Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Integrative and systematic genomic approaches to improve diagnosis in rare and undiagnosed diseases: results from the RareBoost project

uğur özbek

ugur.ozbek@ibg.edu.tr

Izmir Biomedicine and Genome Center

Ayca Yigit

Dokuz Eylul University Izmir International Biomedicine and Genome Institute

Mert Pekerbas

Izmir Biomedicine and Genome Center

Barıs Salman

Izmir Biomedicine and Genome Center

Kutay Bulut

Dokuz Eylul University Izmir International Biomedicine and Genome Institute

Emre Ozzeybek

Dokuz Eylul University Izmir International Biomedicine and Genome Institute

Mehmet Mert Topaloglu

Izmir Biomedicine and Genome Center

Burcu Akman

Izmir Biomedicine and Genome Center

Ravza Nur Yildirim

Izmir Biomedicine and Genome Center

Sinem Aktug

Izmir Biomedicine and Genome Center

Tugce Batur

Izmir Biomedicine and Genome Center

Rare and Undiagnosed Diseases Advisory Board Consortium .

https://orcid.org/0009-0002-4174-923X

Article

Keywords: Rare Diseases, Undiagnosed Diseases, Whole Genome Sequencing (WGS), Exome Reanalysis, RNA Sequencing (Transcriptomics), Diagnostic Yield

Posted Date: March 8th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8703835/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is no duality of interest

Version of Record: A version of this preprint was published at European Journal of Human Genetics on July 9th, 2026. See the published version at https://doi.org/10.1038/s41431-026-02177-9.

Abstract

Rare and undiagnosed diseases pose diagnostic challenges due to phenotypic and genetic heterogeneity and the limitations of conventional molecular testing. As part of the RareBoost Project, we implemented an integrated, stepwise genomic strategy in a cohort of 134 patients from 120 families with previously inconclusive genetic testing. Our strategy combined systematic exome with subsequent genome sequencing when indicated, followed by targeted RNA-sequencing, and complemented with comprehensive phenotyping and follow-up segregation analysis. Definitive molecular diagnoses (pathogenic/likely pathogenic variants) were achieved in 25.0% of families, and an additional 25.8% harbored clinically relevant variants of uncertain significance with strong phenotypic correlation. This approach identified non-coding and structural variants while resolving cases through systematic reanalysis of previously negative data. Collectively, our findings demonstrate that integrated genomic and transcriptomic analyses with reanalysis of existing data enhance diagnostic yield in complex rare disease cohorts and provides a scalable framework for implementing genomic medicine in clinical practice.

Background

Rare and undiagnosed diseases affect millions worldwide and represent a growing public health challenge[1]. Between 7,000 and 11,000 identified rare diseases (RDs) affect 3.5–5.9% of the world's population, of which 72–80% are thought to have an underlying genetic background[2]. High rates of first-degree consanguineous marriages (8.1% nationally and reaching up to 20% in southeastern provinces) drive higher RD prevalence in Türkiye[3]. Patients frequently experience prolonged diagnostic journeys because of heterogeneous clinical presentations and limited access to advanced genomic testing[4].

Next-generation sequencing (NGS) technologies have been used as first-tier diagnostic approaches for patients with suspected genetic diseases[5]. Compared to conventional genetic testing, such as clinically specific gene panels and chromosomal microarray (CMA), these high-throughput methods analyze millions of genomic regions, resulting in significantly higher diagnostic rates across rare disease cohorts[6]. In heterogeneous cohorts, the diagnostic yield of exome sequencing (ES) reaches up to 41%, with more than half of the patients without a definitive molecular diagnosis[7].

Unlike ES, genome sequencing (GS) can detect variants in intronic and regulatory regions, as well as copy number variations (CNVs), and other complex genomic alterations[8]. In clinically heterogeneous rare disease cohorts, GS achieves diagnostic yields of about 40%, reflecting its improved ability to capture variant classes frequently missed by ES[8]. Despite GS's broad coverage, definitive diagnosis often remains elusive because of difficulties in interpreting non-coding or regulatory variants, necessitating complementary functional insights from transcriptome-based approaches[9]. In these cases, RNA sequencing offers complementary functional evidence by directly assessing the impact of candidate variants on gene expression and splicing. Across rare disease cohorts, the diagnostic yield of RNA

sequencing shows considerable variability (approximately 8–36%) and is increasingly recognized as a reliable diagnostic approach[9].

In addition, reanalysis of genomic data has become an increasingly important diagnostic strategy. Updates in genetic databases, improved variant annotation and classification frameworks, and evolving patient phenotype information over time enable the re-evaluation of initially undiagnosed cases[10]. Regular and time-aware reanalysis of negative or inconclusive ES/GS results can play an essential role in increasing diagnostic rates, and this strategy should be implemented in clinical practice[7,11].To address these diagnostic challenges and strengthen rare disease research and clinical genomics capacity in Türkiye, the Horizon 2020 ERA Chairs–funded RareBoost initiative was established at the Izmir Biomedicine and Genome Center. One of the main purposes of the RareBoost project is to support the integration of advanced genomic and functional approaches for undiagnosed patients[12]. We applied a stepwise strategy of ES reanalysis, GS, and RNA-seq to 134 patients (120 families) who remained undiagnosed after standard genetic testing. Here, we assess the diagnostic yield of this comprehensive genomic framework in a predominantly exome-negative patient cohort. To operationalize this stepwise diagnostic strategy, the RareBoost project implemented an integrated workflow combining systematic reanalysis, incremental genomic testing, and functional follow-up (Fig. 1).

Materials and Methods

Ethical Approval

The RareBoost Project was approved by the Izmir Biomedicine and Genome Center Non-interventional Research Ethics Board, with protocol number 2023-039, in compliance with the principles of the Declaration of Helsinki[13]. Informed consent was obtained from all participants or their legal guardians prior to participation in the study.

Sample Collection and Biobanking

Whole blood, tissue biopsy, and/or buccal swab samples and clinical/demographic information were collected from patients and their available affected/unaffected parents following informed consent. All samples were processed and stored in accordance with institutional standard operating protocols to ensure traceability, sample integrity, and compliance with applicable ethical and legal regulations. Based on sample availability and clinical indications, DNA and RNA were isolated from available biospecimens, and Primary Peripheral Blood Mononuclear Cells were isolated from whole blood samples and archived under controlled conditions at IBG-Biobank the national biobank node of Türkiye within the BBMRI-ERIC network, which operates in accordance with ISO 20387-2018 “Biotechnology Biobanking – Biobanking requirements”.

Study Cohort

All patients and their family members were referred to the Rare and Undiagnosed Diseases Advisory Board meetings by physician specialists, medical geneticists, pediatricians or other relevant specialists. Eligible participants included (i) probands with a clinically established or suspected rare disease for whom prior exome or genome analyses were considered negative or inconclusive by a certified expert center, and (ii) their family members enrolled for segregation or trio analyses. Proband selection was not restricted by disease category or age of onset. Accordingly, this cohort represents a diagnostically refractory population enriched for unresolved cases, rather than a population-based or first-tier diagnostic cohort.

From September 2023 to November 2025, 134 patients from 120 families were enrolled in the RareBoost Project. Among these families, 38 were enrolled as proband-parent trios, 10 as proband-single-parent duos, 3 as proband-sibling-parents quartets, and 69 as single probands. All clinical data and patient records presented by the referring clinicians were reviewed in detail, and the patients were phenotyped according to Human Phenotype Ontology (HPO)[14] terms to ensure standardized and consistent diagnosis.

Reanalysis of Exome Sequencing

We obtained FASTQ files for ES reanalyses whenever possible from different centers through the referring clinicians.For previously exome-negative or inconclusive cases, prior clinical reports and variant lists were reviewed, followed by raw data reanalysis using updated annotation databases and improved computational tools whenever FASTQ/BAM files were available.

