Genome sequencing reveals the impact of non-canonical exon inclusions in rare genetic disease

Georgia Pitsava[1], Megan Hawley[2], Light Auriga[1], Ivan de Dios[1], Arthur Ko[3], Sofia Marmolejos[3], Miguel Almalvez[1], Ingrid Chen[2], Kaylee Scozzaro[2], Jianhua Zhao[2], Rebekah Barrick[6], Nicholas Ah Mew[3,4], Vincent A. Fusaro[1], Jonathan LoTempio[1], Matthew Taylor[7], Luisa Mestroni[7], Sharon Graw[7], Dianna Milewicz[8], Dongchuan Guo[8], David R. Murdock[8], Kinga M. Bujakowska[9], UCI-GREGoR Consortium, Changrui Xiao[5], Emmanuèle C. Délot[1], Seth I. Berger[3,4], Eric Vilain[1]

[1]Institute for Clinical and Translational Science, University of California, Irvine, CA, USA

[2]Labcorp Genetics Inc, Burlington, North Carolina, USA

[3]Center for Genetic Medicine Research, Children's National Research Institute, Washington, DC, USA

[4]Division of Genetics and Metabolism, Children's National Hospital, Washington, DC, USA

[5]Department of Neurology, University of California, Irvine, CA, USA

[6]Division of Metabolic Disorders, Children's Hospital of Orange County (CHOC), Orange, CA 92868, USA

[7]Cardiovascular Institute and Adult Medical Genetics Program, University of Colorado Anschutz Medical Campus, Aurora, CO, USA

[8]Division of Medical Genetics, University of Texas Health Science Center at Houston (UTHealth) McGovern Medical School, Houston, Texas, USA

[9]Ocular Genomics Institute, Department of Ophthalmology, Massachusetts Eye and Ear Infirmary, Harvard Medical School, Boston, MA, USA

Correspondence: sberger@childrensnational.org

Abstract

Introduction

Advancements in sequencing technologies have significantly improved clinical genetic testing, yet the diagnostic yield remains around 30-40%. Emerging sequencing technologies are now being deployed in the clinical setting to address the remaining diagnostic gap.

Methods

We tested whether short-read genome sequencing could increase diagnostic yield in individuals enrolled into the UCI-GREGoR research study, who had suspected Mendelian conditions and prior inconclusive clinical genetic testing. Two other collaborative research cohorts, focused on aortopathy and dilated cardiomyopathy, consisted of individuals who were undiagnosed but had not undergone harmonized prior testing.

Results

We sequenced 353 families (754 participants) and found a molecular diagnosis in 54 (15.3%) of them. Of these diagnoses, 55.5% were previously missed because the causative variants were in regions not interrogated by the original testing. In 9 cases, they were deep intronic variants, 5 of which led to abnormal splicing and cryptic exon inclusion, as directly shown by RNA sequencing. All 5 of these variants had inconclusive spliceAI scores. In 26% of newly diagnosed cases, the causal variant could have been detected by exome sequencing reanalysis.

Conclusion

Genome sequencing overcomes multiple limitations of clinical genetic testing, such as inability to call intronic variants and technical limitations. Our findings highlight cryptic exon inclusion as a common mechanism via which deep intronic variants cause Mendelian disease. However, they also reinforce that reanalysis of exome datasets can be a fruitful approach.

Keywords: genome sequencing, RNA sequencing, rare disease, intronic variants, cryptic exon

Introduction

The widespread application of clinical genetic testing has tremendously increased the rate at which rare genetic diseases are being diagnosed and new gene-disease associations are identified. Despite this progress, the diagnostic yield of clinical genetic testing remains 30-40% [1, 2]. There are several reasons for this plateau. For many conditions, the causal genes are still unidentified. Even for conditions with known causal genes, interpreting the clinical significance of variants can be challenging. This is because there are a large number of variants with insufficient information pertinent to their pathogenicity, known as variants of unknown significance (VUS). In addition, technical limitations of commonly offered clinical tests can result in missed variants, for example in regions that are not represented in exome sequencing such as deep intronic and untranslated regions. Epigenetic factors that affect gene expression but are not interrogated by standard testing potentially add to the complexity.

Molecular sequencing technologies not traditionally used in the clinical setting are now being deployed in order to overcome these challenges. These include the increasing use of genome sequencing (GS) as well as RNA sequencing. The latter can be especially useful for directly evaluating the functional consequences of intronic variants and variants in splice regions, since despite advances in computational predictors of splicing disruption (such as SpliceAI and Pangolin [3, 4]) it often remains difficult to determine if these variants are pathogenic or not.

The University of California, Irvine (UCI) - GREGoR site, part of the GREGoR (Genomics Research to Elucidate the Genetics of Rare Diseases) Consortium, has been evaluating the diagnostic utility of these newer approaches. Our study has been enrolling participants whose clinical presentation strongly suggests a genetic cause, but who remain without a molecular diagnosis, despite having undergone conventional genetic testing (gene panels, chromosomal microarray, exome sequencing - ES).

Here, we describe our center's experience from short-read GS of a cohort of 353 such families (754 participants). To help disambiguate previously or newly detected variants (VUS or single heterozygous pathogenic variants in recessive disease genes), we deployed additional methods, including RNA sequencing. We also provide an overview of our cohort and focus on lessons learned from informative, successfully solved cases.

Methods

Participants

The Mendelian Genomics Research Center was launched in 2021, between Children's National Hospital, Invitae, and more recently UCI. The study was approved by the Institutional Review Board (IRB) of Children's National Hospital (Pro00015852). Study participants included probands with a suspected Mendelian condition and prior non-diagnostic clinical testing, such as gene panels, chromosomal microarray, or ES. Testing on additional family members was performed when available. Individuals were enrolled over a three-year period, from July 2021 to April 2024, through referrals by physicians and collaborators, or self-referral via our website

(https://research.childrensnational.org/labs/pediatric-mendelian-genomics-research-center). Written informed consent was available in both English and Spanish and was obtained for all families by study staff or through a study-specific automated chatbot [5]. Workflows for remote enrollment and sample collection were established, eliminating the need for participants to travel. Following consent, buccal swabs and/or peripheral blood samples were collected from probands and their relevant available family members. Some participants provided clinically generated genome data for reanalysis [6]. In addition, our study population involved two collaborative cohorts, one focusing on aortopathy and another on dilated cardiomyopathy; for these their prior genetic testing may not have been as comprehensive or well documented. Phenotypic categories are described in **Table 1**.

Short-read Genome Sequencing

Blood and buccal samples were sent to Invitae® laboratories where they were processed and underwent genome sequencing using Illumina NovaSeq 6000. Sequencing files were subsequently aligned to the hg38 human genome assembly with bwa-mem using soft clipping enabled for supplementary alignments (-Y), processed at 120,000,000 input bases per batch for reproducibility (-K) [7] and genotyped using Google DeepVariant [8]. Structural variant calling was performed with Illumina Manta [9]. Alignments and variant call files are available on the NHGRI Analysis Visualization and Informatics Lab-space (AnVIL; AnVIL Portal) on the GREGoR Consortium workspace.

Short-read RNA sequencing

Blood samples were preserved in PAXgene tubes processed with the WatchMaker Genomics RNA library prep kit with the Polaris depletion module to create stranded and ribosomal/globin-depleted libraries. They then underwent RNA sequencing using the Illumina NovaSeq 6000 S4 flowcells generating 100-200 million 150-bp read pairs. Reads were aligned to hg38 using STAR (v2.7.10a) [10] with the GENCODEv41 annotation. We employed the basic two-pass mode and allowed up to 3 mismatches and a minimum aligned length of 100 bp. Alignments were visualized using integrative genomics viewer (IGV) [11-14]. Alignments are available on AnVIL (AnVIL Portal).

