Unraveling undiagnosed rare disease cases by HiFi long-read genome sequencing

Wouter Steyaert,[1,36] Lydia Sagath,[1,36] German Demidov,[2] Vicente A. Yépez,[3] Anna Esteve-Codina,[4,5] Julien Gagneur,[3,6,7] Kornelia Ellwanger,[2,8] Ronny Derks,[1] Marjan Weiss,[1] Amber den Ouden,[1] Simone van den Heuvel,[1] Hilde Swinkels,[1] Nick Zomer,[1] Marloes Steehouwer,[1] Luke O'Gorman,[1] Galuh Astuti,[1] Kornelia Neveling,[1] Rebecca Schüle,[9,10] Jishu Xu,[9] Matthis Synofzik,[9,11] Danique Beijer,[9] Holger Hengel,[9] Ludger Schöls,[9,11] Kristl G. Claeys,[12,13] Jonathan Baets,[14,15,16] Liedewei Van de Vondel,[14,15] Alessandra Ferlini,[17] Rita Selvatici,[17] Heba Morsy,[18] Marwa Saeed Abd Elmaksoud,[19] Volker Straub,[20] Juliane Müller,[21] Veronica Pini,[21] Luke Perry,[21,22] Anna Sarkozy,[21] Irina Zaharieva,[21] Francesco Muntoni,[21,22] Enrico Bugiardini,[23] Kiran Polavarapu,[24] Rita Horvath,[25] Evan Reid,[26] Hanns Lochmüller,[27,28,29] Marco Spinazzi,[30] Marco Savarese,[31,32] Solve-RD DITF-ITHACA, Solve-RD DITF-Euro-NMD, Solve-RD DITF-RND, Solve-RD DITF-EpiCARE, Leslie Matalonga,[4,5] Steven Laurie,[4,5] Han G. Brunner,[1,33] Holm Graessner,[2,8] Sergi Beltran,[4,34] Stephan Ossowski,[2] Lisenka E.L.M. Vissers,[1] Christian Gilissen,[1,37] Alexander Hoischen,[1,35,37] and on behalf of the Solve-RD consortium

Solve-RD is a pan-European rare disease (RD) research program that aims to identify disease-causing genetic variants in previously undiagnosed RD families. We utilized 10-fold coverage HiFi long-read sequencing (LRS) for detecting causative structural variants (SVs), single-nucleotide variants (SNVs), insertion-deletions (indels), and short tandem repeat (STR) expansions in previously studied RD families without a clear molecular diagnosis. Our cohort includes 293 individuals from 114 genetically undiagnosed RD families selected by European Reference Network (ERN) experts. Of these, 21 families were affected by so-called "unsolvable" syndromes for which genetic causes remain unknown and for which prior testing was not a prerequisite. The remaining 93 families had at least one individual affected by a rare neurological, neuromuscular, or epilepsy disorder without a genetic diagnosis despite extensive prior testing. Clinical interpretation and orthogonal validation of variants in known disease genes yielded 12 novel genetic diagnoses due to de novo and rare inherited SNVs, indels, SVs, and STR expansions. In an additional five families, we identified a candidate disease-causing variant, including an *MCF2*/*FGF13* fusion and a *PSMA3* deletion. However, no common genetic cause was identified in any of the "unsolvable" syndromes. Taken together, we found (likely) disease-causing genetic variants in 11.8% of previously unsolved families and additional candidate disease-causing SVs in another 5.4% of these families. In conclusion, our results demonstrate the potential added value of HiFi long-read genome sequencing in undiagnosed rare diseases.

[Supplemental material is available for this article.]

[1] *Radboud University Medical Center, Department of Human Genetics, Research Institute for Medical Innovation, 6500 HB Nijmegen, Netherlands;* [2] *Universitätsklinikum Tübingen - Institut für Medizinische Genetik und angewandte Genomik, 72076 Tübingen, Germany;* [3] *TUM School of Computation, Information and Technology, Technical University of Munich, 85748 Garching, Germany;* [4] *Centro Nacional de Análisis Genómico (CNAG), 08028 Barcelona, Spain;* [5] *Universitat de Barcelona (UB), 08007 Barcelona, Spain;*

[36] These authors contributed equally to this work.
[37] These authors jointly supervised this work.
Corresponding author: alexander.hoischen@radboudumc.nl
Article published online before print. Article, supplemental material, and publication date are at https://www.genome.org/cgi/doi/10.1101/gr.279414.124. Freely available online through the *Genome Research* Open Access option.

[6]Institute of Human Genetics, School of Medicine, Technical University of Munich, 81675 Munich, Germany; [7]Computational Health Center, Helmholtz Center Munich, 85764 Neuherberg, Germany; [8]Center for Rare Diseases, University Hospital Tübingen, 72076 Tübingen, Germany; [9]Hertie-Institute for Clinical Brain Research and Center of Neurology, University of Tübingen, 72076 Tübingen, Germany; [10]Division of Neurodegenerative Diseases, Department of Neurology, Heidelberg University Hospital and Faculty of Medicine, 69120 Heidelberg, Germany; [11]German Center of Neurodegenerative Diseases (DZNE), 72076 Tübingen, Germany; [12]Department of Neurology, University Hospitals Leuven, 3000 Leuven, Belgium; [13]Department of Neurosciences, Laboratory for Muscle Diseases and Neuropathies, KU Leuven, and Leuven Brain Institute (LBI), 3000 Leuven, Belgium; [14]Translational Neurosciences, Faculty of Medicine and Health Sciences, University of Antwerp, 2610 Antwerp, Belgium; [15]Laboratory of Neuromuscular Pathology, Institute Born-Bunge, University of Antwerp, 2610 Antwerp, Belgium; [16]Neuromuscular Reference Center, Department of Neurology, Antwerp University Hospital, 2650 Antwerp, Belgium; [17]Unit of Medical Genetics, Department of Medical Sciences, University of Ferrara, 44121 Ferrara, Italy; [18]Department of Neuromuscular Diseases, UCL Queen Square Institute of Neurology and The National Hospital for Neurology and Neurosurgery, London WC1N 3BG, United Kingdom; [19]Neurology Unit, Department of Pediatrics, Faculty of Medicine, Alexandria University, Alexandria 5372066, Egypt; [20]John Walton Muscular Dystrophy Research Centre, Translational and Clinical Research Institute, Newcastle University and Newcastle Hospitals NHS Foundation Trust, Newcastle upon Tyne NE1 3BZ, United Kingdom; [21]Dubowitz Neuromuscular Centre, UCL Great Ormond Street Hospital, London WC1N 3JH, United Kingdom; [22]NIHR Great Ormond Street Hospital Biomedical Research Centre, London WC1N 1EH, United Kingdom; [23]Department of Neuromuscular Diseases, UCL Queen Square Institute of Neurology, London WC1N 3BG, United Kingdom; [24]Children's Hospital of Eastern Ontario Research Institute, University of Ottawa and Division of Neurology, Department of Medicine, The Ottawa Hospital, Ottawa, ON K1H 8L1, Canada; [25]Department of Clinical Neurosciences, John Van Geest Centre for Brain Repair, School of Clinical Medicine, University of Cambridge, Cambridge CB2 0PY, United Kingdom; [26]Cambridge Institute for Medical Research and Department of Medical Genetics, University of Cambridge, Cambridge CB2 0XY, United Kingdom; [27]Children's Hospital of Eastern Ontario, University of Ottawa, Ottawa, Ontario, ON K1H 8M8, Canada; [28]Brain and Mind Research Institute, University of Ottawa, Ottawa, Ontario, ON K1H 8M5, Canada; [29]The Ottawa Hospital Research Institute, University of Ottawa, Ottawa, Ontario, ON K1Y 4E9, Canada; [30]Department of Neurology, Centre Hospitalier Universitaire d'Angers, 49933 Angers, France; [31]Folkhälsan Research Center, 00250 Helsinki, Uusimaa, Finland; [32]Faculty of Medicine, University of Helsinki, 00014 University of Helsinki, Uusimaa, Finland; [33]Department of Clinical Genetics, Maastricht University Medical Center, 6229 HX Maastricht, Netherlands; [34]Departament de Genètica, Microbiologia i Estadística, Facultat de Biologia, Universitat de Barcelona (UB), 08028 Barcelona, Spain; [35]Radboud University Medical Center, Department of Internal Medicine; Radboud Expertise Center for Immunodeficiency and Autoinflammation and Radboud Center for Infectious Disease (RCI), 6500 HB Nijmegen, Netherlands

Rare diseases (RDs) affect 400 million people worldwide (Nguengang Wakap et al. 2020). It is estimated that 80% of these diseases have a genetic origin (Sernadela et al. 2017). Pinpointing the disease-causing genetic variant is important for RD families, because it ends an often time-consuming, stressful, and costly diagnostic odyssey (Biesecker and Green 2014). In addition, several disease management strategies and treatment options depend on the specific disease gene or variant (Pogue et al. 2018).