A total of 75 previously negative/inconclusive families underwent ES reanalysis. Sequencing data (FASTQ files) from reanalyzed ES were processed using a Gennext[15] bioinformatic workflow. Reads were aligned to the GRCh38 human reference genome using BWA-MEM2[16]. Single-nucleotide variants (SNVs) and small indels were identified using DeepVariant[17].

Genome Sequencing

For participants without accessible prior raw data or whose reanalysis was negative, GS was performed. When parent samples were available, Trio GS was preferred. Genomic DNA was extracted from whole blood, tissue biopsy or buccal swab samples collected in the IBG-Biobank. DNA underwent library preparation using the Illumina DNA PCR-Free Library Prep Tagmentation kit, and sequencing was performed on the NovaSeq 6000 platform (Illumina, CA, USA) by Health Institutes of Türkiye (TUSEB). A total of 59 families were included in the GS cohort. Among the enrolled patients, 31 families were enrolled as a proband-parent trio (52.5%, 31 of 59), 5 families were enrolled as a proband-single parent duo (8.5%, 5 of 59), and 23 participants were enrolled as proband-only (39%, 23 of 59).

GS results were processed through Illumina DRAGEN Software (v4.3)[18] for alignment and variant calling steps. Further tertiary annotations (including cohort-based variant frequencies) were added using an in-

house-built local run nextflow pipeline, RarePipe v1.0[19]. Additionally, SNVs are annotated using Genebe API[20]. Raw CNV segments were filtered by log2 ratio (<0.8 or >1.2) and aggregated to generate a cohort-level frequency map using BEDTools[21]. Rare variants, defined as occurring in fewer than four individuals, were stratified by size (>10kb and >100kb) and type (deletion or duplication) prior to sample-level intersection. Finally, the pathogenicity of these filtered segments was evaluated according to ACMG guidelines using ClassifyCNV on the hg38 genome build[22].

RNA Sequencing

To further investigate unsolved cases and assess the functional consequences of candidate variants, RNA-seq was performed in a subset of 11 patients. To investigate the functional effect of candidate variants. The aim was subsequently to validate the detected effects in parents using PCR-based targeted methods. Total RNA samples were routinely isolated from each individual enrolled in RareBoost for these analyses.

Total RNA was extracted from the collected samples using the QIAamp RNA Blood or Tissue Mini Kit. Library preparation was performed using the TruSeq Stranded mRNA kit, and sequencing was performed on the NovaSeqX platform. The raw data were processed and analyzed with the Genomics&More platform (PhiTech Inc.). Raw read counts from the patient and control samples were jointly analyzed following normalization using the Trimmed Mean of M-values (TMM) approach as described by Robinson et al. (2010)[23]. A gene or transcript was defined as an expression outlier if its expression level in the patient exceeded the 95th percentile of the GTEx[24] control dataset. FRASER[25].

Variant Filtering and Prioritization

The obtained VCF files were annotated using Gennext[15], Genomize (https://seq.genomize.com/), Franklin (https://franklin.genoox.com), NGS analysis platforms, as well as locally developed RarePipe v1.0[19]. This includes *in silico* prediction tools, including AlphaMissense[26], PromoterAI[27], SpliceAI[28] and allele frequencies from including gnomAD[29] and Turkish Variome[30].

To determine the clinical impact, the variants were assessed based on recommendations determined by the standards and guidelines of the American College of Medical Genetics and Genomics, the Association for Molecular Pathology (ACMG/AMP)[31,32]. Aligned reads were viewed with Integrative Genomics Viewer[33], and identified causative variants were validated using Sanger Sequencing for every available family member to confirm the mode of inheritance.

Stability and Flexibility Change Assessment

The structure of the protein was taken from the AlphaFold[42] Protein Structure Database and submitted to the DynaMut2[34] web application along with the information regarding the variant. Visualizations were composed out of PyMOL software[35] sessions created by DynaMut2[34].

RESULTS

Cohort Characteristics

In the study cohort, 134 unsolved cases (76 male/58 female) from 120 families, aged 0-57 years, were enrolled between September 2023 and November 2025, with suspected genetic diseases and no conclusive genetic diagnosis. Parental consanguinity was 43.3% in the overall cohort. Patients had prior negative genetic test results, including karyotype 49/120 (40.8%), CMA 52/120 (43.3%), targeted gene panel 56/120 (46.7%), and ES 107/120 (89.2%) analysis (Table 1).

Clinically, the cohort primarily presented with neurodevelopmental features, as reflected by the most frequent HPO terms, including seizures (HP:0001250), global developmental delay (HP:0001263), intellectual disability (HP:0001249), EEG abnormality (HP:0002353), and microcephaly (HP:0000252). Metabolic and multisystemic findings were also observed, supporting the cohort's overall clinical heterogeneity.

The cohort was classified into eight distinct clinical categories based on the primary phenotype at presentation (n=120). Multiple Congenital Anomalies (MCA) constituted the largest subgroup (n=38), followed by Neurodevelopmental Disorders (NDD, n=30), Inborn Errors of Metabolism (IEM, n=27), Neuromuscular Disorders (NMD, n=14), Immunology/Rheumatology (n=6), Hematology (n=2), Cardiovascular (n=2) and Adult-onset Neurodegeneration (n=1).

Diagnostic Yield

A stepwise genomic analysis strategy was applied, beginning with ES reanalysis followed by GS and RNA-seq in unsolved cases to maximize diagnostic resolution in this previously undiagnosed cohort. Overall, this combined genomic and transcriptomic approach resulted in a diagnostic yield of 25% (30/120 families), identifying pathogenic or likely pathogenic (P/LP) variants reported in Online Mendelian Inheritance in Man (OMIM) as disease-causing, and yielded clinically relevant candidate VUS variants for 25.8% (31/120) of the families. (Fig. 2, Supplementary Table 1 and 2).

When evaluating the diagnostic yields based on the type of genomic testing performed, ES reanalysis of 75 families identified P/LP variants in 12 families (16%) and clinically relevant VUSs strongly correlated with the clinical phenotype in an additional 19 families (25.3%). Overall, reanalysis provided significant findings in 41.3% (31/75) of the cases. Short-read GS was subsequently applied in cases without a prior molecular diagnosis or in whom ES reanalysis did not yield a conclusive result. GS established a diagnosis in 18 of 59 families (30.5%) based on P/LP variants and identified clinically relevant strong VUSs in 12 families (20.3%)(Supplementary Table 1 and 2).

The Multiple Congenital Anomalies (MCA) group had a definitive diagnostic yield of 21.1% (8/38) and a negative rate of 55.2% (n=21) . Neurodevelopmental Disorders (NDD, n=30) showed a definitive yield of 13.3% (4/30), while 33.3% (n=10) of cases were classified as harboring strong clinically relevant VUS variants. Inborn Errors of Metabolism (IEM, n=27) achieved the highest definitive diagnostic yield among

the major categories at 44.4% (12/27). Neuromuscular Disorders (NMD, n=14) followed with a 28.6% (4/14) definitive yield. In the smaller groups, definitive diagnoses were only reached in Cardiovascular (1/2) and Adult-onset Neurodegeneration (1/1) cases, whereas the Immunology/Rheumatology (n=6) and Hematology (n=2) subgroups yielded no definitive pathogenic variants (Fig. 2b).