Mini gene assays

A minigene splicing assay was performed using a mini-gene split Green Fluorescent Protein (GFP) construct [15], in which N- and C-terminal parts of the *GFP* gene were separated by *SMN1* introns 7 and 8 (NM_000344). Reference and mutated gene fragments containing 1000 bp of *DBT* intron 10 surrounding the c.1282-4218G>A variant flanked with 30 bp vector homology arms were synthetized (TWIST Bioscience, USA) and cloned into the mini-gene construct (Gibson Assembly Master Mix, New England Biolabs). After Sanger sequencing verification of all constructs, they were transfected into HEK293 cells (Lipofectamine 3000, Thermo Fisher Scientific). Forty-eight hours post transfection, total RNA was extracted from the transfected cells (RNAeasy Mini Kit, Qiagen) and cDNA was generated using random hexamer primers (SuperScript IV Synthesis Kit, Thermo Fisher Scientific). Subsequently, the minigene transcripts were amplified from the cDNA using primers specific to the split *GFP* fragments (F: 5'-CACACTGGTGACAACATTTACATAC-3'; R: 5'- GAAATCGTGCTGTTTCATGTGATC-3').

The PCR products were column-purified (DNA Clean & Concentrator-5, Zymo Research) and analyzed with next-generation amplicon sequencing (MiSeq, Illumina, OGI Genomics Core). The splicing pattern analysis was performed by aligning the sequence reads to the hybrid reference of the split *GFP* construct containing the *DBT* intron 10 (STAR Aligner) [16] and visualizing the reads in (IGV) [11-14]. The same process was followed to test the effect of the *HFE* VUS found in participant PMGRC-124-124-0.

Variant prioritization and analysis

Phenotypic data for probands was collected via a manual review of electronic medical records or provided by collaborating researchers. The phenotypes of each proband were converted to Human Phenotype Ontology (HPO) terms.

Variants were reviewed by expert curators using the MOON™ software. MOON implements a natural language processing approach to prioritize variants in genes that are likely to be related to an individual's phenotype based on published literature [17, 18]. Cases were reviewed with the goal of identifying pathogenic, likely pathogenic, and VUS in known disease genes consistent with the phenotype of the proband. In addition to sequence and structural variants prioritized by MOON, rare sequence variants in the following categories were reviewed: those expected to result in loss of function, *de novo* variants, compound heterozygous or homozygous variants, hemizygous variants, variants classified as likely pathogenic or pathogenic in ClinVar, and variants in genes associated with conditions with a high clinical overlap for the patient.

Cases were considered "solved" when a disease-causing variant or multiple disease-causing variants in a gene associated with a condition consistent with the phenotype of the proband were detected. The number and phase of variants needed to be consistent with the inheritance pattern of the condition. For example, a single likely pathogenic or pathogenic variant in an autosomal dominant condition or biallelic likely pathogenic or pathogenic variants in an autosomal recessive condition were needed for a case to be considered "solved". Cases were considered "probably solved" when one or more of the variants were classified as a VUS but there was a strong phenotypic overlap with the clinical presentation of the patient. Cases were considered "partially solved" if likely pathogenic or pathogenic variants consistent with the inheritance pattern of the condition were detected, but the result only explained part of the proband's phenotype.

Finally, we note that analysis is ongoing; the data reported on this manuscript are as of October 1, 2024.

Results

Genome Sequencing leads to a molecular diagnosis in 15.3% of previously unsolved cases

A total of 353 families underwent GS during the designated period; demographic information is shown in **Figures 1a and 1b**. Our study included several different family structures; 41.9% (n=148) were proband-only cases, 13.3% (n=47) were duos or duos plus another family member (duo+), 43.3% (n =153) were trios or trios with another family member (trio+) and 1.4% (n=5)

consisted of other family combinations (**Figure 2a**). This accounts for a total of 754 genomes analyzed.

A genetic diagnosis was found in 54 (15.3%) cases (solved/probably solved; Supplemental Table 1). As expected, the diagnostic yield was higher in trios/trios+ compared to duos/duos + and proband-only families (21% vs 13% and 10%; **Figure 2b**). This difference was statistically significant ($p = 0.03$, chi-square test).

In addition, 6 cases (1.7%) were partially solved; that is, the identified variant could explain part, but not the entirety, of the individual's phenotype. Furthermore, we found a candidate variant in approximately 9% of the cases that remained unsolved (27 out of the 299; Supplemental Table 2).

To better understand why the diagnosis was previously missed in the 54 solved/probably solved cases, we examined the type of prior testing these participants had undergone (**Figure 3a**). In 26% (14 out of 54) of solved cases, the causal variant had been identified by initial testing but was not reported either because the gene-phenotype association was unknown at the time of that testing or because the variant was not classified as pathogenic at the time (**Figure 3b**). Below, we focus on cases where initial testing did not interrogate the genomic region harboring the causative variant.

Diagnoses in uninterrogated regions of original testing highlight cryptic exon inclusion as a common mechanism by which deep intronic variants cause disease

In the majority of solved/probably solved cases (55.5%; 30 out of 54 cases), the diagnosis was missed before because the original testing method was unable to call variants in the region harboring the causative variant. For example, there were 13 cases where panel testing did not include the causal gene, and 13 cases where the causative variant was non-coding. In one case, a synonymous coding-region variant was predicted by spliceAI to lead to intron retention in *HFE*. This intron retention event leading to a stop-gain was confirmed using a mini gene assay (Supplemental Figure 1).

With respect to non-coding causative variants, these were deep intronic in 9 cases. In five of those, the causative variant altered the canonical splicing pattern via the inclusion of a cryptic exon (**Table 2**). Variant pathogenicity was revealed by RNAseq, which directly demonstrated the effect at the transcript level. For example, one case was due to an intronic variant in *PEX1*, in a patient with a known biochemical diagnosis of Zellweger Spectrum Disorder and a carrier of a heterozygous known pathogenic variant in *PEX1*. The abnormal cryptic exon inclusion was revealed by RNAseq (Supplemental Figure 2). Another example was a newborn female who was biochemically diagnosed with thiamine-responsive maple syrup urine disease (MSUD) based on the same diagnosis in her sister. However, all prior genetic testing (panel testing of 9 genes, *ASL* single-gene testing, and ES) was negative. GS revealed a homozygous deep (-4230) intronic single nucleotide variant predicted to lead to aberrant splicing of *DBT*, a gene known to cause MSUD type II (OMIM# 620699), which can sometimes be thiamine-responsive. RNAseq confirmed the cryptic exon inclusion, establishing the molecular diagnosis (**Figure 4a**). This variant was also confirmed by minigene assay (**Figure 4b-e**). In another case we found a deep intronic variant in a case of Coffin-Siris, which we validated by episignature in long-read sequencing [19].

We found that, in all five cases with a cryptic exon inclusion, both the acceptor gain and the donor gain spliceAI scores fell within a range generally considered inconclusive with regards to impact on splicing (range 0.16 - 0.43 for acceptor gain scores; 0.19 - 0.55 for donor gain scores). This observation, while preliminary, suggests that the combination of such acceptor gain and donor gain scores may be a marker for variants leading to this splicing alteration. Of note, commonly used pre-calculated spliceAI scores limit their search to 50 bp from the variant, and therefore in each of these examples the new splice acceptor would not have been detected with pre-calculated spliceAI scores alone. It required a direct calculation of a spliceAI score with a search of at least 500 bp from the variant. Pangolin scores were generally higher than spliceAI scores.

Four additional cases harbored variants in the spliceosomal non-coding RNA *RNU4-2* and contributed to the discovery of ReNU syndrome (OMIM#620851) [20]. Variants in this gene represent a new, surprisingly common etiology of neurodevelopmental disorders, explaining approximately 0.4% of such cases.

Finally, an informative example in which the causal variant was missed due to variant filtering parameters driven by limitations of ES was a woman in her 30s with glomuvenous malformations, conductive hearing loss, and midline malformations. Previous testing included a cancer 38-gene panel, *GLA* single-gene testing, ES, and mitochondrial testing; all were negative. GS revealed a known pathogenic exonic frameshift variant in *GLMN* (Glomuvenous Malformation Syndrome; OMIM#138000). Upon review of the previous clinical ES, we discovered that this same variant was detectable in the original data. However, the variant had not been reported as the exon was excluded from the lab's clinically validated regions due to a propensity for sequencing artifacts in this exon. Clinicians ordering clinical exome sequencing may not be aware of the coverage limitations of clinically relevant genes or exons important for the testing indication.