With routinely used short-read sequencing (SRS) technologies, such as exome and genome sequencing, diagnostic yields vary between 8% and 70%, depending on the diseases studied and inclusion criteria used (Wright et al. 2018). Besides incomplete knowledge of the functional and phenotypic consequence of genetic variation, shortcomings at the variant identification level may substantially contribute to the fact that many RD patients remain genetically undiagnosed. Indeed, SRS technologies result in an almost complete characterization of short genetic variants (single- and multinucleotide substitutions and small insertions and deletions) in the unique portions of an individual's genome, but the analysis of duplicated and repetitive genomic regions and particularly the identification of structural variants (SVs) and short tandem repeat (STR) expansions remain far from complete (Chaisson et al. 2019; Chintalaphani et al. 2021; Porubsky et al. 2023). Several recent studies demonstrate that long-read sequencing (LRS) technologies uncover a whole new reservoir of (structural) genetic variation (Chaisson et al. 2019; Zook et al. 2020; Beyter et al. 2021; Pauper et al. 2021; Kucuk et al. 2023). Now that these LRS technologies produce high-quality sequencing reads at steadily dropping costs, researchers are able to evaluate the hypothesis that part of the genetically undiagnosed RDs is caused by variants that remain hidden from previously used technologies. The exploration and interpretation of SVs in undiagnosed RD families by LRS have indeed shown to be successful in the past couple of years for several disease phenotypes (Merker et al. 2018; Sanchis-Juan et al. 2018; Mizuguchi Suzuki et al. 2019a,b; Zeng et al. 2019; Fadaie et al. 2021). Here, as part of the Solve-RD consortium effort, we applied long-read genome sequencing to 293 individuals from 114 previously undiagnosed RD families to show the potential additional benefits of LRS for resolving the genetic cause in rare disease patients.

Results

Our complete study cohort consists of two different subcohorts (Fig. 1; Supplemental Table S1). Firstly, a subcohort of 21 families (including 16 trios) with clinically well-recognized, so-called "unsolvable" syndromes, including Aicardi (MIM ID % 304050), Hallermann–Streiff (%234100), Gomez–Lopez-Hernandez (%601853), Pai (%155145), and syndromes belonging to the oculoauriculovertebral spectrum, all of which remain genetically elusive despite huge global efforts to identify the disease cause. These patients had not necessarily undergone previous testing. The second subcohort consisted of 232 individuals from 93

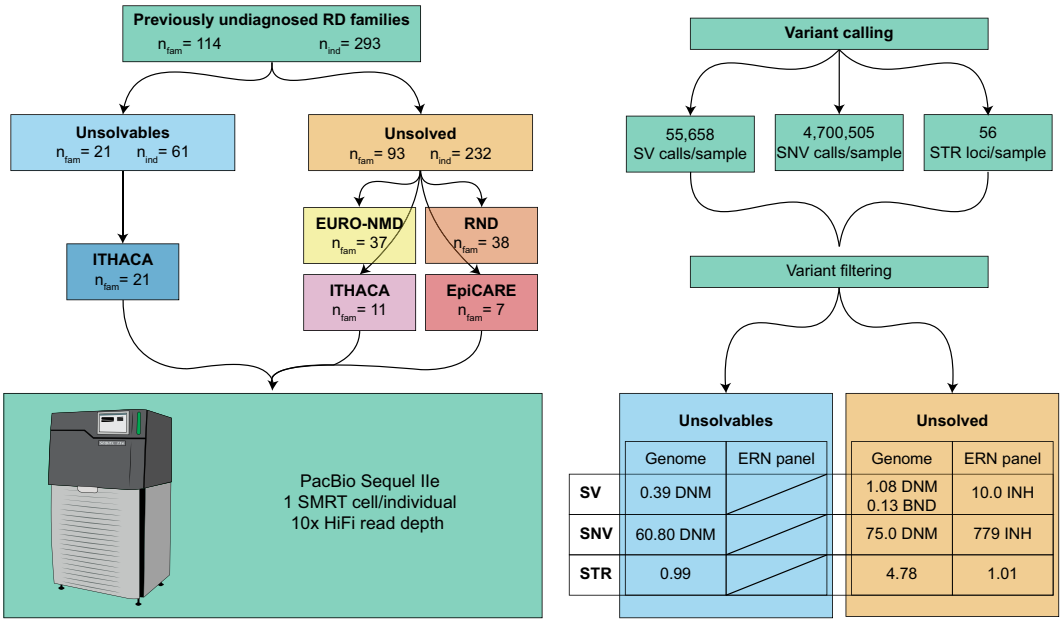

Figure 1. HiFi LRS in a unique cohort of 293 individuals from 114 RD families. The study cohort consists of two subcohorts: the "unsolvables" (families affected by clinically well-recognizable syndromes for which the cause is yet unknown) and the "unsolved" (families affected by a rare neurological, neuromuscular, or epilepsy disease). All patients were recruited via four European Reference Networks and subsequently sequenced using a single SMRT cell of sequencing data per individual. Genome-wide calling of SVs and SNVs was conducted, and STRs were genotyped at 56 known disease-associated loci. (ERN) European Reference Network, (BND) breakend call, (INH) inherited variant, (DNM) de novo mutation.

families with rare neurological, neuromuscular, or epilepsy disorders selected by experts from ERN-EURO-NMD, ERN-EpiCARE, ERN-RND, and ERN-ITHACA. While most of these patients are affected by conditions for which several genetic causes are known, these particular families remained "unsolved"; as extensive diagnostic and/or research testing, including prior exome or genome sequencing, had failed to yield a diagnosis (Supplemental Table S2).

Within known disease-relevant genes (ERN-specific gene lists; Methods), we assessed all types of SVs. Outside these gene panels, we focused our analysis on putative de novo events in parent-offspring trios and, due to the lack of effective population databases for SVs, on large inherited SVs (>100 kb [corresponding to breakend calls]) since these events are more likely to affect the phenotype (Methods). We also genotyped 56 known disease-associated STR loci and assessed all rare SNVs within known disease-relevant genes and, in parent-offspring trios, on putative de novo events in the complete genome (Methods).

On average, we identified 55,658 SVs (≥20 bp; 23,385 SVs ≥ 50 bp) and 4,700,505 SNVs per individual (Methods; Supplemental Table S3). Of these, 13,481 SVs and 43,172 SNVs are private to a single family. In the 42 parent-offspring trios included in the study, we identified an average of 32.6 candidate de novo SVs, which were reduced to 8.5 after quality filtering. Following visual curation in the Integrative Genomics Viewer (IGV), this number was further refined to an average of 0.81 per trio.

From the 18 visually curated putative de novo SVs for which flanking sequencing primers could be designed (0.43 per trio on average), four were confirmed as de novo variants in the child. Of the remaining calls, five were false positives, six were true variants inherited from a parent, and three were true positives where the parental sequences failed (Methods; Supplemental Fig. S1; Supplemental Tables S4 and S5).

Identification of (likely) pathogenic variants in previously undiagnosed RD

Unsolvable syndromes

In the subcohort consisting of 21 families with "unsolvable" syndromes, we could not identify a gene or locus in which rare (de novo) variants were present in multiple families with the same syndrome. However, in a sporadic female patient (P0185637) initially diagnosed with Aicardi syndrome, presenting with global developmental delay, partial agenesis of the corpus callosum, and abnormalities with the vasculature and innervation of the eye, we identified a de novo missense variant in *TUBA1A* (tubulin alpha 1a, MIM ID *602529, NM_006009.4, Chr12:g.49,185,725C>T, c.641G>A, p.(Arg214His); Fig. 2; Table 1). The variant has previously been described as a cause of lissencephaly 3 (LIS3, MIM ID #611603; Bahi-Buisson et al. 2014). Clinical reassessment of the patient's phenotype confirmed the new diagnosis.

Disease-causing variants identified in rare neurological and neuromuscular diseases, and epilepsies

After prioritizing and clinically interpreting genetic variants in the 93 families from the "unsolved" subcohort, we established a genetic diagnosis in 11 of them (Methods; Table 1).