Cases Highlighting the Impact of ES Reanalysis and Up-to-date Genetic Analysis

The significant contribution of exome reanalysis to diagnostic yield is mainly due to the time elapsed since the initial tests. During this period, newly described syndromes that explain the phenotype emerged in the literature. A representative case illustrating the importance of the timing of genomic analysis using the most up-to-date data was Patient 36. He had previously undergone ES at another center in 2021, which yielded inconclusive results. Upon enrollment, since the ES data was inaccessible for reanalysis, we proceeded directly to GS. GS identified a likely pathogenic homozygous missense variant (p.Arg88Cys) in the *MRPL49* gene (Fig. 3a). The association of this gene with recessive combined oxidative phosphorylation deficiency was established only in 2025 by Thomas et al. in a cohort of 13 patients, leading to a recent update of its phenotype in the OMIM database[36]. Consequently, this variant was not reported in the initial 2021 analysis, as the gene-disease association had not yet been established. To further investigate the pathogenicity of the variant, stability and flexibility change assessment was performed using Dynamut2 and a Predicted Stability Change ($\Delta\Delta G^{Stability}$) score of -0.64 kcal/mol (Destabilizing) further confirmed this to be the causative variant (Fig. 3b).

Supported by in-silico pathogenicity predictions, this variant ended the 17-year-old patient's 14-year diagnostic odyssey. Notably, although the variant is theoretically detectable by standard ES, the diagnosis was ultimately achieved through GS combined with up-to-date literature curation.

Diagnoses Benefitting From Genome Sequencing

In our cohort, one advantage of GS over ES was its ability to detect both structural variants and copy number variations. For example, Patient 62, in whom GS identified compound heterozygous variants in *CNTN1* gene: an 18.6 kb maternally inherited exonic deletion involving exons 21–23 (seq[GRCh38] 12q12(41,021,576_41,040,185)x1) and a paternally inherited truncating splice-site variant (c.703+1G>A) (Fig. 3c-3e). The deletion variant may not have been detectable by ES approaches since detecting CNVs with ES is less successful than GS, whereas array-based methods would not have captured the splice-site variant. The simultaneous detection of a structural variant and a sequence-level variant within the same gene underscored the advantage of GS over ES and array-based methods in resolving pathogenic variation of different molecular classes with a single diagnostic test.

For patient 51, GS identified a homozygous synonymous c.630G>C (p.Pro210=) variant in the *CHKA* gene, located at the terminal nucleotide of exon 4. The prior exome analysis was negative due to a misannotation of the splice variant as synonymous. As a synonymous change, this variant would not typically be prioritized in standard variant interpretation. However, during GS analysis, SpliceAI predicted loss of the canonical donor splice site (score: 0.74), likely resulting in exon skipping. The patient

presented with global developmental delay, microcephaly, hypotonia, seizures, and dysmorphic features, a phenotype consistent with *CHKA*-related "Neurodevelopmental disorder with microcephaly, movement abnormalities, and seizures" (MIM #620023), previously described in individuals with bi-allelic pathogenic variants in *CHKA* gene. We subsequently confirmed the functional impact of the variant through RT-PCR (gel electrophoresis) and RNA-seq, both of which demonstrated the predicted splicing aberration (Fig. 3f-3h).

Another GS assisted case was Patient 5 with a 270 kb deletion upstream of *LMNB1*, responsible for autosomal dominant leukodystrophy (MIM #619179), which has previously been established as pathogenic. This large deletion encompassed four genes (*GRAMD2B, ALDH7A1, PHAX, TEX43)*. Although ES-based CNV algorithms can detect exonic copy-number changes, their identification remains unreliable due to limited sensitivityfor variants larger than 100 kb[37].

Genome sequencing is also highly effective in characterizing mitochondrial genome alterations, successfully identifying SNVs and CNVs alongside complex structural variants. We established a diagnosis in one of the cases by identifying a 7.7 kb heteroplasmic mitochondrial deletion causing a single large-scale mitochondrial deletion syndrome through GS performed on a cardiac biopsy sample obtained during an interventional procedure. This deletion was previously undetectable in blood-based tests, including mitochondrial multiplex ligation-dependent probe amplification, demonstrating the limitations of peripheral blood as a diagnostic sample for disorders involving somatic mosaicism or tissue-specific heteroplasmy. Similarly, for transcriptomic investigations, blood-derived RNA sequencing provides limited diagnostic value for genes with restricted or tissue-specific expression, underscoring the need for tissue-appropriate sampling in functional genomic analyses.

RNA Sequencing Leads to Definitive Diagnoses

To further investigate unsolved cases and assess the functional consequences of candidate variants, RNA-seq was performed and this transcriptome-level mRNA analysis revealed aberrant splicing in 4 cases, supporting the pathogenicity of variants and contributing to the overall diagnostic yield. Notably, functional analysis of the compound heterozygous *RNU4ATAC* variants revealed severe splicing defects in Patient 16, characterized by significant minor intron retention events in 354 of the 691 minor intron–containing genes presented in the Olthof et al.[38](Fig 4a). We utilized the FRASER θ metric, where decrease in splicing efficiency scores indicates aberrant intron retention. This transcriptome-wide outlier approach, adopted from Arriaga al. (2025)[39] provided critical functional support for the pathogenicity of the non-coding variant detected by GS (Fig 4b).

Among the identified variants, missense variants represented the most frequent variant type, followed by truncating, splice-site, non-coding, and CNVs. Consistent with these characteristics, autosomal recessive inheritance predominated, with 50.8% (31/61) of diagnosed probands carrying homozygous or compound heterozygous variants, while autosomal dominant inheritance accounted for 39.3% (24/61), X-linked inheritance for 8.1% (5/61), and mitochondrial inheritance (1.6%) (1/61). Notably, 25.8% (8/31)

of families with VUS harbored variants in strong candidate genes for novel disease associations. In addition, several P/LP or clinically relevant VUS were identified in genes not previously associated with human disease in OMIM, including *CAMTA1, BECN2, FGF22, ATP1B4, BSN, LRP1,* and *PHKG1*, expanding the current gene–disease landscape (Table 2). Following comprehensive genomic and transcriptomic analyses, 59 families remained undiagnosed and were identified as candidates for further genomic investigation using long-read genome sequencing.

DISCUSSION

Our findings demonstrate the diagnostic value of a comprehensive genomic approach, particularly in a phenotypically heterogeneous patient cohort in which many individuals had previously undergone extensive genetic testing yet still lacked a definitive molecular diagnosis. A significant proportion of patients were referred after inconclusive evaluations at other centers, enriching the cohort with diagnostically refractory cases and leaving few, if any, easily diagnosable cases. Even within this challenging context, the systematic integration of deep phenotyping and the stepwise application of advanced genomic technologies enabled a relatively high diagnostic yield.

In previous studies, ES reanalysis has been reported to provide an additional diagnostic yield of 10–25% across various cohorts, whereas the diagnostic yield of GS over ES varies widely, ranging from 5% to 40%[7,8,46]. Consistent with these findings, our study achieved a 16% definitive diagnostic rate through ES reanalysis. GS yielded a 25% diagnostic rate in our predominantly ES-negative cohort, which is higher than that reported in many studies. This can be attributed to several factors. The initial ES analyses were performed several years ago, and we had access only to the clinical reports. Second, the lack of access to raw sequencing data for these cases precluded reanalysis. Furthermore, our cohort originated from a population with a high rate of consanguinity (43.3%), resulting in a predominance of autosomal recessive inheritance over autosomal dominant inheritance (50.8% vs. 39.3%). This high rate of consanguinity likely contributed to our enhanced diagnostic yield compared to other studies in the literature.