Detection of structural variants

We also identified structural variants not detected by previous testing. This includes a complex rearrangement of *OCA2* with deep intronic breakpoints [6] and deletion of the first exon of *CREBBP* in a child with features of Rubinstein-Taybi syndrome 1 (OMIM#180849) and previous negative ES. It is important to note that while ES has made strides in detection of structural variants, copy-neutral variants and single-exon deletions may still not be detected.

Syndromic phenotypes had the highest solve rate

When examining the solve rate separately for each phenotypic category, the largest number of diagnoses were obtained in individuals with syndromic phenotypes (32 solved cases out of 128 total cases; 25%), while the solve rate in non-syndromic cases was 9.8% (**Figure 5A**). When further stratifying non-syndromic cases according to the affected organ system (**Figure 5B**), individuals whose phenotype was categorized as 'Cardiovascular' were the most likely to receive a diagnosis in our entire cohort (7 out of 19, 36.8 %). All but one of these individuals who received a diagnosis were affected by dilated cardiomyopathy and were part of a legacy cohort. The genes we found were not known to be associated with dilated cardiomyopathy at the time of testing, which explains why they were initially missed.

Discussion

Our study provides a characterization of the potential of short-read GS, combined with other confirmatory analyses such as RNAseq, to yield genetic diagnoses for previously undiagnosed cases. GS provides a molecular diagnostic advantage over ES in that it can detect variants in regions not interrogated by ES such as introns or non-coding genes, while RNAseq can assist in establishing the pathogenicity of intronic splice-altering variants by directly demonstrating the impact on the RNA product. Here, using a combination of GS and RNAseq we discovered cryptic exon inclusion as a common (among intronic variants) molecular mechanism with a loss-of-function impact via perturbed splicing.

Regarding causative variants in non-protein coding genes, while some examples of disease-causing non-coding RNAs were previously known, this number is now expanding thanks to the implementation of GS. Recently discovered examples include the aforementioned *RNU4-2*, encoding the U4 small nuclear RNA (snRNA) component of the major spliceosome [20, 21], and the long non-coding RNA *CHASERR,* which was shown to cause a novel syndromic neurodevelopmental disorder [22].

Going forward, we envision that improvements in our ability to evaluate the impact of variants in regions traditionally missed by conventional genetic testing will continue to increase the diagnostic yield of GS. It is important to recognize, however, that in some cases it may be cost-effective to reanalyze ES data instead of resorting to GS, especially given advances in variant interpretation driven by better predictive models of variant effects and an expanded understanding of tolerated genetic variation. Indeed, many causal variants in our study could have been detected by a reanalysis of ES data, using new information provided by variant reclassification or improved analytic pipelines; a finding that is in alignment with results from previous studies [23-25]. In addition, testing of other family members can also aid in variant reinterpretation by showing if a variant is *de novo* or if it segregates with the phenotype in the family. Finally, we speculate that newer long-read sequencing technologies will help further close the diagnostic gap by capturing variants not captured by short-read technologies such as certain structural variants and repeat expansions [26].

Data Availability

Data available in the GREGoR workspace in AnVIL (AnVIL Portal).

Funding Statement

The study was supported by the National Institutes of Health grant U01HG011745, as part of the GREGoR Consortium. KMB was supported by the GREGoR Consortium Research Grant from

the GREGoR Data Coordinating Center [U24HG011746]; National Eye Institute [R01EY035717 (KMB) and P30EY014104 (MEEI core support)], Iraty Award 2023 (KMB), Lions Foundation (KMB) and the Research to Prevent Blindness Unrestricted Grant (KMB).

Disclosures/Conflict of interest

M.H, J.Z. and K.S. are currently employees of Labcorp Genetics Inc, formerly known as Invitae Corp, I.C. is a former employee of Invitae Corp. All other authors declare no conflicts of interest.

Acknowledgements

We thank the participants and referring physicians for participating in this study.

Ethics Declaration

This study was approved by the Children's National Hospital Institutional Review Board (IRB) under protocol Pro00015852. Informed consent was obtained from all participants as required by the IRB.

Table 1. Definitions of Phenotypic Categories

Multisystem Syndromic Disorders	Proband presenting with symptoms affecting multiple organ systems, often accompanied by distinctive facial features.
Non-syndromic Disorders	Proband with conditions that predominantly affect a single organ system, without significant involvement of other systems.
Subcategories	
Neurodevelopmental Disorders	Probands exhibiting conditions that primarily affect the development and function of the nervous system, including the brain, spinal cord, and peripheral nerves. This category encompasses disorders impacting cognitive function, behavior, motor skills, and neurological processes.
Cardiovascular Disorders	Proband with primary involvement of the heart and blood vessels, such as structural heart defects, cardiomyopathies, arrhythmias, and vascular diseases.
Connective Tissue & Skeletal Disorders	Proband with symptoms affecting bones, joints, and connective tissues such as skin, tendons, and ligaments.
Other	Proband who do not clearly fit any of the aforementioned categories.

337 Table 2. Participants with cryptic exon inclusion.

Participant ID	Variant	Gene	SpliceAI score		Distance to variant		Pangolin score (Gain)
			Acceptor	Donor	New Acceptor	New Donor	
PMGRC-146-146-0	c.3235+700C>G	*ARID1B*	0.31	0.19	-143 bp	-5 bp	0.54
PMGRC-220-220-0	c.1390-515_1390-499del	*HADHB*	0.27	0.27	-85 bp	27 bp	0.35
PMGRC-332-332-0	c.789+973C>G	*CRPPA*	0.43	0.55	118 bp	5 bp	0.46
PMGRC-403-403-0	c.1359+601A>G	*PEX1*	0.30	0.54	168 bp	0 bp	0.36
PMGRC-658-658-0	c.1282-4218G>A	*DBT*	0.16	0.20	161 bp	3 bp	0.66

338 SpliceAI scores were calculated using the spliceAI lookup tool with a maximum distance of 500 bp.

339

340

341

342

343

344

345

346

347

348

349

350

351

352

353

354

355

356

357

358 Supplemental Table 1. Solved/Probable solved cases

Participant ID	Phenotypic Category	Gene	Variant	Patient Phenotype	Reason diagnosis was missed on previous testing	Associated Disorder (OMIM)
PMGRC-13-13-0	Syndromic	*PTHLH*	c.4C>T, p.Gln2*	Pseudohypoparathyroidism, Chiari 1 malformation, shortened metacarpals	Limited Original Testing	Brachydactyly, type E2 (#613382)
PMGRC-32-32-0	Other	*FLG*	c.1297_1298del, p.Asp433fs c.2282_2285del, p.Ser761fs	Eczematoid dermatitis, numerous arcuate plaques with erythematous border, perivascular and interstitial dermatitis with neutrophils on biopsy, ichthyosiform xerosis	Limited Original Testing	Ichthyosis vulgaris (#146700)
PMGRC-43-43-0	Syndromic	*HECTD1*	c.2082C>G, p.Phe694Leu	Hypotonia, oral dysphagia, macrocephaly, autism spectrum disorder, somatic overgrowth syndrome, hypothalamic obesity, tube feeding, esotropia, stereotypic movements, abnormal brain MRI, dysmorphic facial features	New Information/Reclassification/Candidate gene now linked to syndrome	HECTD1-neurodevelopmental disorder (under review)
PMGRC-46-46-0	Syndromic	*PPP1R3F*	c.554_555del, p.His185fs*66	Hypotonia, tube feeding, autism spectrum disorder, speech/language delays, developmental delay	New Information/Reclassification/Candidate gene now linked to syndrome	X-linked PPP1R3F neurodevelopmental disorder [27]
PMGRC-88-88-0	Syndromic	*NARS2*	c.749G>A, p.Arg250Gln	Critically ill neonate with multisystemic disease. A few months later, new findings of interstitial lung disease and refractory seizures	Limited Original Testing	Combined oxidative phosphorylation deficiency 24 (#616239)
PMGRC-101-101-0	Syndromic	*CBL*	c.1259G>A, p.Arg420Gln	Hemihypertrophy, nystagmus, club foot	Previously known	Noonan syndrome-like disorder with or without juvenile myelomonocytic leukemia (#613563)
PMGRC-114-114-0	Connective tissue & Musculoskeletal disorders	*ACAN*	c.4753del, p.Asp1585fs c.757+4A>G	Severe scoliosis, spondyloepimetaphyseal dysplasia, genu valgum, short stature	Previously known	Spondyloepimetaphyseal dysplasia, aggrecan type (#612813)