Structural variants

In two unrelated male patients (P0078963 and P0695060; Fig. 3A–D) with muscular dystrophy, we identified disease-explanatory inversions breaking *DMD* (dystrophin, MIM ID *300377, NM_004006.3; Fig. 3A,C). The breakpoints in patient P0078963 are ChrX:23,308,848 and ChrX:32,004,110 GRCh38 (hg38) resulting in an inversion of 8.7 Mb, which breaks *DMD* in intron

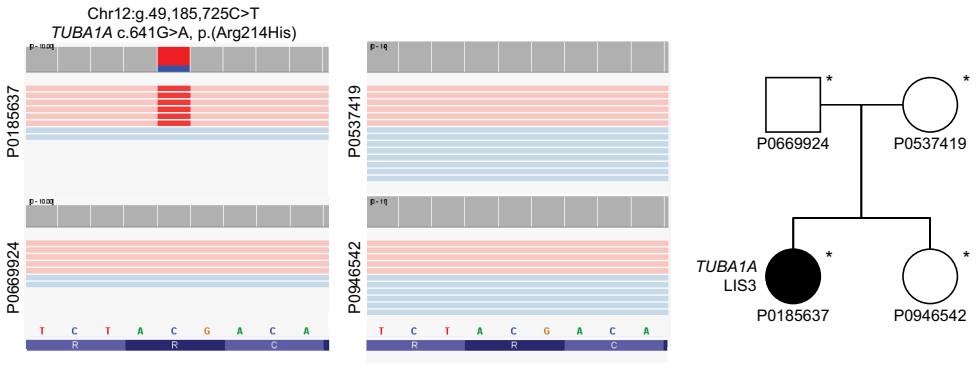

Figure 2. Visualization of the *TUBA1A* de novo missense variant in P0185637 using IGV, and a pedigree of the family. The variant had earlier been described as a cause of lissencephaly. Healthy family members do not carry the variant. Sequenced individuals are marked with an asterisk (*) in the pedigrees.

44, resulting in a truncated transcript (Fig. 3A). This event was initially discovered by optical genome mapping (OGM) and LRS detected the exact breakpoints of the event. In patient P0695060, an inversion of ChrX:17,398,320–32,130,845 was identified (Fig. 3C). This event breaks *NHS* (NHS actin remodeling regulator, MIM ID *300457, NM_0129186.2) in intron 1 and *DMD* in intron 44. This event was confirmed by Sanger sequencing, which also highlighted the insertion of a short ATAAT sequence in the first intron of *NHS*, and of a 38 nt sequence in the intron 44 of *DMD*, which likely favored the inversion. As these genes have opposite orientations on the chromosome, the inversion results in two theoretical fusion genes in which the exon orientation is conserved. However, neither theoretical gene product is in frame much past the fusion breakpoint. The likely disruption of both genes is also in line with the patient's phenotype, who in hindsight presents not only with a dystrophy but also with a cataract, which is a characteristic feature of Nance–Horan syndrome (MIM ID #302350) caused by loss-of-function variants of *NHS*.

In a duo consisting of an affected father and affected son (P0011782 and P0011781, respectively; Fig. 3E–H) presenting with hereditary spastic paraplegia (HSP), we detected a 1.2 kb deletion encompassing the entire exon 6 of *REEP1* (receptor accessory protein 1, MIM ID *609139, NM_001371279.1), Chr2: g.86,232,216–86,233,399del, eventually leading to a frameshift mutation p.(Gly140Cysfs*18) that removes exons 6–9. Both the father and the son are heterozygous carriers of this deletion (Fig. 3F, G). Variants in *REEP1* have been described in autosomal recessive distal hereditary neuronopathy (MIM ID #620011) and autosomal dominant (AD) spastic paraplegia 31 (SPG31, MIM ID #610250). Of SNVs, frameshift variants are the most common causative variant type in SPG31 cases (Beetz et al. 2008). In addition, single-exon deletions in *REEP1* of exons 2 and 3 have been described as pathogenic (Goizet et al. 2011). Additionally, two cases with deletions encompassing more than one exon have been described (Ishiura et al. 2014), neither affecting exon 6.

In a singleton patient with adult-onset distal myopathy (P0657753, Fig. 3I–K), a 65 kb duplication involving *MYOT* (myotilin, MIM ID *604103, Chr5:137,832,296–137,897,203) had earlier been identified by a gene panel for myofibrillar myopathy. Because we did not observe this variant in our filtered variant calls, we reverted to the raw sequencing data for this specific case, which allowed us to determine that the duplication was in tandem. The variant also segregates with the probands similarly affected sibling. Heterozygous variants in *MYOT* are a known cause of myofibrillar myopathy 3 (MIM ID #609200), a slowly progressive muscle disorder with an adult onset. A separate report on the family, including functional validation showing increased expression of myotilin, has been published (Spinazzi et al. 2025).

Repeat expansions

In two families with autosomal dominant ataxia (AD-ATX), we identified disease-explanatory heterozygous expansions of the GGCCTG motif in intron 1 of the *NOP56* gene (Table 1; NOP56 ribonucleoprotein, NM_006392.4, MIM ID *614154). Repeat expansions in *NOP56* are a known cause of AD spinocerebellar ataxia 36 (SCA36, MIM ID #614153; Kobayashi et al. 2011). The hexanucleotide motif count in a duo consisting of two affected siblings (P0016368 and P0018504; Fig. 4A) was estimated at >1200. In the other family, consisting of two affected family members in two generations and one unaffected family member (P0018996, P0019023, and P0019024, respectively, Fig. 4B), the motif count was >34 in the affected mother and >45 in the affected child. The pathogenic repeat threshold of *NOP56* is generally regarded to be 650 hexanucleotide repeats; however, shorter repeats are also known to be causative (Obayashi et al. 2015). The repeat expansion in the latter family was also discovered by reduced expression through RNA-seq and whole genome sequencing by parallel efforts in Solve-RD (Supplemental Fig. S2).

In a family presenting with AD-ATX (P0016356, P0019033), we found a repeat expansion in *DAB1* (DAB adaptor protein 1, NM_001365792.1, MIM ID *603448), a known causal gene for Spinocerebellar ataxia 37 (SCA37, MIM ID #615945). Age-dependent penetrant alleles have been reported to have an insertion of 31–75 ATTTC repeats, while the normal motifs are usually uninterrupted and consisting of 7–400 units of ATTTT (Matilla-Dueñas and Volpini 1993). The analysis of LRS data indicated the presence of two alleles in the index case P0016356, one with seven ATTTT repeats and another with a complex structure of (estimated) 615 ATTTT motifs, followed by 117 ATTTC repeats and then again by 29 ATTTT repeats (Fig. 4C), supported by five high-quality spanning LRS reads.

Finally, a homozygous repeat of the pathogenic AAGGG motif in the *RFC1* gene (replication factor C, NM_002913.5, MIM ID *102579) was found in a patient with ataxia (P0019027). Repeat expansions in *RFC1* are known to cause CANVAS-spectrum disorder ("cerebellar ataxia, neuropathy, and vestibular areflexia syndrome," MIM ID #614575), being very consistent with the observed phenotype. The number of AAGGG motifs was estimated

Table 1. Overview of pathogenic (P) and likely-pathogenic (LP) genetic variants, and candidate disease-causing variants (variants of unknown significance [VUS]) in the complete study cohort