The diagnostic yield observed across phenotypic subgroups reflects the heterogeneous genetic architecture of the rare diseases within this cohort. The relatively high rate of negative results in the MCA group (55.2%) underscores the inherent complexity of structural malformations, which are frequently driven by structural variants, or low-level mosaicism that current bioinformatics pipelines may not capture. In contrast, the 44.4% definitive success rate in the IEM group highlights the critical efficiency of assessing biochemical and genomic data together; in these cases, specific metabolic abnormalities significantly helped the variant prioritization and clinical interpretation. The NDD subgroup presented a distinct challenge, characterized by a burden of clinically relevant VUSs (33.3%). This high prevalence suggests that, although the likely molecular etiology is frequently identified through genomic analyses, a definitive classification is often limited by insufficient functional evidence or a well-established clinical association. Consequently, these NDD cases represent a primary candidate pool for longitudinal follow-up and functional assays to bridge the gap between variant detection and definitive diagnosis.

Given this multi-tiered analytical framework, our first step was to re-evaluate the case to assess the diagnostic potential of systematic reanalysis (when previous data are available) before implementing a genome-wide approach. ES reanalysis was a cost-effective initial step, providing a definitive molecular diagnosis for 16% (12/75) of the families and detecting clinically relevant VUSs in 25.3% (19/75) of the families. Several cases had not been reanalyzed for years, and many exome data were generated using outdated capture kits, resulting in variable coverage and inconsistent quality.

These findings indicate that most variants identified through GS were not truly “genome-required” but rather “exome-detectable” under improved data quality and updated interpretation frameworks, including the re-analyzing of variants with up-to-date literature, emphasizing the need for systematic reanalysis. Among the 59 families with GS, 30.5% (18/59) achieved a definitive molecular diagnosis, while 20.3% (12/59) carried clinically relevant VUSs. Notably, 14.8% (8/59) of the diagnosed families harbored causal variants that were not detected by their prior ES analyses. Three of these eight cases involved variants that, in principle, could have been captured by ES if sequencing coverage had been sufficient. For example, Patient 16 carried bi-allelic pathogenic variants in *RNU4ATAC* (MIM #210710, #616651), a non-coding RNA gene that has recently been implicated in disease[47]. This locus was not covered by the outdated ES capture kit used for that case.

Collectively, these findings highlight the diagnostic value of GS in uncovering causative variants located in non-coding, regulatory, and structurally complex regions of the genome. GS thereby enables a more comprehensive mechanistic interpretation of disease and facilitates future functional and translational research. At the same time, the subset of cases that might have been resolved through optimized ES underscores the importance of standardized coverage and strict quality-control metrics within clinical sequencing workflows.

RNA-seq served as a functional complement to DNA-seq and contributed to diagnosis in 4 families. Specifically, the diagnosis was confirmed in two patients when RNA-seq analysis demonstrated that splice-region VUS variants identified by GS led to abnormal splicing in the relevant gene. In the third patient, a homozygous synonymous variant previously detected by ES but with a very low SpliceAI prediction score was shown to cause exon skipping. In the fourth patient, transcriptome analysis revealed a global splicing defect in minor intron-containing genes with compound heterozygous variants in *RNU4ATAC*.

Notably, among the 61 families who received a molecular diagnosis or a clinically relevant VUS, 38 had been on a diagnostic journey exceeding five years prior to enrollment in the RareBoost study, highlighting the prolonged diagnostic timelines typically experienced by individuals with rare diseases. By combining GS with standardized variant interpretation and detailed phenotypic assessment, our study enabled diagnoses in these undiagnosed cases and facilitated improved clinical characterization. These findings underscore the value of comprehensive and coordinated genomic approaches in resolving complex cases within rare disease populations.

While this study demonstrates the value of an integrated, stepwise genomic diagnostic framework, several limitations should be taken into account when interpreting the findings. First, ES reanalysis was complicated by variability in data quality, as some raw sequencing files originated from older platforms or lacked information on library preparation. In addition, both GS analyses and ES reanalyses were limited to short-read technology, potentially missing repeat expansions and complex structural variants. In several cases, segregation or functional follow-up could not be performed due to loss of patient contact. RNA-seq was performed using blood-derived RNA, which may have limited the detection of expression and splicing abnormalities in genes with tissue-specific expression and contributed unresolvedcases[48]. While functional studies are underway for a subset of VUSs in our cohort, definitive classification of these variants remains a worldwide challenge. Practical constraints, most notably time and resources, prevent comprehensive reclassification of all clinically relevant VUSs in our study. Taken together, these limitations indicate that the diagnostic yields reported here should not be directly compared with those from routine diagnostic settings.

In conclusion, this study demonstrates the effectiveness of a comprehensive, stepwise genomic approach in improving diagnostic outcomes in rare and undiagnosed diseases. By combining GS, ES, RNA-seq, and systematic reanalysis, diagnostic resolution was increased in a previously unresolved cohort. Importantly, this work contributes to the development of rare disease diagnostic capacity and research infrastructure in Türkiye and provides a model for cohort-based investigations that can be expanded with emerging technologies. Future efforts will focus on further increasing diagnostic yield through the integration of long-read sequencing and multi-omic approaches, as well as functional evaluation of strong candidate variants to resolve remaining unsolved cases and accelerate novel gene discovery.

Declarations

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author upon reasonable request, in accordance with ethical approvals and applicable data protection regulations.

Availability of data and materials

Not applicable.

Declaration of Interest

The authors declare no conflict of interest.

Funding

We would like to acknowledge the support of the RareBoost project, funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 952346. This initiative has

substantially enhanced the research and innovation capacity in the field of rare diseases at the Izmir Biomedicine and Genome Center (IBG) and across the broader region of Türkiye, fostering progress in diagnostics and further studies of rare and undiagnosed conditions.

Contributions

Conceptualization: AY, MP, KB, BS, EO, MMT, BA, UO. Data curation: AY, MP, KB, BS, EO, MMT, BA. Formal analysis: AY, MP, MMT, EO, BS, RNY. Funding acquisition: UO. Investigation: AY, MP, MMT, BS, EO, BA, RNY. Methodology: AY, MP, MMT, BS, EO, RNY. Project administration: UO, SA. Resources: UO, Rare and Undiagnosed Diseases Advisory Board Consortium. Software: EO, BS, MP. Supervision: UO, BA. Validation: AY, SA, TB. Visualization: AY, MP, EO, BS, KB. Writing—original draft: AY, MP, MMT, BS. Writing—review & editing: all authors. All authors read and approved the final manuscript.

Rare and Undiagnosed Diseases Advisory Board Consortium:

Ahmet Okay Caglayan[1], Aliye Kubra Unal[1], Inci Yaprak[1], Beste Ozkalay[2,6], Caglar Celebi[5,6], Evin Iscan[1], Miray Sevinin[1], Mehmet Baysan[1], Nese Atabey[6], Sanem Tercan Avci[6], Yavuz Oktay[1], Ayse Ipek Polat[7], Ayse Semra Hiz[7], Uluc Yis[7], Huseyin Bahadır Senol[7], Mustafa Halk[7], Elif Naz Kadem[7], Dilek Sonmezoglu[7], Asude Durmaz[8], Ayca Aykut[8], Haluk Akin[8], Pinar Gencpinar[9], Safa Mete Dagdas[9], Mehmet Semiz[9], Ozlem Giray[10], Semra Gursoy[10], Seda Gunes[11], Ozlem Akgun Dogan[12], Bulent Kara[13], Esra Isik[14], Tahir Atik[14], Merve Bilen[15], Nur Arslan[15], Pelin Teke[15], Abdurrahman Akgun[16], Arzu Akyay[17], Aslihan Uzun[18], Aydan Mengubas Erbas[19], Berk Ozyilmaz[20], Fatma Silan[21], Gokhan Ozan Cetin[22], Ilknur Surucu Kara[23], Korcan Demir[24], Ozlem Unal Uzun[25], Sevil Yildiz[26], Tolga Polat[27], Sezgin Sahin[28], Ercan Mihci[29], Gonca Kilic Yildirim[30], Fehime Erdem[31]

[1]Rare and Undiagnosed Diseases Group, Izmir Biomedicine and Genome Center, Izmir, Türkiye.