PMGRC-124-124-0	Syndromic	*HFE*	c.187C>G, p.His63Asp c.892G>A, p.Glu298Lys	Liver disease, hyperammonemia, encephalopathy, hyperbilirubinemia	Limited Original Testing	Hemochromatosis, type 1 (#235200)
PMGRC-146-146-0	Syndromic	*ARID1B*	c.3235+700C>G	Developmental delay, gross motor delay, abnormal brain MRI, mixed receptive-expressive language disorder, dysmorphic facial features	Limited Original Testing	Coffin-Siris syndrome 1 (#135900)
PMGRC-148-148-0	Syndromic	*RNU4-2*	n.64_65insT	Hypotonia, global developmental delay, speech delay, febrile seizures, optic nerve hypoplasia, tube feeding, dysmorphic facial features	Limited Original Testing	ReNU syndrome (#620851)
PMGRC-151-151-0	Syndromic	*GLMN*	c.157_161del, p.Glu52_Lys53insTer	Clinical diagnosis of glomuvenous malformations	Limited Original Testing	Glomuvenous malformations (#138000)
PMGRC-158-158-0	Syndromic	*COQ2*	c.421G>C, p.Val141Leu c.138dup, p.Ala47fs	Retinal atrophy, horseshoe kidney, nephrotic syndrome	New Information/Reclassification	Coenzyme Q10 deficiency, primary, 1 (#607426)
PMGRC-170-170-0	Neurodevelopmental	*GRIA4*	c.1643G>C, p.Trp548Ser	Hypotonia, feeding difficulties, global developmental delays of motor and language skills, microcephaly, delayed visual attention, abnormal brain MRI	New Information/Reclassification	Neurodevelopmental disorder with or without seizures and gait abnormalities (#617864)
PMGRC-175-175-0	Syndromic	*PTPN11*	c.1510A>G, p.Met504Val	Pulmonary valve stenosis, toe syndactyly, widely spaced nipples	Previously Known	Noonan syndrome 1 (#163950)
PMGRC-176-176-0	Syndromic	*SCN8A*	c.2549G>A, p.Arg850Gln	Refractory seizures, epileptic encephalopathy, infantile spasms, global developmental delay, hypotonia, tube feeding, respiratory insufficiency, hypotonia	Previously Known	Developmental and epileptic encephalopathy 13 (#614558)
PMGRC-178-178-0	Neurodevelopmental	*CHD1*	c.1010C>T, p.Thr337Ile	Motor delay, hypotonia, stereotypic behaviors, autism spectrum disorder	New Information/Reclassification	Pilarowski-Bjornsson syndrome (#617682)

PMGRC-204-204-0	Neurodevelopmental	*FBXO31*	c.1000G>A, p.Asp334Asn	Abnormal brain MRI, gross motor delay, speech delay, hypertonia, lower extremity spasticity, esotropia, mixed receptive-expressive language disorder	New Information/Reclassification	FBXO31-related spastic-dystonic cerebral palsy syndrome [28]
PMGRC-205-205-0	Syndromic	*SLC6A8*	c.1255-3_1255-2delCA	Acute liver failure, acute kidney injury, autism spectrum disorder	Limited Original Testing	Cerebral creatine deficiency syndrome 1 (#300352)
PMGRC-212-212-0	Other	*OCA2*	c.1465A>G, p.Asn489Asp and a complex rearrangement with deep intronic breakpoints	Oculocutaneous albinism	Limited Original Testing	Oculocutaneous albinism, type II (#203200)
PMGRC-220-220-0	Syndromic	*HADHB*	c.1390-515_1390-499del	Hypoparathyroidism, LCHAD, mitochondrial trifunctional protein deficiency, pancytopenia, bone marrow deficiency, nephrotic syndrome, tube feeding	Limited Original Testing	Mitochondrial trifunctional protein deficiency 2 (#620300) [29]
PMGRC-265-265-0	Syndromic	*RNF220*	c.1094G>A, p.Arg365Gln	Bilateral sensorineural hearing loss, leukodystrophy, seizures, mixed receptive-expressive language disorder, photoreceptor degeneration, intention tremor	New Info/Reclassification	Leukodystrophy, hypomyelinating, 23, with ataxia, deafness, liver dysfunction, and dilated cardiomyopathy (#619688)
PMGRC-316-316-0	Syndromic	*PTPN11*	c.794G>A, p.Arg265Gln	Short stature, gross motor delays	Limited Original Testing	Noonan syndrome 1 (#163950)
PMGRC-332-332-0	Neurodevelopmental	*CRPPA*	c.789+973C>G	Global developmental delay, hypotonia, chronic lung disease, retinopathy, severe optic atrophy, tube feeding, abnormal brain MRI	Limited Original Testing	Muscular dystrophy-dystroglycanopathy (congenital with brain and eye anomalies), type A, 7 (#614643)
PMGRC-355-355-0	Neurodevelopmental	*TCF7L2*	c.1156C>T, p.Arg240Cys	Walking difficulties, unsteady gait, lateral lisps, ADHD, distal weakness, hammertoes, gynecomastia, learning disabilities	New Information/Reclassification	TCF7L2-associated syndromic neurodevelopmental disorder [30]

PMGRC-366-366-0	Neurodevelopmental	*DDX17*	c.1077G>A, p.Trp359*	Hypotonia, global developmental delay, gross motor delay, mixed receptive-expressive language disorder, foot drop, steppage gait, pes planus, ataxic gait	New Information/Reclassification	DDX17- associated neurodevelopmental disorder [31]
PMGRC-388-388-0	Syndromic	*RNU4-2*	n.64_65insT	Global developmental delay, hemiparesis, microcephaly, tube feeding, growth delays, truncal hypotonia, autism spectrum disorder, nonverbal, speech apraxia, sensory processing difficulties, abnormal brain MRI	Limited Original Testing	ReNU syndrome (#620851)
PMGRC-392-392-0	Neurodevelopmental	*COQ4*	c.718C>T, p.Arg240Cys c.202+4A>C	Spastic diplegia, gross motor delays, speech articulation delay, unsteady gait, walking difficulties	New Information/Reclassification	Spastic ataxia 10, (#620666)
PMGRC-403-403-0	Syndromic	*PEX1*	c.1359+601A>G	Zellweger Spectrum Disorder, elevated very long-chain fatty acids, decreased plasmalogen	Limited Original Testing	Peroxisome biogenesis disorder 1A (#214100)
PMGRC-418-418-0	Neurodevelopmental	*ZIC2*	c.1030_1032del, p.Trp359*	Alobar holoprosencephaly	Limited Original Testing	Holoprosencephaly 5 (#609637)
PMGRC-437-437-0	Syndromic	*GLI2*	c.1111_1120del, p.Ile371fs	Septopreoptic holoprosencephaly, solitary median maxillary incisor syndrome, low thyroid, underdeveloped pituitary, sensory processing disorder, failure to thrive	Limited Original Testing	Holoprosencephaly 9 (#610829)
PMGRC-440-440-0	Syndromic	*ZIC2*	c.1013del, p.Pro338fs	Deafness, cerebral palsy, holoprosencephaly	Previous testing information not provided	Holoprosencephaly 5 (#609637)
PMGRC-451-451-0	Syndromic	*ZIC2*	c.1377_1406dup, p.Ala461_Ala470dup	Semilobar holoprosencephaly, diabetes insipidus, delayed development	Limited Original Testing	Holoprosencephaly 5 (#609637)
PMGRC-509-509-0	Syndromic	*RNU4-2*	n.64_65insT	Microcephaly, hypotonia, delayed myelination, gross motor delay, brachycephaly, abnormal brain MRI, global developmental delay	Limited Original Testing	ReNU syndrome (#620851)
PMGRC-538-538-0	Syndromic	*AGO2*	c.602G>T, p.Gly201Val	Hip dysplasia, congenital hypotonia, cleft palate, global developmental delay, hearing loss, Pierre Robin sequence, seizures, speech delay, cerebral palsy,	New Information/Reclassification	Lessel-Kreienkamp syndrome (#619149)