Participant ID	ERN	Cohort	Classification	Gene symbol (transcript)	Variant type	Inheritance	HGVS	gnomAD allele frequency	Orthogonal validation
P0185637	ITHACA	Unsolvables	P	*TUBA1A* (NM_006009.4)	SNV	De novo AD	Chr12:g.49185725C>T, c.641G>A, p.(Arg214His)	Absent	–
P0695060	EURO-NMD	Unsolved	P	*DMD, NHS* (NM_004006.3, NM_001291867.2)	SV (inversion)	XLR;XLD	ChrX:g.17398320_32130845inv	Absent	Sanger
P0078963	EURO-NMD	Unsolved	P	*DMD (NM_004006.3)*	SV (inversion)	XLR	ChrX:g.23308848_32004110inv	Absent	OGM
P0016368	RND	Unsolved	P	*NOP56* (NM_006392.4)	STR	AD	Chr20:g.2652734_2652756GGCCTG [1200]	Absent	RNA-seq
P0018996	RND	Unsolved	P	*NOP56* (NM_006392.4)	STR	AD	Chr20:g.2652734_2652756GGCCTG[34]	Absent	RNA-seq
P0016356	RND	Unsolved	P	*DAB1* (NM_001365792.1)	STR	AD	Chr1:g.57367044_57367118AAAAT[29] GAAAT[117]AAAAT[615]	Absent	–
P0019022	RND	Unsolved	P	*RFC1* (NM_002913.5)	STR	AR	Chr4:g.39348427_39348476delins AAGGG[1181]; Chr4:g.39348427_39348476delins AAGGG[271]	Absent	RNA-seq
P0008178	EURO-NMD	Unsolved	P	*DMD* (NM_004006.3)	SNV (deep intronic)	XLR	ChrX:g.33174335C>T, c.31 + 36947G>A	Absent	Sanger
P0016160	RND	Unsolved	LP	*SPAST* (NM_041946.4)	SNV (intronic)	AD	Chr2:g.32115840G>A, c.1004 + 5G>A, p.(spl)	6.24×10^{-4}	ES, exon-skipping, Sanger
P0631224	RND	Unsolved	P; P	*TTN* (NM_001267550.2)	SNV	AR, maternally inherited, de novo on paternal allele	Chr2:g.178530761dup, c.105854dup, p.(Pro35286Thrfs*13); Chr2:g.178640613del, c.40652del, p.(Pro13551Glnfs*47)	Absent; Absent	SRS
P0657753	EURO-NMD	Unsolved	LP	*MYOT* (NM_006790.3)	SV (tandem duplication)	AD	Chr5:g.137832296_137897203dup	Absent	SRS
P0011781	RND	Unsolved	LP	*REEP1* (NM_001371279.1)	SV (deletion)	AD	Chr2:g.86232216_86233399del, c.418-597_595 + 409del, p.(Gly140Cysfs*18)	Absent	PCR + LRS
P0237528	EURO-NMD	Unsolved	VUS	*REEP1* (NM_001371279.1)	SNV (deep intronic)	AD	Chr2:g.86327804T>C, c.32 + 9675A>G	6.57×10^{-6}	–
P0936700	EURO-NMD	Unsolved	VUS	*FGF13, MCF2, and F9* (NM_004114.5, NM_001171876.2, NM_000133.4)	SV (duplication)	De novo AD/XLR	ChrX:g.139164887_139679311dup	Absent	PCR + LRS + cDNA + RNA-seq
P0021581	EURO-NMD	Unsolved	VUS	*PSMA3* (NM_002788.4)	SV (deletion)	De novo AD	Chr14:g.58268649_58283944del	Absent	PCR + Sanger
P0537031	ITHACA	Unsolved	VUS	*CPE, TLL1, NEK1, CLCN3, …*	SV (5 Mb duplication)	N/A	Chr4:g.165447976_170473344dup	Absent	Array CGH, ES
P0016165	RND	Unsolved	VUS	*ARMC9, NCL*	SV (300 kb duplication)	AD	Chr2:g.231348004_231684006dup	Absent	–

Allele frequency databases used: gnomAD v.4.1.0 (SNVs), gnomAD SVs v.4.1.0 (SVs), and gnomAD v.3.1.2 (STRs).

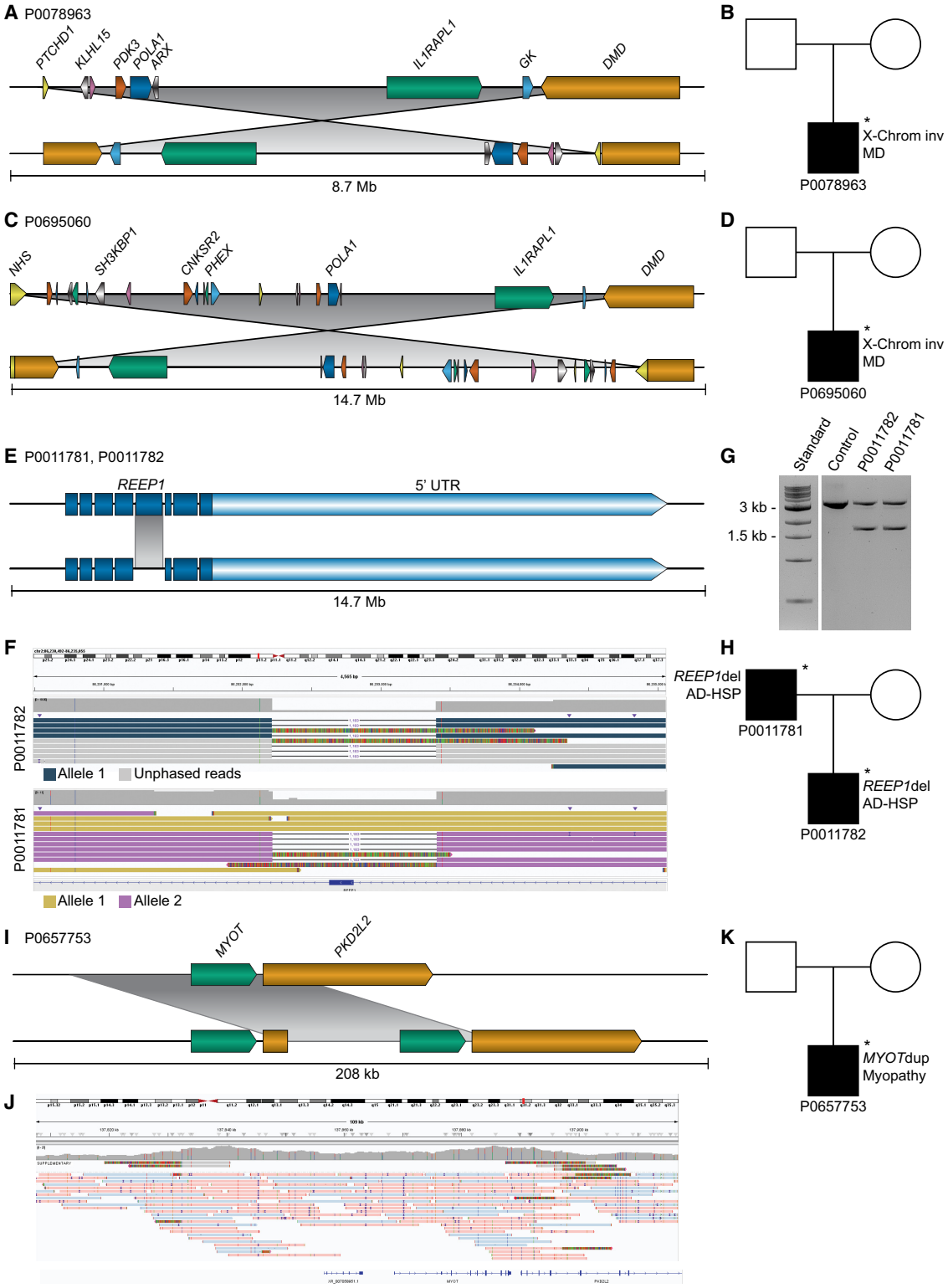

Figure 3. Visualization of disease-causing SVs in the "unsolved" subcohort in the form of cartoons and/or IGV screenshots, along with corresponding pedigrees. In two unrelated male patients (P0078963 in *A*,*B*, P0695060 in *C*,*D*) with muscular dystrophy, we found X-Chromosomal inversions (*A–D*). In both cases, *DMD* is disrupted (*A*,*C*), in one a second gene disruption adds to the phenotype (*C*). In a father and son with hereditary spastic paraplegia, we detected a deletion of *REEP1* exon 6 (*E–H*). The deletion in P001782 and P0011781 is shown here as a cartoon (*E*) and as a screenshot in IGV (*F*). The deletion was also visualized by agarose gel electrophoresis, which confirms that both patients are heterozygous for the deletion (*G*). The pedigree of the family is shown in *H*. In a patient with adult-onset distal myopathy, a 65 kb duplication involving *MYOT* (*I*) was confirmed to be in tandem by LRS (*J*). The pedigree of the family is shown in *K*. Sequenced individuals are marked with an asterisk (*) in the pedigrees (*B*,*D*,*G*,*H*,*K*). (MD) Muscular dystrophy, (AD-HSP) autosomal dominant hereditary spastic paraplegia.