[2]Izmir International Biomedicine and Genome Institute, Dokuz Eylul University, Izmir, Türkiye.

[3]Institute of Graduate Studies in Health Sciences, Istanbul University, Türkiye.

[4]Aziz Sancar Institute of Experimental Medicine,Istanbul University, Türkiye.

[5]Ege University, Institute of Health Sciences, Izmir, Türkiye.

[6]Biobank and Biomolecular Resources, Izmir Biomedicine and Genome Center, Platform, Izmir, Türkiye.

[7]Department of Pediatric Neurology, Faculty of Medicine, Dokuz Eylül University, Izmir, Türkiye.

[8]Department of Medical Genetics, Faculty of Medicine, Ege University, Izmir, Türkiye.

[9]Department of Pediatric Neurology, Izmir Bayrakli City Hospital, Izmir, Türkiye.

[10]Department of Pediatric Genetics, Faculty of Medicine, Dokuz Eylül University, Izmir, Türkiye.

[11]Department of Pediatric Metabolism, Faculty of Medicine, Harran University, Sanliurfa, Türkiye.

[12]Department of Medical Genetics, Faculty of Medicine, Acibadem Mehmet Ali Aydinlar University, Istanbul, Türkiye.

[13]Department of Pediatric Neurology, Faculty of Medicine, Kocaeli University, Kocaeli, Türkiye.

[14]Department of Pediatric Genetics, Faculty of Medicine, Ege University, Izmir, Türkiye.

[15]Department of Pediatric Metabolism, Faculty of Medicine, Dokuz Eylül University, Izmir, Türkiye.

[16]Department of Pediatric Metabolism, Faculty of Medicine, Fırat University, Elazıg, Türkiye

[17]Department of Pediatric Hematology, Faculty of Medicine, Inonu University, Malatya, Türkiye.

[18]Department of Pediatric Rheumatology, Faculty of Medicine, Dokuz Eylül University, Izmir, Türkiye.

[19]Department of Medical Genetics, Faculty of Medicine, Aydın Adnan Menderes University, Aydın, Türkiye.

[20]Department of Medical Genetics, Izmir Bayrakli City Hospital, Izmir, Turkiye

[21]Department of Medical Genetics, Faculty of Medicine, Canakkale Onsekiz Mart University, Canakkale, Türkiye.

[22]Department of Medical Genetics, Faculty of Medicine, Pamukkale University, Denizli, Türkiye.

[23]Department of Pediatric Metabolism, Faculty of Medicine, Gaziantep University, Gaziantep, Türkiye

[24]Department of Pediatric Endocrinology, Faculty of Medicine, Dokuz Eylül University, Izmir, Türkiye.

[25]Department of Pediatric Metabolism, Faculty of Medicine, Kocaeli University, Kocaeli, Türkiye.

[26]Department of Pediatric Metabolism, Bursa Yüksek İhtisas Training and Research Hospital, Bursa, Türkiye.

[27]Department of Neurology, Faculty of Medicine, Dokuz Eylül University, Izmir, Türkiye.

[28]Department of Pediatric Rheumatology, Faculty of Medicine, Istanbul University-Cerrahpasa, Istanbul, Türkiye.

[29]Department of Pediatric Genetics, Faculty of Medicine, Akdeniz University, Antalya, Türkiye.

[30]Department of Pediatric Metabolism, Eskisehir Osmangazi University, Eskisehir, Türkiye.

[31]Department of Pediatric Metabolism, Van Training and Research Hospital, Van, Türkiye.

Acknowledgements

The authors thank the Izmir Biomedicine and Genome Center Biobank and Biomolecular Resources Platform (IBG-Biobank) for the use of its facilities and services, as well as for collecting, processing, storing, and making available the biological samples and data used in this study. This work was supported by the RareBoost Project, funded by the European Union's Horizon 2020 research and innovation programme (Grant No. 952346; https://rareboost.ibg.edu.tr/), and by the ERDERA Project (Grant No. 101156595).

We also gratefully acknowledge TÜSEB for enabling genome sequencing analyses for our patients.We sincerely thank all clinicians, researchers, and members of the Rare and Undiagnosed Disease Platform (RUDiP) who contributed to the Rare and Undiagnosed Patients Advisory Board meetings, as well as the patients and their families for their invaluable participation. We further acknowledge Gennext and Genomics and More for their genomic and transcriptomic services. Finally, we thank Prof. Dr. Sibel Ugur Iseri for the critical reading of the manuscript.

References

1. Chung, C. C. Y., Hong Kong Genome Project, Chu, A. T. W. & Chung, B. H. Y. Rare disease emerging as a global public health priority. *Front. Public Health* 10, 1028545 (2022).
2. Nguengang Wakap, S. *et al.* Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. *Eur. J. Hum. Genet.* 28, 165–173 (2020).
3. Turkish Statistical Institute (TÜİK). (2025). Statistics on Family, 2024. Retrieved December 22, 2025, from https://data.tuik.gov.tr/Bulten/Index?p=Statistics-on-Family-2024-53898.
4. Hatirnaz Ng, O. *et al.* Obstacles and expectations of rare disease patients and their families in Turkiye: ISTisNA project survey results. *Front Public Health* vol. 10 1049349 (2022).
5. Slaba, K., Pokorna, P., Jugas, R. et al. Diagnostic efficacy and clinical utility of whole-exome sequencing in Czech pediatric patients with rare and undiagnosed diseases. Sci Rep 14, 28780 (2024). https://doi.org/10.1038/s41598-024-79872-4.
6. Pandey, R. *et al.* A meta-analysis of diagnostic yield and clinical utility of genome and exome sequencing in pediatric rare and undiagnosed genetic diseases. *Genet. Med.* 27, 101398 (2025).
7. Fung, J. L. F. *et al.* A three-year follow-up study evaluating clinical utility of exome sequencing and diagnostic potential of reanalysis. *NPJ Genom Med* vol. 5 37 (2020).
8. Lionel, A. C. *et al.* Improved diagnostic yield compared with targeted gene sequencing panels suggests a role for whole-genome sequencing as a first-tier genetic test. *Genet. Med.* 20, 435–443 (2018).

9. Yépez, V. A. *et al.* Clinical implementation of RNA sequencing for Mendelian disease diagnostics. *Genome Med.* 14, 38 (2022).
10. Bullich, G. *et al.* Systematic Collaborative Reanalysis of Genomic Data Improves Diagnostic Yield in Neurologic Rare Diseases. *J Mol Diagn* vol. 24 529–542 (2022).
11. Liu, P. *et al.* Reanalysis of Clinical Exome Sequencing Data. *N Engl J Med* vol. 380 2478–2480 (2019).
12. Durmus, S. *et al.* Management of rare and undiagnosed diseases: insights from researchers and healthcare professionals in Türkiye. *Frontiers in Public Health* vol. 12 (2025).
13. World Medical, A. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. *JAMA* vol. 310 2191–4 (2013).
14. Kohler, S. *et al.* The Human Phenotype Ontology in 2021. *Nucleic Acids Res* vol. 49 D1207–D1217 (2021).
15. Geniva Informatics. Gennext workflows [computer software]. Zenodo; 2025. DOI: 10.5281/zenodo.17426020.
16. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* vol. 25 1754–60 (2009).
17. Poplin, R. *et al.* A universal SNP and small-indel variant caller using deep neural networks. *Nat Biotechnol* vol. 36 983–987 (2018).
18. Behera, S. *et al.* Comprehensive genome analysis and variant detection at scale using DRAGEN. *Nat. Biotechnol.* 43, 1177–1191 (2025).
19. Barış Salman, Emre Özzeybek. (2026). Ozbek-Lab/rarepipe: Initial Release (v0.0.1). Zenodo. https://doi.org/10.5281/zenodo.18169888.
20. Stawiński, P. & Płoski, R. Genebe.net: Implementation and validation of an automatic ACMG variant pathogenicity criteria assignment. *Clin. Genet.* 106, 119–126 (2024).
21. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842 (2010).
22. Gurbich, T. A. & Ilinsky, V. V. ClassifyCNV: a tool for clinical annotation of copy-number variants. *Sci. Rep.* 10, 20375 (2020).
23. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* 11, R25 (2010).
24. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* 45, 580–585 (2013).
25. Mertes, C. *et al.* Detection of aberrant splicing events in RNA-seq data using FRASER. *Nat. Commun.* 12, 529 (2021).
26. Cheng, J. *et al.* Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* vol. 381 eadg7492 (2023).
27. Jaganathan, K. *et al.* Predicting expression-altering promoter mutations with deep learning. *Science* 389, eads7373 (2025).

