids long lysosomal peptide which is composed of two domains including **ABC** transporter 1 (478-713) and **ABC** transporter 2 (1290-1533). The protein determines the fate of cholesterol derived from lipoprotein. The functional loss of this enzyme lead to glycogen storage disease. In addition to **MPS** phenotypes associated with variants in **SGSH**, the patient showed low hemoglobin concentration which is found in glycogen storage disease caused by pathogenic variant in **GAA**. These findings, suggest that the disease phenotypes in the present patient could be due to the contribution of variants in both genes, **SGSH** and **GAA**.

In family H, a homozygous missense variant (c.264G>T; **p.D88E**) was identified in **ALDOB**. **ALDOB** encodes fructose-1,6-bisphosphate aldolase, a tetrameric glycolytic enzyme that catalyzes the reversible conversion of fructose-1,6-bisphosphate to glyceraldehyde 3-phosphate and dihydroxyacetone phosphate. Pathogenic variants in the gene have been associated with fructose intolerance. The specific variant (c.264G>T) identified in family H was previously reported by Retterer et al. (2016) in a family segregating abnormality of central nervous system. Interestingly, the affected individual in our family H, showed delay in achieving milestones, abdominal distention, coarse facial features, aggressive behavior, short stature, respiratory/ear infections, hip dysplasia, joint stiffness, hepatosplenomegaly, difficulty in walking, craniofacial disproportion, J-shaped sella, L1 vertebral body is hypoplastic and slightly posteriorly displaced and exaggerated lumbar lordosis. In addition to these phenotypes, the patient showed hypoglycemia, which is an indication of fructosuria. The amino acid aspartic acid D at position 88 make salt bridge with R57, L92 and L321, the mutant residue glutamic acid E may not be able to interact. Wild type residue D is in its preferred secondary structure, a turn, while mutant residue E prefers to be in another secondary structure, which may therefore destabilize the protein conformation. Due to mutated fructose-1,6-bisphosphate aldolase, accumulation of fructose-1,6-bisphosphate in different organs of the body may lead to multi-organ abnormality.

Family J revealed two homozygous variants, one is a missense variant c.5941G>A; p.V1981I in **HERC1** gene and other is splice site variant (c.454 + 3A>G) in **TRAPPC4**. **HERC1** protein is involved in the membrane trafficking via guanine nucleotide exchange factors (GEF) and previously reported as a novel candidate gene for causing intellectual disability (Ortega-Recalde et al., 2015). Recently, Van Bergen and colleagues (Van Bergen et al., 2020) reported the same splice variant underlying early-onset seizures, developmental delay, microcephaly, sensorineural deafness, spastic quadriaparesis and progressive cortical and cerebellar atrophy in families of Caucasian, Turkish and French-Canadian ethnicities. Other authors also identified the same splice site variant in a family of Indian origin having progressive encephalopathy and muscle involvement (Kaur et al., 2020). The primary phenotypes of our patient in family J include delay in achieving milestones, aggressive behavior with multiple crying episodes, respiratory/ear infection, unable to walk, and intellectual disability. Interestingly, phenotypes like muscles weakness, deafness, microcephaly and seizures were not found in our case. The phenotypic variability might be due to different familial backgrounds of the families and effect of rare SNPs in modifier genes. The variant is predicted to cause frameshift of the protein due to mis-splicing of mRNA.

In affected individual of family L, a novel homozygous missense variant c.439C>T; p.R147W in **RRAS2** gene and a novel non-sense variant (c.1810C>T; Q604*) was found in **SMARCAL1**. The substitution in gene **RRAS2** of a positively charged arginine at position 147 with a neutral residue tryptophan may lead to the loss of ionic interactions with surrounding residues. It also leads to distortion of shape as mutant residue tryptophan is bigger in size than wild type residue arginine. **RRAS2** protein is involved in the regulation of MAPK signaling pathway (Capri et al., 2019). The involvement of **RRAS2** gene in some disease is still to be explored and linked with disease etiology in human genome mutation.
We recruited patients who have died and families were recontacted for more detailed evaluation. Most of the patients had MPS II (Sukegawa et al., 1995). The authors found a smaller precursor protein in cells transfected with W345*. This variant predicted to cause loss of function of the protein either through non-sense-mediated mRNA decay or production of a truncated protein.

Regarding families diagnosed as MPS II (M and N) we found pathogenic variants (c.1264C>G; C422S in family M and c.1035C>T; W345* in family N) in IDS segregating with the phenotypes. The novel missense variant C422S affects a non-conserved residue among human sulfatases. Cysteine is more hydrophobic in nature than serine, which causes loss of hydrophobic interactions. Cysteine at position 422 is involved in the sulphide bridge formation which will be lost after this change with serine thus it will affect the stability of protein structure and ultimately loss of function. The other non-sense variant (W345*) in IDS was previously identified in Japanese patient affected with severe phenotypes of MPS II (Sukegawa et al., 1995). The authors found a smaller precursor protein in cells transfected with W345*. Therefore, it is predicted that the variant leads to the production of truncated enzyme. The affected individual in family N showed mild phenotypes of short stature, coarse facial features and hernia as compared to those reported by Sukegawa et al. (1995).