				scoliosis, congenital heart defects, dysmorphic facial features		
PMGRC-540-540-0	Syndromic	*KMT2C*	c.5668C>T, p.Arg1890*	Hypotonia, fine motor delay, gross motor delay, sensory processing disorder, speech delay, decreased oral tone, selective mutism, ADHD, developmental coordination disorder, dysmorphic facial features	Limited Original Testing	Kleefstra syndrome 2 (#617768)
PMGRC-569-569-0	Neurodevelopmental	*STAG2*	87kb deletion of chrX:123,946,386-124,003590 Exon 1-4 deletion	Holoprosencephaly, global developmental delay, scoliosis, apraxia of speech, vision problems, vertebral anomaly	Previously Known	Holoprosencephaly-13, X-linked (#301043)
PMGRC-599-599-0	Neurodevelopmental	*RNU4-2*	n.69C>T	Small for gestational age, abnormal visual development/eye movement irregularity, global developmental delay, hypotonia, abnormal brain MRI	Limited Original Testing	ReNU syndrome (#620851)
PMGRC-635-635-0	Syndromic	*CDK13*	c.425dup, p.Leu143fs	Congenital heart disease, dysmorphic facial features, short stature, ADHD, global developmental delay	New Information/Reclassification	Congenital heart defects, dysmorphic facial features, and intellectual developmental disorder (#617360)
PMGRC-658-658-0	Syndromic	*DBT*	c.1282-4218G>A	Prenatal diagnosis of Argininosuccinic acid lyase deficiency and thiamine-responsive maple syrup urine disease	Limited Original Testing	Argininosuccinic aciduria, Maple syrup urine disease, type II (#620699)
PMGRC-677-677-0	Syndromic	*TPM3*	c.835C>G, p.Leu279Val	Pierre Robin sequence, tube feeding, gastroparesis, mannose-binding lectin deficiency, recurrent aspiration pneumonia and sinusitis, abnormal brain MRI, hypotonia	Limited Original Testing	Congenital Myopathy 4A, Autosomal Dominant (#255310)
PMGRC-700-700-0	Syndromic	*OFD1*	c.2413dup, p.Gln805fs	Semilobar holoprosencephaly, cleft lip and palate, bilateral postaxial polydactyly of hands and feet, bilateral cryptorchidism, Dandy-Walker malformation, ventriculomegaly, hypotelorism, seizures	Limited Original Testing	Joubert syndrome 10 (#300804)

PMGRC-753-753-0	Syndromic	*MSL2*	c.535G>T, p.Glu179*	Gross motor delay, speech delay, hypotonia, microcephaly, frontal bossing, fine hair, ptosis, pectus excavatum, low muscle tone, Mongolian spot in lower back, global developmental delay	New Information/Reclassification	Karayol-Borroto-Haghshenas neurodevelopmental syndrome (#620985)
PMGRC-812-812-0	Syndromic	*NDST1*	c.1850 C>T, p.Thr617Ile	Ankyloglossia, gross motor delays, thoracolumbar kyphosis, global developmental delay, axial hypotonia	New Information/Reclassification	Intellectual Developmental Disorder, Autosomal Recessive 46 (#616116)
PMGRC-817-817-0	Syndromic	*MT-TL1*	n.14A>G (m.3243A>G)	Fatigue, headache, recurrent fever, tinnitus, recurrent episodes of infection, sudden loss of hearing	Limited Original Testing	Diabetes-deafness syndrome (#520000)
PMGRC-890-890-0	Cardiovascular	*TNN*	c.59604_59607delAAAG p.Gly19871fs	Dilated Cardiomyopathy	Limited Original Testing	Dilated Cardiomyopathy, 1G (#604145)
PMGRC-897-897-0	Cardiovascular	*SCN5A*	c.3911C>T, p.Thr1304Met	Dilated Cardiomyopathy	Previously Known	Dilated Cardiomyopathy, 1E (#601154)
PMGRC-904-904-0	Cardiovascular	*DMD*	c.9851G>A, p.Trp3284*	Dilated Cardiomyopathy	Previously Known	Dilated Cardiomyopathy, 3B (#302045)
PMGRC-911-911-0	Cardiovascular	*MYH6*	c.2489C>T, p.Pro830Leu	Dilated Cardiomyopathy	Limited Original Testing	Dilated Cardiomyopathy, 1EE (#613252)
PMGRC-914-914-0	Cardiovascular	*TNNT2*	c.517C>T, p.Arg173Trp	Dilated Cardiomyopathy	Limited Original Testing	Dilated Cardiomyopathy, 1D (#601494)
PMGRC-917-917-0	Cardiovascular	*RBM20*	c.2497dupA, p.Arg833fs	Dilated Cardiomyopathy	Limited Original Testing	Dilated Cardiomyopathy, 1DD (#613172)
PMGRC-921-921-0	Cardiovascular	*TNNI3K*	c.2302G>A, p.Glu768Lys	Dilated Cardiomyopathy	Limited Original Testing	Cardiac conduction disease with or without dilated cardiomyopathy (#616117)

PMGRC-991-991-0	Connective tissue & Musculoskeletal disorders	*TGFBR1*	c.1355del, p.Pro452fs	Thoracic aortic aneurysm, Aortic dissection	Previous testing information not provided	Loeys-Dietz syndrome 1 (#609192)
PMGRC-1040-1040-0	Connective tissue & Musculoskeletal disorders	*FBN1*	c.4096G>A, p.Glu1366Lys	Thoracic aortic aneurysm, Aortic dissection	Previous testing information not provided	Marfan syndrome (#154700)
PMGRC-1091-1091-0	Syndromic	*CREBBP*	exon 2 deletion	Failure to thrive, global developmental delay, pectus excavatum, axial hypotonia, abnormal brain MRI, dysmorphic features	Limited Original Testing	Rubinstein-Taybi syndrome 1 (#180849)

ADHD attention deficit hyperactivity disorder, *LCHAD* long-chain 3-hydroxyacyl-CoA dehydrogenase enzyme deficiency, *MRI* magnetic resonance imaging

Supplemental Table 2. Candidate genes identified in unsolved cases. All genes are deposited in GeneMatcher.