28. De Sainte Agathe, J.-M. *et al.* SpliceAI-visual: a free online tool to improve SpliceAI splicing variant interpretation. *Hum. Genomics* 17, 7 (2023).
29. Chen, S. *et al.* A genome-wide mutational constraint map quantified from variation in 76,156 human genomes. *bioRxiv* 2022.03.20.485034 (2022) doi:10.1101/2022.03.20.485034.
30. Kars, M. E. *et al.* The genetic structure of the Turkish population reveals high levels of variation and admixture. *Proc. Natl. Acad. Sci.* 118, e2026076118 (2021).
31. Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med* vol. 17 405–24 (2015).
32. Riggs, E. R. *et al.* Technical standards for the interpretation and reporting of constitutional copy-number variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics (ACMG) and the Clinical Genome Resource (ClinGen). *Genet. Med.* 22, 245–257 (2020).
33. Thorvaldsdottir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* vol. 14 178–92 (2013).
34. Rodrigues, C. H., Pires, D. E. & Ascher, D. B. DynaMut: predicting the impact of mutations on protein conformation, flexibility and stability. *Nucleic Acids Res.* 46, W350–W355 (2018).
35. Rosignoli, S. & Paiardini, A. Boosting the Full Potential of PyMOL with Structural Biology Plugins. *Biomolecules* 12, 1764 (2022).
36. Thomas, H. B. *et al.* Biallelic variants in MRPL49 cause variable clinical presentations, including sensorineural hearing loss, leukodystrophy, and ovarian insufficiency.
37. Nmezi, B. *et al.* An oligodendrocyte silencer element underlies the pathogenic impact of lamin B1 structural variants. *Nat. Commun.* 16, 1373 (2025).
38. Olthof, A. M., Hyatt, K. C. & Kanadia, R. N. Minor intron splicing revisited: identification of new minor intron-containing genes and tissue-dependent retention and alternative splicing of minor introns. *BMC Genomics* 20, 686 (2019).
39. Arriaga, M. T. *et al.* Transcriptome-wide outlier approach identifies individuals with minor spliceopathies. Preprint at https://doi.org/10.1101/2025.01.02.24318941 (2025).
40. Quiles JM, Najor RH, Gonzalez E, Jeung M, Liang W, Burbach SM, Zumaya EA, Diao RY, Lampert MA, Gustafsson ÅB. Deciphering functional roles and interplay between Beclin1 and Beclin2 in autophagosome formation and mitophagy. Sci Signal. 2023;16(770):eabo4457. doi: 10.1126/scisignal.abo4457. Epub 2023 Jan 31. PMID: 36719945; PMCID: PMC10019900.
41. Singh, R. *et al.* Fibroblast Growth Factor 22 Contributes to the Development of Retinal Nerve Terminals in the Dorsal Lateral Geniculate Nucleus. *Front. Mol. Neurosci.* 4, (2012).
42. Miyake, A., Ohmori, T. & Murakawa, Y. Fgf22 and Fgfr2b are required for neurogenesis and gliogenesis in the zebrafish forebrain. *Biochem. Biophys. Res. Commun.* 681, 212–217 (2023).

43. Sun, J., Zheng, Y., Chen, Z. & Wang, Y. The role of Na^+ -K^+ - ATPase in the epileptic brain. *CNS Neurosci. Ther.* 28, 1294–1302 (2022).
44. Ye, T. *et al.* Variants in *BSN* gene associated with epilepsy with favourable outcome. *J. Med. Genet.* 60, 776–783 (2023).
45. Mark, P. R. *et al.* Autosomal recessive LRP1-related syndrome featuring cardiopulmonary dysfunction, bone dysmorphology, and corneal clouding. *Mol. Case Stud*.
46. Lee, S. *et al.* Clinical utility of genome sequencing in rare diseases: lessons from a single-center study of 1,452 Korean families. *Npj Genomic Med.* https://doi.org/10.1038/s41525-025-00538-9 (2025) doi:10.1038/s41525-025-00538-9.
47. Tabib, A., Richmond, C. M. & McGaughran, J. Delineating the phenotype of RNU4ATAC-related spliceosomopathy. *Am. J. Med. Genet. A.* 191, 1094–1100 (2023).
48. Hou, H. *et al.* Assessing the diagnostic impact of blood transcriptome profiling in a pediatric cohort previously assessed by genome sequencing. Preprint at https://doi.org/10.1101/2024.12.03.24317221 (2024).

Tables

Tables are available in the Supplementary Files section.

Figures

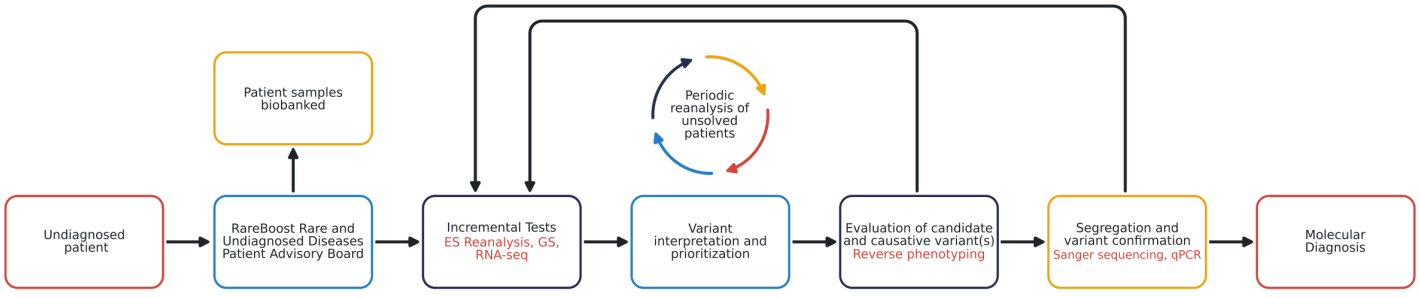


Figure 1

RareBoost diagnostic workflow for rare and undiagnosed diseases. Overview of the RareBoost diagnostic pathway for patients with rare and undiagnosed conditions. Undiagnosed patients are evaluated by the Rare and Undiagnosed Diseases Patient Advisory Board, followed by centralized variant

interpretation and systematic reanalysis of available prior data. Unresolved cases undergo incremental genomic and transcriptomic testing, including ES reanalysis, GS, and RNA-seq. Candidate variants are assessed through reverse phenotyping, segregation analysis, and confirmatory testing, with functional studies to support molecular diagnosis.