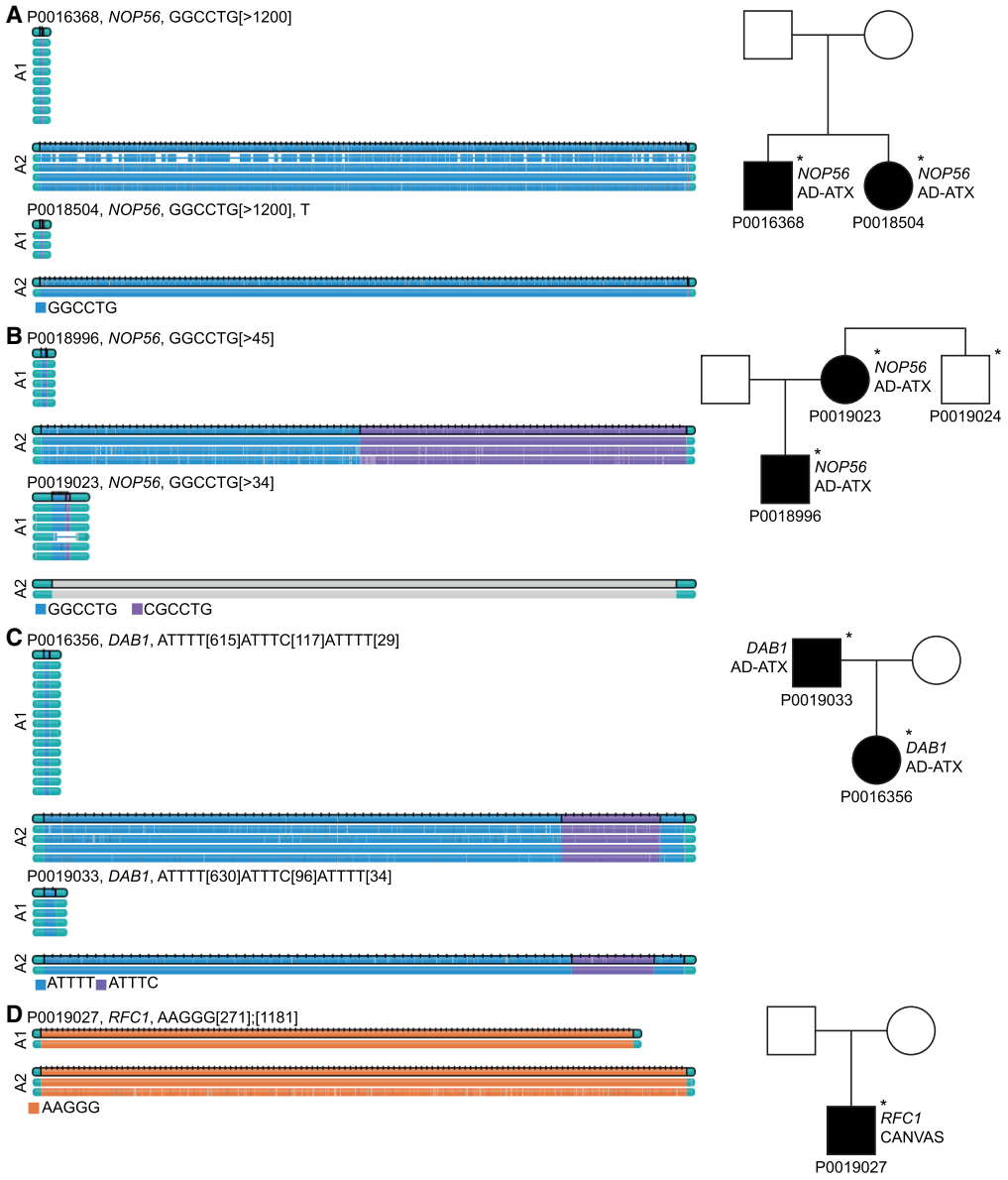

Figure 4. Visualizations were produced using the PacBio TRGT tool and pedigrees for the families with pathogenic STR expansions. In siblings P0016368 and P0018504, a heterozygous GGCCTG expansion in *NOP56* was detected (*A*). In another family, an expansion including the motifs GGCCTG and CGCCTG in *NOP56* was detected in one generation (P0018996), and the STR expansion was subsequently also identified in the mother (*B*). In patient P0016356 and their father, we identified heterozygous STR expansion *DAB1*, including both ATTTT and ATTTC motifs (*C*). In another patient, we identified homozygous STR expansions in *RFC1* (*D*). Alleles are denoted by "A1" and "A2." Sequenced individuals are marked with an asterisk (*) in the pedigrees. (AD-ATX) Autosomal dominant ataxia.

by the tool to be 271 on one allele and 1181 on the other allele (Fig. 4D). However, it is possible that the first allele is longer than 271 pathogenic repeats, since it was inferred based on soft clipped reads, not reads spanning the full repeat. The visualization of LRS data from this patient in IGV indicated that no normal alleles were present. The further validation of this likely causative repeat is to be described elsewhere.

Single-nucleotide variants

In a sporadic male patient with suspected titinopathy (P0008178), presenting with progressive proximal muscle weakness and myopathic features in his muscle biopsy, we identified a deep intronic SNV in *DMD* (ChrX:33,174,335C>T) (Fig. 5A). This variant has previously been shown to be a cause of Becker muscular dystrophy (BMD, MIM ID #300376) through exonization of a 149 bp sequence within intron 1 of *DMD* (Okubo et al. 2020). Clinical reassessment of the patient's phenotype confirmed the BMD diagnosis.

In a duo consisting of an affected mother and daughter (P0859417, P0016160) with AD HSP, we identified a variant in intron 6 of *SPAST* (spastin, MIM ID *604277, NM_014946.4, Chr2: g.32,115,840G>A, c.1004+5G>A) (Fig. 5B). Variants in *SPAST* are known to cause HSP4 (MIM ID #182601; Hazan et al. 1999) and while the same variant has not been previously recorded, a

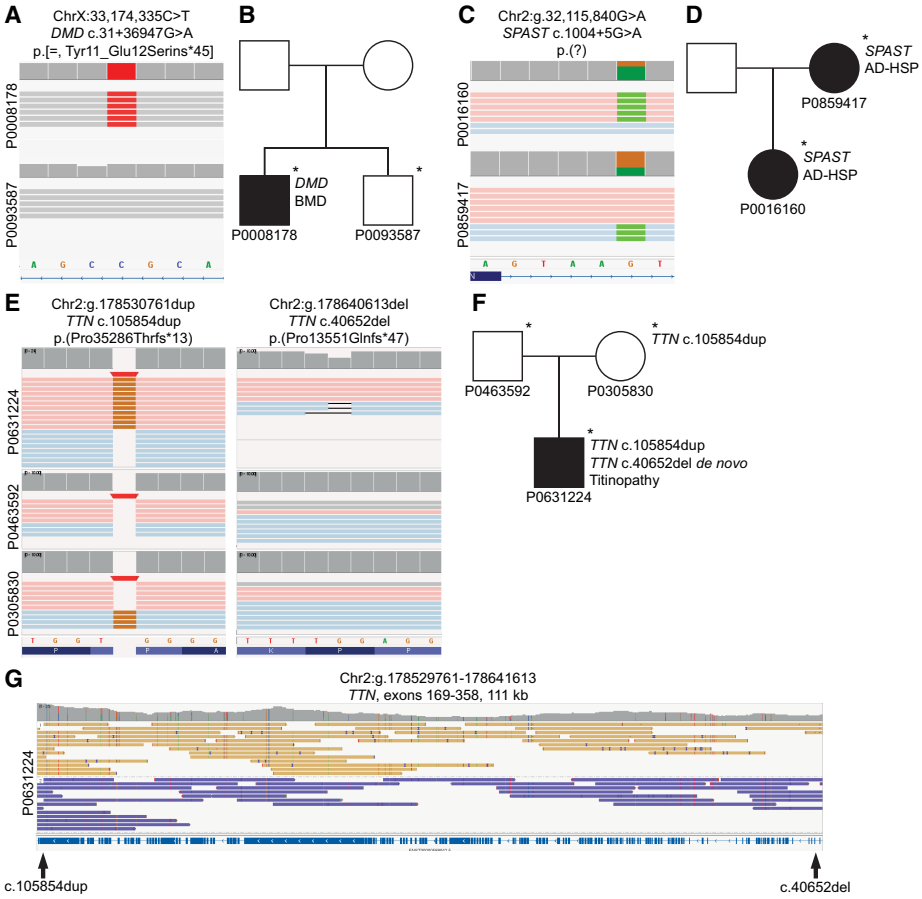

Figure 5. Visualization of disease-causing SNVs and indels in the "unsolved" subcohort in the form of IGV screenshots, along with corresponding pedigrees. In a sporadic patient with suspected titinopathy, we identified a deep-intronic variant in *DMD* (*A*). The nonaffected sibling did not carry the variant (*A*,*B*). In a duo consisting of an affected mother and affected daughter with HSP, we identified a noncanonical splice site variant in *SPAST* (*C*,*D*). In a patient with titinopathy (P063122), a maternally inherited and a de novo variant had been identified earlier (*C*–*F*). The two variants are located 109 kb apart, but the alleles were successfully phased through the entire region by LRS (*G*). The reads are colored by haplotag; pink and light blue, or yellow and purple represent different alleles in *A*, *C*, *E*, and *G*. Unphased reads, such as X-Chromosomal reads in males, are shown in gray (*A*,*E*). Sequenced individuals are marked with an asterisk (*) in the pedigrees (*B*,*D*,*F*). (BMD) Becker muscular dystrophy, (AD-HSP) autosomal dominant hereditary spastic paraplegia.

variant affecting the same base has previously been evaluated as pathogenic in ClinVar (variation ID 989101). The variant was identified in parallel by the referring laboratory but was initially considered to be of uncertain significance. Subsequent RNA analysis eventually demonstrated skipping of exon 6 showing that the variant is likely pathogenic through loss-of-function exon-skipping.