Participant ID	**Candidate Gene**	**Variant**
PMGRC-86-86-0	*NRXN2*	c.2107G>A, p.Gly703Ser
PMGRC-104-104-0	*GALNT16*	c.502+1G>A
PMGRC-116-116-0	*ARHGAP21; PDPK1*	c.3692_3693del; c.1588G>A, p.Gly530Arg
PMGRC-137-137-0	*CSMD3*	c.3158dup, p.Arg1530*
PMGRC-217-217-0	*SLC16A13; HIVEP1; MHY7*	c.410G>A, p.Arg137Gln, c.395A>G, p.Tyr132Cys; c.6212dupA, p.Tyr2071*; c.4187G>T, p.Arg1396Leu
PMGRC-239-239-0	*FGD5*	c.3372_3373del, p.Gly1125Glufs*54
PMGRC-255-255-0	*NRXN2; STRN3*	c.2605C>T, p.Arg869Trp; c.1875C>A, p.Tyr625*
PMGRC-279-279-0	*VWA3B*	c.59G>T, p.Gly20Val c.1737+1G>A
PMGRC-291-291-0	*NSL1*	c.314-2_315delAGAT
PMGRC-312-312-0	*TJP1*	c.3927-2_3927-1del
PMGRC-363-363-0	*CEP350*	c.2258del, p.Gly753Glufs*11
PMGRC-423-423-0	*FAM193A; MEX3D; CLPTM1*	c.1390-2A>G; c.595+1G>C; c.953A>G, p.Tyr318Cys
PMGRC-446-446-0	*TP53BP2*	c.40G>A, p.Val14Met c.3240del, p.Asp1080Glufs*5
PMGRC-485-485-0	*HEATR1; PRR30*	c.3613C>T, p.Gln1205*; c.275del, p.Pro92Glnfs*33, c.-365-6C>T
PMGRC-525-525-0	*PLXNC1*	c.416del, p.Leu139Argfs*4
PMGRC-746-746-0	*DOP1A; PCDH9*	c.3559dup, p.Ile1187Asnfs*5; c.3291del, p.Ser1098Leufs*25

PMGRC-772-772-0	*MUC5B*	c.15477+1G>C
PMGRC-782-782-0	*CD177*	c.99G>A, p.Trp33*
PMGRC-832-832-0	*TLN2; IL9*	c.4645C>T, p.Asp1549Tyr; c.183+6T>C, p.Ser893Arg, c.202T>C, p.Cys68Arg
PMGRC-949-949-0	*GNRH2*	c.213_214insTGTC
PMGRC-985-985-0	*FMNL1*	c.1379del, p.Pro460Glnfs*19
PMGRC-992-992-0	*PRR14L*	c.4166del, p.Tyr318Cys
PMGRC-999-999-0	*SLC17A9*	c.287C>G, p.Tyr132Cys
PMGRC-1017-1017-0	*NXF3*	c.1231G>T, p.Glu411*
PMGRC-1030-1030-0	*CORO1C*	c.474G>A, p.Trp158*
PMGRC-1048-1048-0	*NEURL4*	c.1606G>T, p.Glu536*
PMGRC-1050-1050-0	*CUL1*	c.571_575del, p.Arg191Trpfs*2

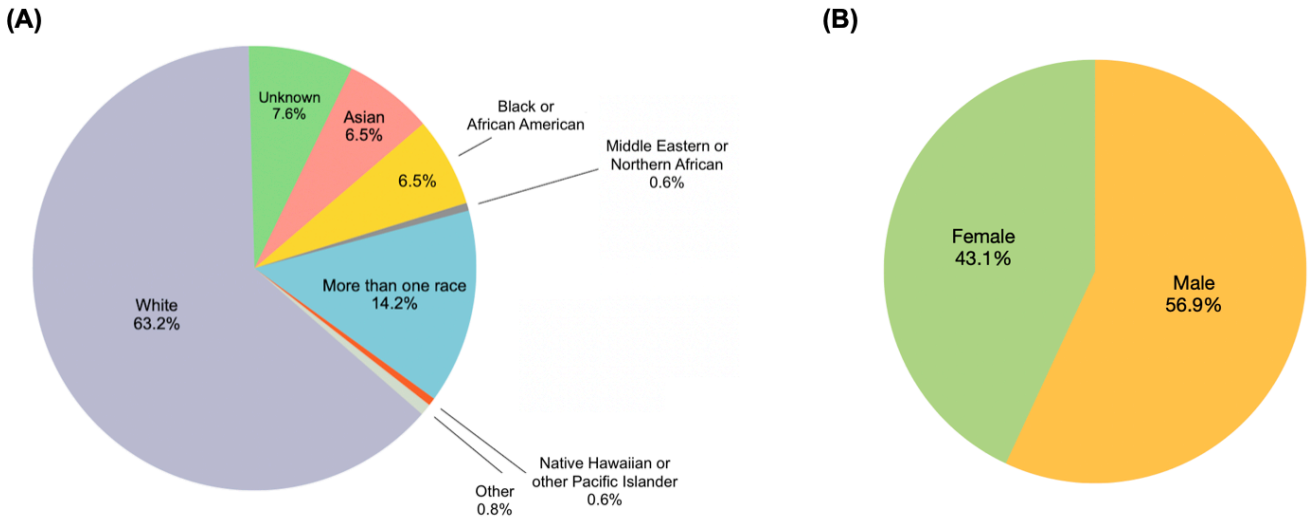


Figure 1. (A), (B) Sex and race distribution of probands that underwent genome sequencing.

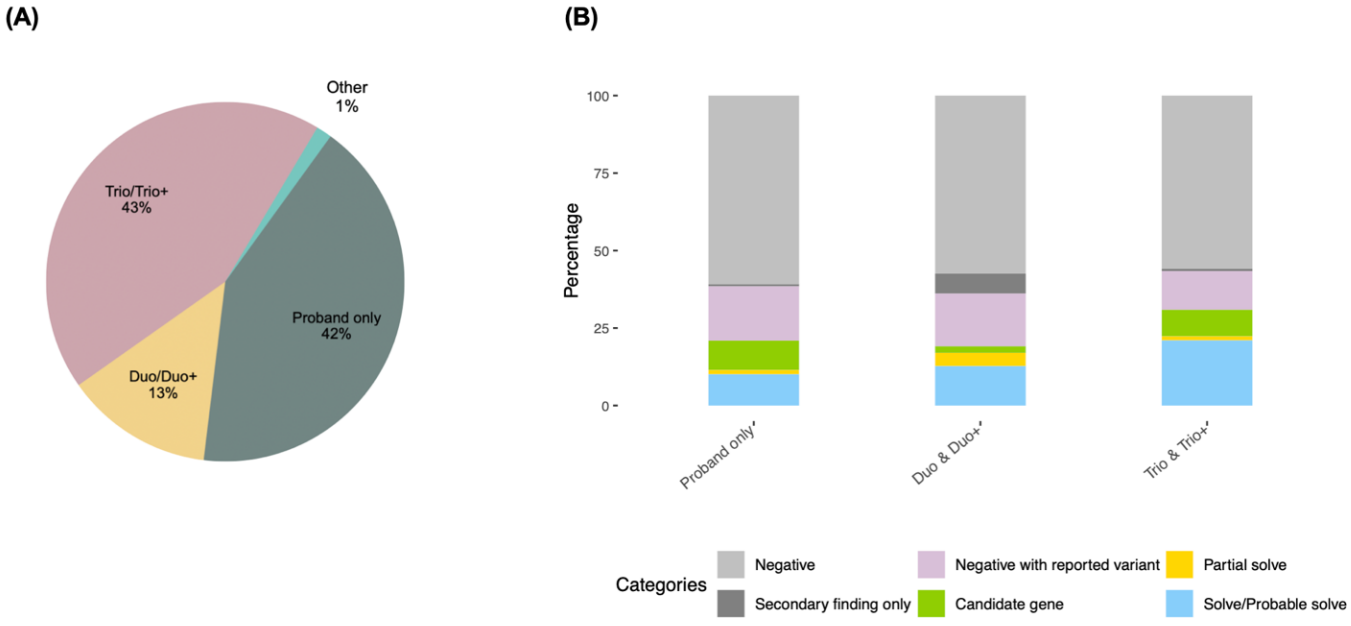
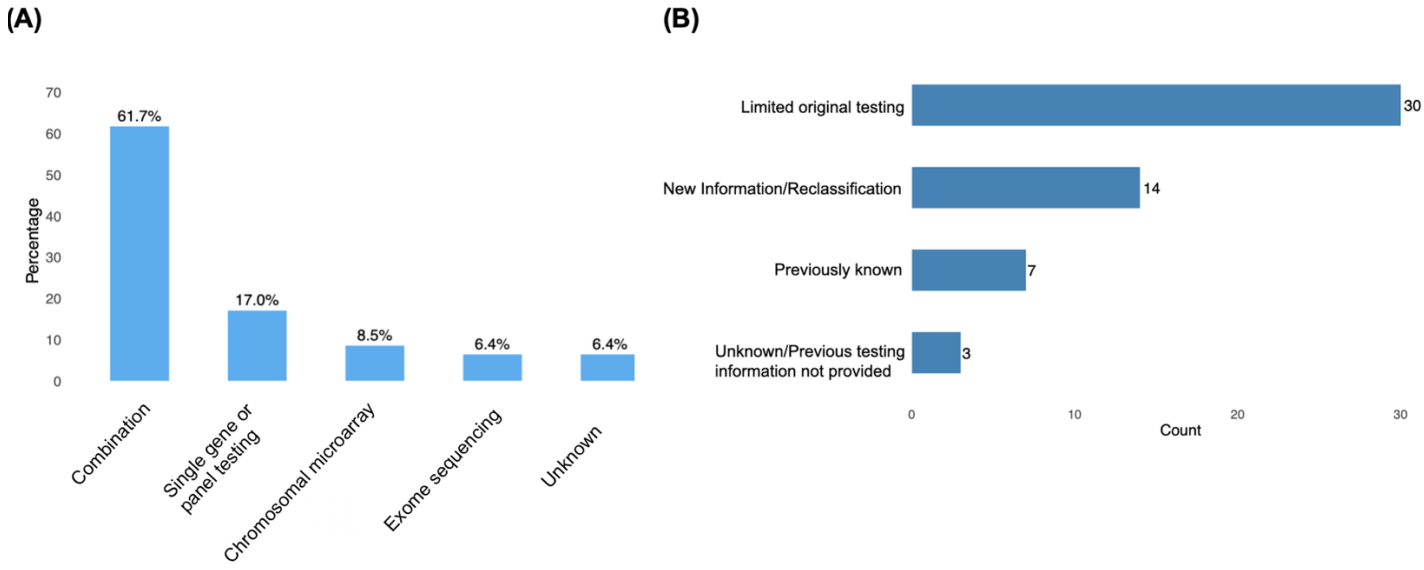