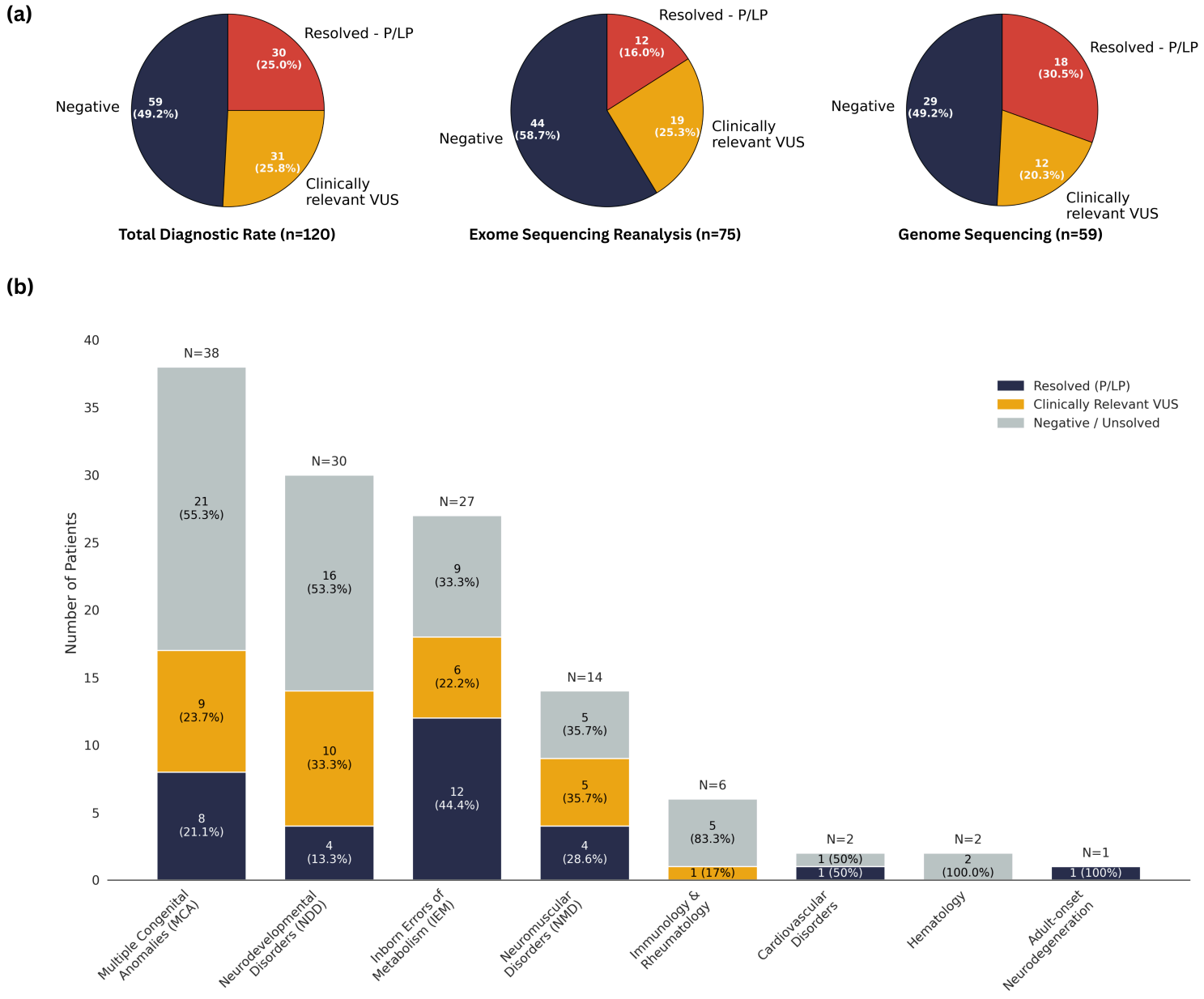


Figure 2

Diagnostic yields of ES Reanalysis/GS and distribution of diagnostic yields by disease groups. (a) Pie charts display the rates of Pathogenic/Likely Pathogenic (P/LP) definitive diagnoses and clinically relevant VUSs for the entire cohort, as well as ES reanalysis and GS separately (b) The diagnostic rates and clinically relevant VUS rates are categorized according to specific phenotypic disease groups, illustrating the performance of genomic testing across different clinical presentations.

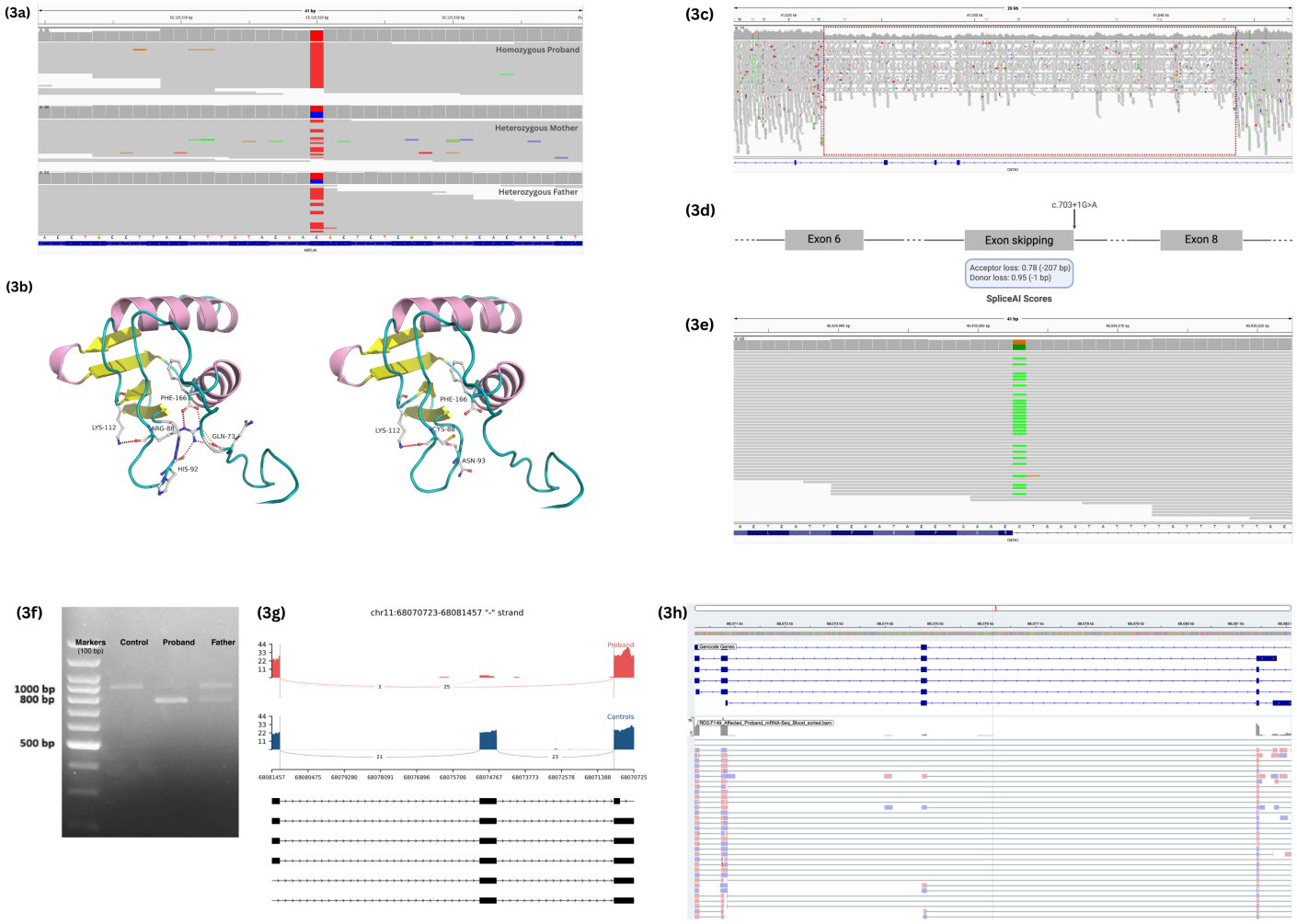


Figure 3

Representative genomic findings and supporting functional evidence from solved cases identified through genome sequencing