Our study increases the knowledge of genetic factors involved in rare diseases. For this purpose, the Pakistani population is unique due to the high number of families with consanguineous marriages and the high frequency of large pedigrees. Furthermore, we have used WGS instead of phenotype-based diagnosis to diagnose subtypes of disorders having overlapping phenotypes. Thus, WGS is a fruitful strategy to identify novel causal homozygous variants in inbred populations.

The findings of this study have to be taken with some limitations. For example, limited clinical details were gathered from the recruited families due to limited resources like testing facilities at tertiary care hospital at the time of recruitment of the patients. The overlapping phenotypes among the recruited cases presented in this study is one of the main reasons for pseudo-identification of various pathologies diagnosed initially as MPS. After getting results from sequencing data, the families were recontacted for more detailed evaluation but most of the recruited patients have died and families flatly declined to provide more information. Thus, the present study could not include functional enzymatic studies and comprehensive analysis of 3D protein structure. Performing in-vitro and in-vivo functional studies would help in examining detailed biological effects of the identified variants. Further studies including more consanguineous families with similar phenotypes and combined with functional analysis should be performed to clarify the effect of the identified variants.

Genetic findings in the affected families showed that the diagnosis based on phenotypic presentation is not an optimal method. For this reason, the incorporation of next-generation sequencing (NGS) in consanguineous families represents one of the best ways to elucidate the underlying causes in rare diseases such as MPS I and MPS II and establish a genotype-phenotype relationship (Hoffman et al., 2013; Umair et al., 2018). NGS will not only be helpful in the correct diagnosis of the disease but it will also help in prenatal testing, family planning, carrier testing and genetic counseling in families affected with LSDs (Komlosi et al., 2016). Similarly, dual molecular diagnosis has been reported in several cases of consanguineous unions (Umair et al., 2017), which cannot be detected through conventional sequencing or enzymatic testing. Therefore, NGS is required to identify all the pathogenic and modifier variants in heterogeneous disorders like LSDs. Sequencing highly homologous regions (for example, GBA1 locus) using Sanger or short-read NGS methods in suspected cases of LSDs can miss other mutant alleles. Because, the shorter reads cannot be mapped uniquely to the reference genome, especially in cases where there are recombinant alleles aligning to the homologous region. Therefore, long-read NGS is recommended in such cases to discriminate the functional genes from their pseudogenes (Woo et al., 2021).

Accurate and early diagnosis of LSDs in children represents a helpful step for designing therapeutic strategies to save different organs from permanent degeneration. Prenatal screening and identification of carriers’ status in an affected family for LSDs will be helpful for genetic counselling of the family. The identification of a novel MPS I gene (ABCA5) based on WGS proves to be helpful for expanding the limits of targeted therapies to treat LSDs.

Data availability statement

The data presented in the study are deposited in the ClinVar repository, accession numbers: SCV003842213 (IDS:c.1035G>T); SCV003842212 (IDS:c.1264T>G); SCV003842211 (MASP1:c.73A>T); SCV003842210 (KIAA1109:c.1297G>T); SCV003842209 (ABCA5:c.2569C>T); SCV003842208 (SMARCAL1:c.1810G>T); SCV003842207 (HERC1:c.5941G>A); SCV003842206 (IDS:c.353T>A); SCV003842205 (GALC:c.1930G>T); SCV003844061 (RRAS2:c.439C>T).

Ethics statement

The studies involving human participants were reviewed and approved by Bioethical Review board, Quaid-i-Azam University, Islamabad, Pakistan. Written informed consent to participate in this study was obtained from all participants.

Ethics statement
study was provided by the participants’ legal guardian/next of kin. The animal study was provided by the participants and approved by Bioethical Review board, Quaid-i-Azam University, Islamabad, Pakistan.

Author contributions

SF and TH designed and supervised the study experiments. RG and MH collected data and conducted the clinical assessments. RG performed initial lab work. MS, AU, EP, AT, FS, AG, and AA performed WGS and data analysis. RG, AU, and EP wrote the manuscript. All authors revised and approved the article.

Funding

The study was supported by the Novo Nordisk Foundation through the Challenge Program (NNF18OC0033950), and Higher Education Commission, Islamabad, Pakistan under International Research Support Initiative Program (IRISP).

Acknowledgments

We would like to thank all participating families, staff and colleagues who helped us during this study.

References


Clinical application of whole-exome sequencing across clinical indications.