In a sporadic patient with suspected titinopathy (P0631224), two pathogenic frameshift variants in *TTN* (titin, MIM ID *188840, NM_001267550.2) had been previously identified before submission to the Solve-RD collection. Of these, one was maternally inherited, Chr2:g.178530761dup, c.105854dup, p.(Pro35286Thrfs*13), and one de novo, Chr2:g.178640613del, c.40652del, p.(Pro13551Glnfs*47) (Fig. 5C). Both variants are located in ubiquitously expressed exons; the maternally inherited variant affects the constitutional exon 308, and the de novo variant affects exon 221, which is expressed in 99% of *TTN* transcripts (Savarese et al. 2018). Previous SRS efforts had not been successful in identifying on which allele the de novo event had occurred. Using our approach, we were able to successfully differentiate between the alleles and confirmed the two frameshift variants to be in *trans*, thus explanatory for the patient's phenotype (Fig. 5D).

Candidate disease-causing variants identified in rare neurological, neuromuscular, and epilepsy diseases

In addition to the pathogenic variants identified above, in which the disease gene is well established and fits the patient's phenotype according to clinical experts, our analyses revealed novel, candidate disease-causing aberrations in five additional families (Table 1).

De novo duplication on Chromosome X

In a female patient (P0936700) presenting with arthrogryposis multiplex congenita, thoracolumbar scoliosis, and restrictive ventilatory defect, we discovered a 500 kb X-chromosomal tandem duplication (ChrX:g.139,164,887–139,679,311dup), which was confirmed de novo in the patient by gel electrophoresis and sequencing (Fig. 6A–D). The breakpoints of the duplication disrupt two genes: *FGF13* (fibroblast growth factor 13, MIM ID *300070, NM_004114.5), and *MCF2* (MCF.2 cell line derived transforming sequence, *311030, NM_001171876.2).

FGF13 modulates the function and location of voltage-gated sodium channels in the brain (Fry et al. 2021). Mutations in the

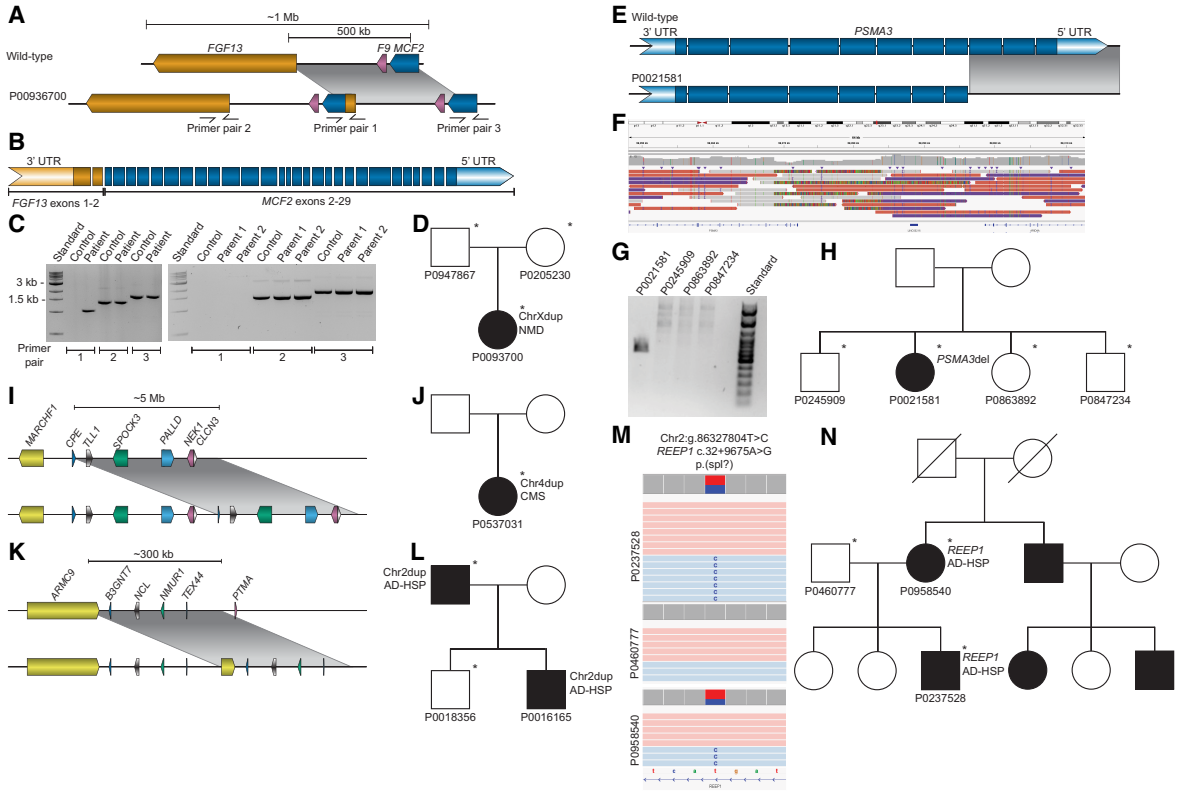

Figure 6. Visualization of candidate disease-causing SVs in the "unsolved" cohort. In a sporadic patient (P0093700) with arthrogryposis multiplex congenita, we detected an X-Chromosomal tandem duplication (*A–D*). The duplication spans from intron 1 of *MCF2* to intron 2 of *FGF13* and also includes F9 (*A*). The result of the duplication is a hypothetical fusion gene, including *FGF13* exons 1–2 and *MCF2* exons 2–29 (*B*). The duplication was validated by PCR and agarose gel electrophoresis (*C*) using primers targeting the breakpoints of the duplication, and a combination of the *MCF2* forward and *FGF13* reverse primers (primer pair 1, *A*,*C*). In a sporadic patient with psychomotor development delay (P0021581), we detected a deletion of *PSMA3* exons 9–11, here shown as a cartoon (*E*) and as a screenshot using IGV (*F*). The deletion was validated by agarose gel electrophoresis (*G*) and Sanger sequencing. The index P0021581 has three healthy siblings who do not carry the deletion (*H*). In a sporadic female patient (P0537031) with congenital malformation syndrome, we detected a 5 Mb tandem duplication on Chromosome 4, visualized here as a cartoon (*I*). The pedigree is shown in *J*. In a sporadic male patient (P0016165) with autosomal dominant spastic paraplegia, with a similarly affected father, we detected a 300 kb tandem duplication on Chromosome 2, visualized here as a cartoon (*K*). The pedigree is shown in *L*. In a family with an affected mother and son (P0847234 and P0958540), we detected an intronic variant in *REEP1* (*M*). The reads are colored by haplotag; pink and light blue represent different alleles in *M*. The index patient also has an affected uncle, whose sample was not sequenced (*N*). Sequenced individuals are marked with an asterisk (*) in the pedigrees (*D*,*H*,*J*,*L*,*N*). (NMD) Neuromuscular disorder, (CMS) congenital malformation syndrome, (AD-HSP) autosomal dominant hereditary spastic paraplegia.

gene have been linked to developmental and epileptic encephalopathy 90 (MIM ID #301058) and intellectual developmental disorder (MIM ID #301095). *MCF2* is an oncogene belonging to a family of GDP-GRP exchange factors, the role of which is to modulate the activity of small GTPases in the Rho family. In addition to brain tissue, it has relatively high expression in the adrenal gland, the testes, and the ovaries. Molinard-Chenu and colleagues reported a putative pathogenic missense mutation in *MCF2* in a patient presenting with complex perisylvian syndrome (Molinard-Chenu et al. 2020).

The duplication results in a hypothetical *FGF13–MCF2* fusion gene, in which the breakpoint resides within the second intron of *FGF13* and the first intron of *MCF2* (Fig. 6B). The entire fusion gene product is in-frame. The putative pathogenic mechanism of this fusion gene will be subject for another study.