(3a-3b) Homozygous missense variant in the *MRPL49* gene identified via GS with up-to-date literature (a) IGV representation of the detected variant. (b) Visualization of the wild type (left) and R88C mutant (right) *MRPL49* protein structure. With a Predicted Stability Change ($\Delta\Delta G^{Stability}$) score of -0.64 kcal/mol (Destabilizing), the change from positively charged arginine seen in the wild type protein (left) to polar cysteine in the mutated version (right) leads to loss of hydrogen bonds between Cys-88 and Gln-73, His-92 and Phe-166. **(3c-3e) Compound heterozygous *CNTN1* CNV and splice-site variant identified by GS** (c) An 18.6 kb intragenic deletion spanning exons 21, 22, and 23 in the maternal allele (shown in the red-dotted rectangle). The IGV (Integrative Genomics Viewer) tracks show the loss of read coverage and discordant reads across the affected region. (d) The splice-site variant c.703+1G>A affects the splice donor GT sequence at the end of exon 7. SpliceAI analysis predicts a high-confidence loss of both the splice donor (score: 0.95) and splice acceptor (score: 0.78) sites. (e) IGV representation of the splice variant in the paternal allele. **(3f-3h) Functional validation of the splicing defect using RT-PCR and RNA-seq.** (f) RT-PCR analysis of the *CHKA* transcript. Using forward and reverse primers located in exons 3 and 5, RT-PCR amplified a 980 bp wild-type (WT) band and an 866 bp mutant band corresponding to the

skipping of exon 4. The gel electrophoresis demonstrates the expected distribution: a single WT band in the control, predominantly the mutant band in the homozygous proband, and both bands in the heterozygous father. Since the skipping of the 114 bp (38 amino acid) exon 4 maintains the open reading frame (in-frame deletion), the mutant transcripts are not targeted by nonsense-mediated decay (NMD), as evidenced by the robust amplification of the 866 bp product. (g) Sashimi plot visualization of the RNA-seq data, illustrating exon 4 skipping. The absence of junctional reads between exons 3–4 and 4–5 in almost all the present reads, combined with the presence of direct reads between exons 3 and 5, confirms the aberrant splicing. (h) IGV tracks of the RNA-seq data, showing a nearly complete lack of coverage across exon 4, consistent with a homozygous exon skipping event in the patient.

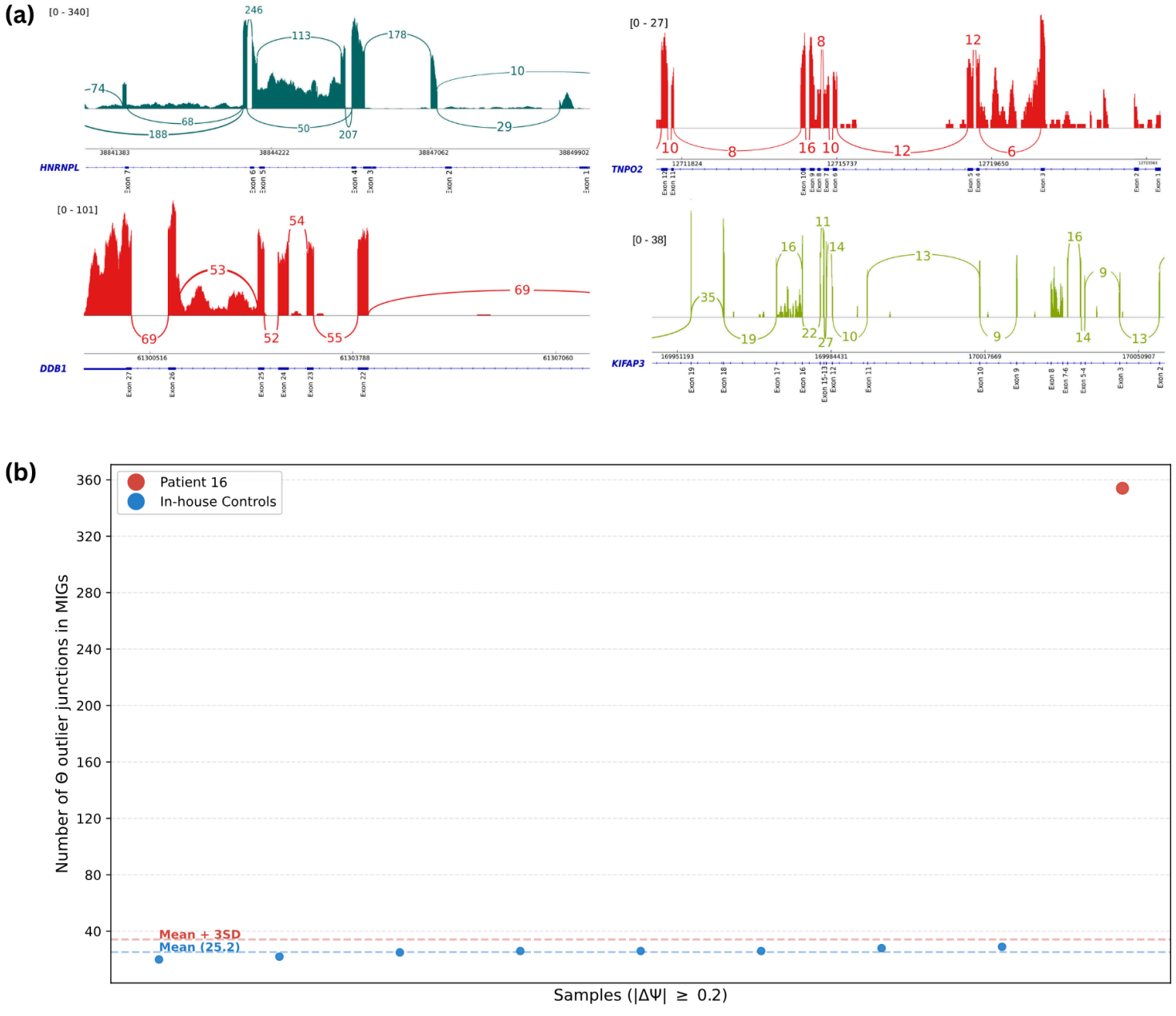


Figure 4

Transcriptome-wide assessment of global minor intron retention events in the RNU4ATAC-opathy patient. Scatter plot displaying the number of significant splicing efficiency (θ) outliers specifically within minor intron-containing genes (MIGs) per individual. The analysis was performed using the FRASER pipeline adapted from Arriaga et al. (2025). The proband (red dot), carrying compound heterozygous *RNU4ATAC* variants, exhibits a very high number of outliers (>400), indicating widespread intron retention. In contrast, the in-house control cohort (blue dots) (n=62) shows significantly lower outlier numbers (< +3 SD). Horizontal dashed lines represent the cohort mean, mean + 2 standard deviations (SD), and mean + 4 SD. Outliers were filtered based on an FDR-adjusted p-value < 0.05 and a PSI cutoff of ≥0.2. The proband's position above the +4 SD threshold confirms a severe, specific functional defect in the minor spliceosome complex.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SUPPLEMENTARYTABLES12.xlsx
- GenomicpaperTables2601TABLE1.docx
- GenomicpaperTables2601TABLE2.docx