Figure 2. **(A)** Family structures corresponding to the participants **(B)**. Trio/Trio+ had the highest diagnostic yield (21%) whereas Duo/Duo+ and Proband only families reached a diagnostic yield of 13% and 10% respectively (p = 0.03, chi-square test).

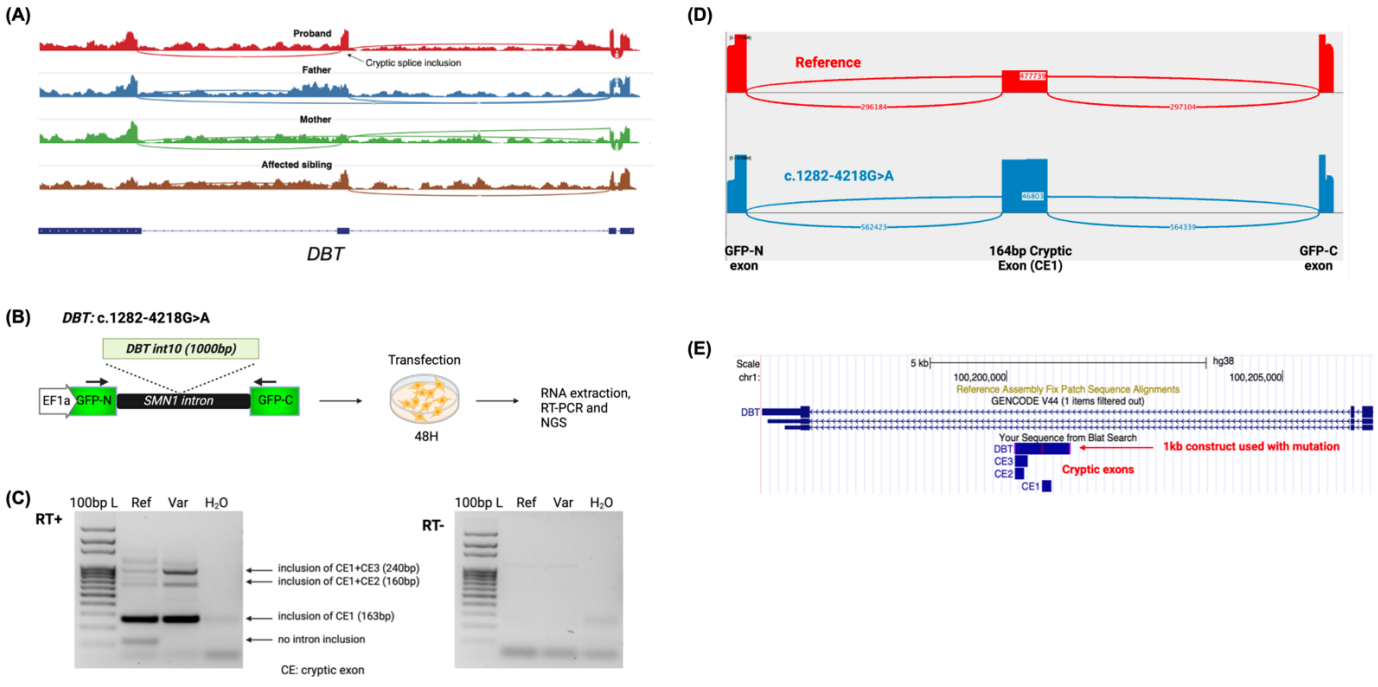


401

402 **Figure 3. (A)** Types of prior testing undergone by participants that were deemed solved or probably
403 solved. **(B)** Reasons diagnosis was previously missed on clinical genetic testing. Limited original testing
404 includes gene panels that lacked the causative gene or did not detect the causative variant due to its
405 location in a genomic region not covered by the test.

406

407



408

409 **Figure 4. (A)** RNA sequencing reads visualized with sashimi plot to indicate splicing events showing the
410 cryptic splice inclusion in *DBT* in the proband (PMGRC-658-658-0) and the affected sibling as a result of

a deep (-4230) homozygous intronic single nucleotide variant. **(B-E) Mini-gene splicing assay for the *DBT* c.1282-4218G>A variant. B)** Reference and c.1282-4218G>A variant *DBT* intron 10 fragments were cloned into a mini-gene split *GFP* construct, in which N and C-terminal parts of the *GFP* gene were separated by *SMN1* introns 7 and 8 (NM_000344). The construct was expressed in HEK293 cells for 48 hours followed by mRNA extraction, cDNA generation and NGS amplicon sequencing. Black arrows on the construct image indicate primer placement. **C)** RT-PCR gel electrophoresis demonstrates inclusion of the cryptic exon in both the wild type and reference transcript. **D)** Sashimi plots of reference and variant hybrid DBT-GFP constructs demonstrating a preferential cryptic exon (10B) inclusion in the sequences containing the c.1282-4218G>A variant. **E)** Inclusion of cryptic exon 10B leads to a premature stop codon at the end of the exon, which most likely escapes nonsense-mediated decay and leads to a truncated protein with a C-terminus different by 16-residues.