PSMA3 carboxy-terminal deletion

In a sporadic female patient (P0021581), we identified a de novo 15.3 kb deletion on Chromosome 14 affecting the three last exons (ex 9–11) of *PSMA3* (proteasome 20S subunit alpha 3, MIM ID *176843, NM_002788.4) (Fig. 6E–H), Chr14:58,268,649–58,283,944. The phenotype consists of a marked delay of psychomotor development resulting in the achievement of independent walking at the age of 3 years. The patient displays facial dysmorphism and marked intellectual disability. From the age of 21, there was a progressive worsening of motor functioning. Nerve conduction studies revealed an axonal sensorimotor neuropathy. Unaffected siblings of the index patient did not carry the deletion, and haplotyping of *PSMA3* suggests that the deletion has arisen de novo in the patient. The deletion breakpoints were confirmed by Sanger sequencing, and its absence in the siblings was confirmed by PCR and gel electrophoresis (Fig. 6G).

PSMA3 is expressed in tissues throughout the body, including skeletal muscle and nerve tissues. As a proteasome subunit, the role of *PSMA3* is to contribute to the proteolytic pathway of aberrant proteins and/or proteins with high turnover rates in the ubiquitin-proteasome system (UPS). Variants in *PSMA3* have not previously been linked to disease and no exonic *PSMA3* deletions have been described in gnomAD SVs v4.1.0. However, variants in genes contributing to the UPS have been linked to several neurodegenerative diseases caused by the aggregation of neurotoxic

proteins in the absence of a functioning UPS. Biran and colleagues have proposed that the *PSMA3* carboxy-terminal region targets intrinsically disordered proteins for degradation, and would thus play an important role in the UPS (Biran et al. 2022). The loss of function observed/expected upper bound fraction (LOEUF) for *PSMA3* is 0.28 (gnomAD v.2.1.1), suggesting that the gene is likely important for normal function. Therefore, the referring clinician chose to submit the patient to the MME platform, making the patient discoverable for potential genetic matches (Philippakis et al. 2015). Also, the *PSMA3* gene has been submitted as a candidate disease gene within the GPAP platform (Laurie et al. 2022).

De novo duplication on Chromosome 4

In a singleton female patient (P0537031) with a congenital malformation syndrome, we identified a 5 Mb tandem de novo duplication on Chromosome 4 (Fig. 6I,J). The patient presented with a complex phenotype involving growth delay, facial syndromic features with optical and neurological involvement, cleft palate, and tonic-clonic seizures. The duplicated sequence is Chr4:165,447,976–170,473,341, and involves several known disease-causing genes, among which *NEK1* (MIM ID *604588), and *CLCN3* (MIM ID *600580).

While none of the known disease-causing genes within the duplicated region can be directly tied back to the phenotype of the patients, some overlap is present. Variants in *NEK1* are a known cause of a form of thoracic dysplasia (short-rib thoracic dysplasia 6 with or without polydactyly, MIM ID #263520). This syndrome involves cleft palate, and enlargement of the lateral ventricles; however, it is also characterized by several clinical manifestations not present in the patient. In turn, missense variants in *CLCN3* are a known cause of autosomal dominant neurodevelopmental disorder with seizures and brain abnormalities (MIM ID #619512).

300 kb duplication on Chromosome 2

In a family presenting with HSP, we identified a 300 kb tandem duplication on Chromosome 2 in the affected father (P0016174) and an affected son (P0016165) (Fig. 6K,L). The nonaffected brother of the son (P0018356) does not carry the duplication. The duplicated sequence is Chr2:231,348,004–231,684,006, with a breakpoint within *ARMC9* (armadillo repeat containing 9, MIM ID *617612, NM_001352754.2) and containing *NCL* (nucleolin, MIM ID *164035, NM_005381.3), among other genes. RNA-seq confirmed upregulation *NCL* within the duplicated sequence.

Nucleolin is a ubiquitously expressed, major nucleolar protein in growing eukaryotic cells, and plays a role in the regulation of ribosomal RNA transcription, ribosome maturation and assembly, and transportation of ribosomal components between the nucleus and cytoplasm. It is predicted to be intolerant to loss-of-function variants (pLI 1.00) and dosage-sensitive (LOEUF 0.18). In addition, variants in *ARMC9* are a known cause of Joubert syndrome 30 (MIM ID #213300), a recessively inherited and genetically heterogeneous neurodevelopmental ciliopathy (Van De Weghe et al. 2017). Individuals with Joubert syndrome present with ataxia, along with hypotonia, abnormal eye movements, and cognitive impairment (Latour et al. 2020).

Deep-intronic SNV in REEP1

In a family with suspected autosomal dominant hereditary spastic paraplegia (AD-HSP), we identified a deep intronic substitution in the first intron of *REEP1* (Chr2:g.86327804T>C, NM_001371279.1:c.32+9675A>G), segregating in the affected mother and son (Fig. 6M,N). Previous genetic analysis with HSP and hereditary neuropathy panels was negative. Loss-of-function variants including splice-altering intronic variants in *REEP1* have previously been reported as causative in AD-HSP families (Züchner et al. 2006).

Discussion

We conducted HiFi long-read genome sequencing for 293 carefully selected patients and healthy relatives from 114 previously undiagnosed rare disease families. Whereas sequencing was performed at a relatively modest coverage of ∼10-fold, we identified and orthogonally validated pathogenic variants of all classes: SNVs, indels, SVs, and STRs. Although our approach is not ideally suited for obtaining highly comprehensive SNV call sets, strict filtering, interpretation, and validation of calls show that previously unidentified and/or misclassified SNVs contribute to the diagnostic yield in our study. In two cases, our study did not identify a novel variant but provided additional information about previously identified candidate variants. In the case of a titinopathy patient, a pathogenic maternally inherited mutation and a single base pair de novo deletion in *TTN* had already been identified. However, the diagnosis was inconclusive because the allele on which the de novo mutation had occurred could not be determined. The long reads of this study allowed for the phasing of this variant and could confirm that it occurred on the paternal allele, thereby leading to a definite diagnosis. Similarly, a previously detected gain involving *MYOT*, was shown here to be a tandem-duplication.

In total, we identified disease-causing variants in 11 out of 93 families with unsolved disorders (11.8%), but only one in the 21 families (4.8%) with "unsolvable" disorders. The variant found in the "unsolvable" trio is a de novo SNV in *TUBA1A* in a patient initially suspected of Aicardi syndrome. The identified *TUBA1A* variant has previously been associated with LIS3 (Bahi-Buisson et al. 2014). In hindsight, and by reverse phenotyping, the clinical experts in this project also confirmed this as the disease-causing genetic variant in this specific case. Individuals with LIS have severe neurological problems, including intellectual disability and epilepsy, and may appear phenotypically very similar to patients with Aicardi syndrome. The difference in the number of resolved cases between the two cohorts suggests that "unsolvable" syndromes indeed are a special class of syndromes. In such cohorts, other explanations for the disorder should be considered, such as methylation defects, somatic mutations, polygenic origin, larger heterogeneity than expected, or even nongenetic causes of disease (Boycott et al. 2018).

Along with the 12 diagnoses, we also identified five candidate disease-causing SVs: one intragenic deletion in *PSMA3*, two large duplications (a 5 Mb event breaking and involving multiple coding genes, and a 300 kb event affecting the *ARMC9* and *NCL* genes), an X-Chromosomal duplication likely leading to the production of an *FGF13–MCF2* fusion protein, and a deep-intronic *REEP1* SNV.

Our study design, aiming primarily at identifying previously hidden variants that could explain disease in expert-selected rare disease patients and families from all across Europe, does not allow for a comprehensive comparison between the technical abilities of SRS and LRS. However, it has been well established that (complex) SVs, inversions, and STRs are difficult to detect or are incompletely characterized with SR-WGS and that LRS improves on this (Höps

et al. 2025). This is also supported by some of the findings from our current study. For example, we identified two pathogenic inversions, a variant type that is especially challenging to detect with SRS due to its copy-neutral nature, as well as four STR expansions among the disease-causing variants in our study.

Similarly, our study does not allow for a direct comparison of diagnostic rates between LRS and SRS because of differences in the underlying diagnostic approaches that samples previously underwent in different medical genetic centers. Moreover, results from such a comparison would not be generalizable to other studies due to the specific patient inclusion criteria, sequencing methods, depth of sequencing, and analytical approach of our study.