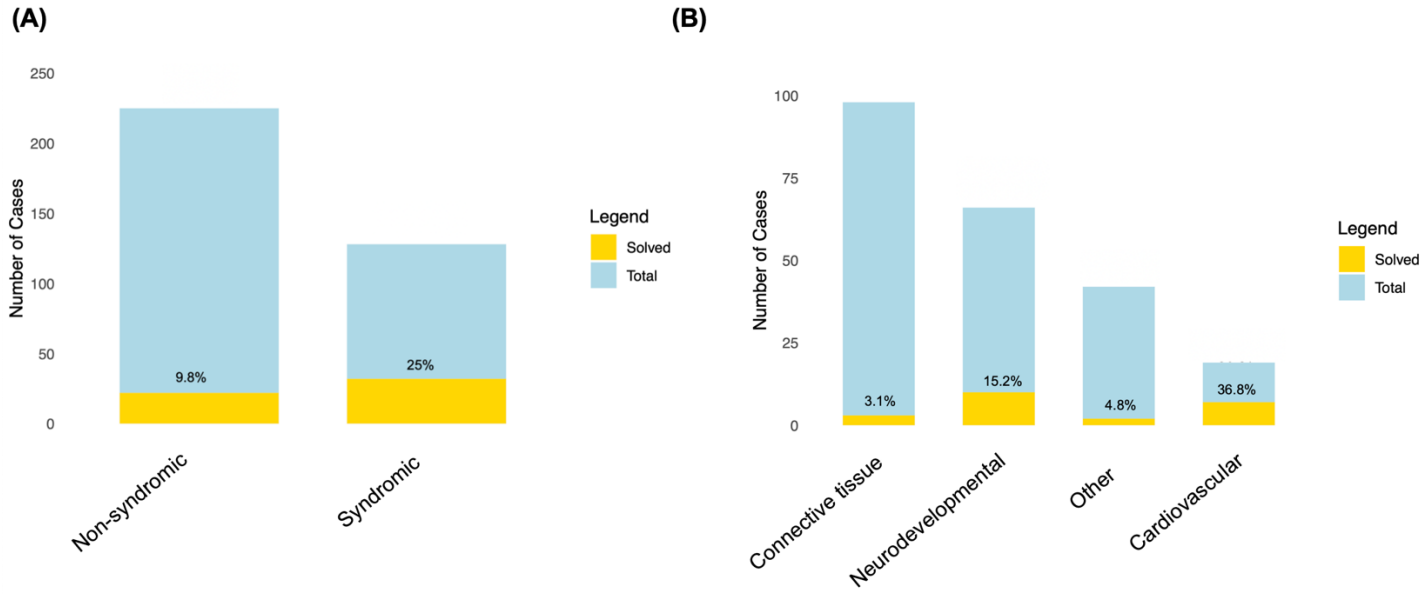


Figure 5. (**A**) The solve rate for syndromic versus non-syndromic cases. (**B**) The solve rate stratified by phenotypic category for the non-syndromic cases.

435

436

437

438

439

440 References

441

442 1. Slavotinek, A., et al., *Diagnostic yield of pediatric and prenatal exome sequencing in a diverse*
443 *population.* NPJ Genom Med, 2023. **8**(1): p. 10.
444 2. Wilke, M., et al., *Diagnostic yield of exome and genome sequencing after non-diagnostic multi-*
445 *gene panels in patients with single-system diseases.* Orphanet J Rare Dis, 2024. **19**(1): p. 216.
446 3. Jaganathan, K., et al., *Predicting Splicing from Primary Sequence with Deep Learning.* Cell, 2019.
447 **176**(3): p. 535-548 e24.
448 4. Zeng, T. and Y.I. Li, *Predicting RNA splicing from DNA sequence using Pangolin.* Genome Biol,
449 2022. **23**(1): p. 103.
450 5. Savage, S.K., et al., *Using a chat-based informed consent tool in large-scale genomic research.* J
451 Am Med Inform Assoc, 2024. **31**(2): p. 472-478.
452 6. Berger, S.I., et al., *Increased diagnostic yield from negative whole genome-slice panels using*
453 *automated reanalysis.* Clin Genet, 2023. **104**(3): p. 377-383.
454 7. Li, H., *Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM*. 2013:
455 arXiv.
456 8. Poplin, R., et al., *A universal SNP and small-indel variant caller using deep neural networks.* Nat
457 Biotechnol, 2018. **36**(10): p. 983-987.
458 9. Chen, X., et al., *Manta: rapid detection of structural variants and indels for germline and cancer*
459 *sequencing applications.* Bioinformatics, 2016. **32**(8): p. 1220-2.
460 10. Dobin, A., et al., *STAR: ultrafast universal RNA-seq aligner.* Bioinformatics, 2013. **29**(1): p. 15-21.
461 11. Robinson, J.T., et al., *igv.js: an embeddable JavaScript implementation of the Integrative*
462 *Genomics Viewer (IGV).* Bioinformatics, 2023. **39**(1).
463 12. Robinson, J.T., et al., *Variant Review with the Integrative Genomics Viewer.* Cancer Res, 2017.
464 **77**(21): p. e31-e34.
465 13. Robinson, J.T., et al., *Integrative genomics viewer.* Nat Biotechnol, 2011. **29**(1): p. 24-6.
466 14. Thorvaldsdottir, H., J.T. Robinson, and J.P. Mesirov, *Integrative Genomics Viewer (IGV): high-*
467 *performance genomics data visualization and exploration.* Brief Bioinform, 2013. **14**(2): p. 178-
468 92.
469 15. Scott, H.A., et al., *A high throughput splicing assay to investigate the effect of variants of*
470 *unknown significance on exon inclusion.* medRxiv, 2023: p. 2022.11.30.22282952.
471 16. Dobin, A. and T.R. Gingeras, *Optimizing RNA-Seq Mapping with STAR.* Methods Mol Biol, 2016.
472 **1415**: p. 245-62.
473 17. Clark, M.M., et al., *Diagnosis of genetic diseases in seriously ill children by rapid whole-genome*
474 *sequencing and automated phenotyping and interpretation.* Sci Transl Med, 2019. **11**(489).
475 18. O'Brien, T.D., et al., *Artificial intelligence (AI)-assisted exome reanalysis greatly aids in the*
476 *identification of new positive cases and reduces analysis time in a clinical diagnostic laboratory.*
477 Genet Med, 2022. **24**(1): p. 192-200.

19. Negi, S., et al., *Advancing long-read nanopore genome assembly and accurate variant calling for rare disease detection.* medRxiv, 2024.
20. Chen, Y., et al., *De novo variants in the RNU4-2 snRNA cause a frequent neurodevelopmental syndrome.* Nature, 2024. **632**(8026): p. 832-840.
21. Nguyen, T.H., et al., *The architecture of the spliceosomal U4/U6.U5 tri-snRNP.* Nature, 2015. **523**(7558): p. 47-52.
22. Ganesh, V.S., et al., *Novel syndromic neurodevelopmental disorder caused by de novo deletion of CHASERR, a long noncoding RNA.* medRxiv, 2024.
23. Wojcik, M.H., et al., *Genome Sequencing for Diagnosing Rare Diseases.* N Engl J Med, 2024. **390**(21): p. 1985-1997.
24. Palmer, E.E., et al., *Diagnostic Yield of Whole Genome Sequencing After Nondiagnostic Exome Sequencing or Gene Panel in Developmental and Epileptic Encephalopathies.* Neurology, 2021. **96**(13): p. e1770-e1782.
25. Alfares, A., et al., *Whole-genome sequencing offers additional but limited clinical utility compared with reanalysis of whole-exome sequencing.* Genet Med, 2018. **20**(11): p. 1328-1333.
26. Wojcik, M.H., et al., *Beyond the exome: What's next in diagnostic testing for Mendelian conditions.* Am J Hum Genet, 2023. **110**(8): p. 1229-1248.
27. Liu, Z., et al., *Hemizygous variants in protein phosphatase 1 regulatory subunit 3F (PPP1R3F) are associated with a neurodevelopmental disorder characterized by developmental delay, intellectual disability and autistic features.* Hum Mol Genet, 2023. **32**(20): p. 2981-2995.
28. Dzinovic, I., et al., *Variant recurrence confirms the existence of a FBXO31-related spastic-dystonic cerebral palsy syndrome.* Ann Clin Transl Neurol, 2021. **8**(4): p. 951-955.
29. Thomas Cassini, S.S., Molly Behan, Cynthia J. Tifft, May Christine Malicdan, David R. Adams, Undiagnosed Diseases Network, Sun-Young Ahn, Debra S. Regier, *Mitochondrial trifunctional protein deficiency caused by a deep intronic deletion leading to aberrant splicing.* JIMD Reports, 2024. **66**(1).
30. Dias, C., et al., *De novo variants in TCF7L2 are associated with a syndromic neurodevelopmental disorder.* Am J Med Genet A, 2021. **185**(8): p. 2384-2390.
31. Seaby, E.G., et al., *Monoallelic de novo variants in DDX17 cause a neurodevelopmental disorder.* Brain, 2024.