One factor currently limiting the diagnostic yield in LRS studies is the clinical interpretation of the large number of identified "rare" SVs. Large catalogs of identified variants from LRS of both affected and unaffected individuals will therefore be of critical importance to improve variant interpretation in such cases. Here, Solve-RD shares the full data set, including expert-curated pedigree and phenotype information as well as a frequency call set of high-quality SVs of the unrelated individuals as a resource for other researchers (Methods; https://github.com/WouterSteyaert/LongReadSequencing.git).

In conclusion, HiFi long-read genome sequencing was conducted for a unique cohort of 293 individuals from 114 previously studied rare disease families. While we did not identify a common genetic cause in any of the "unsolvable" syndromes, we identified causal genetic variants in 11.8% of families from the "unsolved" cohort, and candidate variants in an additional 5.4%, which is comparable to similar efforts (Hiatt et al. 2024). Our study shows the potential and effectiveness of even modest-coverage LRS in rare disease studies.

Methods

Study cohort

HiFi long-read genome sequencing was conducted for 293 individuals from 114 genetically undiagnosed rare disease families (Supplemental Table S1). Patient samples came from two subcohorts: the "unsolvables" ($n = 61$) for which genetic causes remain unknown, and the "unsolved" ($n = 232$) for which a previously hidden genetic variant in a known or yet unknown disease gene is expected to be the cause of disease. All of the patients and healthy relatives were carefully selected by experts from four European Reference Networks: RND (38 families; 95 individuals), EURO-NMD (37 families; 89 individuals), ITHACA (32 families; 88 individuals; including all "unsolvables"), and EpiCARE (seven families; 21 individuals). Depending on the research hypothesis and sample availability 1–7 (un)affected individuals were selected per family for sequencing on a PacBio Sequel IIe instrument. The most represented family structure is the parent-offspring trio ($n_{families} = 42$; $n_{samples} = 126$; 43.0% of cohort). We have used a single SMRT cell of sequencing data per individual which, after read alignment (onto hg38) and read filtering, resulted in a mean HiFi read depth of 9.8 (Supplemental Table S6).

DNA sequencing

Genomic DNA was isolated from peripheral blood according to standard protocol and long-read genome HiFi sequenced using SMRT sequencing technology (Pacific Biosciences, Menlo Park, CA, USA). For every sample, 7–15 µg of DNA was sheared on Megaruptor 2 or 3 (Diagenode, Liège, Belgium) to a target size of 15–18 kb. Libraries were prepared with SMRTbell Prep Kit 2.0 or 3.0. Size selection was performed using a BluePippin DNA size selection system (Sage Science, Beverly, MA, USA) targeting fragments equal to or longer than 10 kb in length. Sequence primer and polymerase were bound to the complex using the Sequel II binding kit 3.2 (PacBio), and sequencing was performed on the Sequel IIe system with 2.0 Chemistry and 30 h movie time per SMRTcell using a single flow cell per sample.

Primary data analysis

All samples were processed in the same fashion using a custom workflow based on standard methods from the Pacific Biosciences analysis pipeline (https://github.com/PacificBiosciences/pb-human-wgs-workflow-snakemake) (Supplemental Fig. S3). Sequencing reads were aligned to the GRCh38 (hg38) genome with pbmm2 (version 1.4.0; Li 2018, 2021) using default parameters. HiFi reads (>QV20) were extracted for all downstream analyzes. Small variant (substitution and indel) calling was performed using DeepVariant (version 1.1.0) with default settings (Poplin et al. 2018). No threshold for maximum size of the indels was applied, and all indel calls were used for further analyses. For parent-offspring trios, GLNexus (version 1.3.1) was used to conduct SNV joint genotyping (Yun et al. 2021).

Small variants were phased using WhatsHap (version 1.1.0) and variants were annotated using an in-house developed pipeline (Martin et al. 2023). This variant annotation was based on the variant effect predictor (VEP V.91) and GENCODE 34 basic gene annotations. STR calling was performed using Tandem Repeat Genotyper (TRGT; version 0.3.3) at 56 known disease-associated STR loci (Supplemental Table S7; Dolzhenko et al. 2024). SV calling was performed using PBSV (version 2.4.0) using default settings with a minimum SV size of 20 bp (https://github.com/PacificBiosciences/pbbioconda). SVs were annotated using AnnotSV (version 3.1.1; Geoffroy et al. 2018).

In each of the 114 RD families that comprise our study cohort, we selected the maximum number of unrelated individuals resulting in a subcohort of 166 unrelated individuals. SVs merging using Jasmine resulted in a call set of 251,672 unique SVs (corresponding to 11,290,783 variant alleles in the subcohort) of which 59,876 are private to one individual (Kirsche et al. 2023). Only 1971 unique SVs (0.78%; 51,433 alleles) in the complete call set affect a coding exon. An additional 2965 unique SVs (1.18%; 111,231 alleles) alter the noncoding sequence of an exon and 95,197 unique SVs (37.8%; 4,445,817 alleles) reside in an intron of a protein-coding gene. Lastly, 35,525 unique SVs (14.1%; 1,638,574 alleles) affect a noncoding gene.

Variant filtering

Structural variants

In parent-offspring trios, we focused on putative de novo variants. For this, we selected sites that are covered by at least eight HiFi reads in each of the members of the trio. Furthermore, at least three HiFi reads should support the variant allele in the child, and only SVs that are unique in the study cohort were retained (Supplemental Table S3). Because of the modest sequencing depth, we subjected all of the resulting SVs to visual inspection using IGV. In this step, we removed SV calls that were unclear in the child (despite the variant call), SVs for which one of the parents had a trace in their sequencing data (for example, supported by one read) and SVs for which both or one of the parents only had one allele sequenced (based on the phased alignments). All of the remaining sites were subjected to primer design for further validation (cf. wet-laboratory validation).

In all of the other family structures, we focused on rare inherited high-quality SVs that cosegregate with disease. To do so, we selected family-unique SV calls, which were observed in all affected members of a given family and absent from all unaffected family members. Furthermore, in at least one of the affected family members, the SVs need to be covered by at least eight HiFi reads of which three support the variant allele. In contrast to SV calls corresponding to well-characterized deletions, inversions, duplications, and insertions, we visually inspected all breakend-calls. We evaluated, based on coverage and on the complexity of the sequence context, whether or not a breakend-call could, together with the linked breakend-call, be a signature of a genetic event that is too large to be characterized as a deletion, inversion, duplication, or insertion by pbsv. In this step, we required that all clipped reads support the same regional split. Since these calls support relatively large genetic events (>100 kb), we clinically assessed them in the complete human genome. In contrast, clinical interpretation of SV calls corresponding to characterized deletions, inversions, duplications, and insertions (size <100 kb) was restricted to events that reside in genes within recently curated ERN-specific gene lists (Laurie et al. 2025).

Single-nucleotide variants

In parent-offspring trios, we focused on putative de novo events. These were selected from the joint calls generated by GLnexus. We considered a variant to be putative de novo when the child is heterozygous (genotype "0/1") and both parents are homozygous for the reference allele (genotype "0/0"). In addition, we require that both parents and the child have ≥8 HiFi reads covering the site of which three reads support the variant allele in the child.

In all other families, we selected for rare inherited SNVs that cosegregate with disease. For this, we selected SNVs that are unique to a single family that are present in all affected family members and absent from all unaffected family members. In contrast to the de novo variant interpretation, we restricted variant interpretation for inherited variants to variants that reside in genes incorporated in the recently curated ERN-specific gene lists.

STR genotypes

STR genotypes were visualized in R per submitter group in comparison with the rest of the cohort in order to facilitate the evaluation of quality of calling per locus and the detection of pathogenically expanded alleles (R Core Team 2022). These results were sent to the groups for clinical interpretation.

Wet-laboratory variant validation

Altogether, 35 variants called as de novo were selected for validation using targeted LRS (Supplemental Fig. S1). Primers for the validations were designed using the online Primer3 design tool as per the manufacturer's suggestions. Primers were selected to be 18–21 nt in length with a GC content ranging from 40% to 60%. While an annealing temperature of 60°C was proposed to be optimal, annealing temperatures between 57°C and 61°C were considered to be acceptable as well. Sizes of the products ranged between 1000 and 4000 nt, to ensure capture of the full region and compatibility with PacBio LRS.

In three cases, two adjacent SVs could be covered by one primer pair, and for the large X-Chromosomal duplication, altogether three primer pairs were designed (Fig. 6). For nine variants, the primer design was not possible, resulting in a total number of 23 primer pairs designed for de novo variants (Supplemental Table S5). In addition to these, primers were also designed for the inherited candidate exon 6 deletion in *REEP1* and a 50 bp deletion in *MAPK8IP3* segregating with disease in P0016368 and P0018504.

All successfully amplified patient samples were validated by targeted LRS. Subsequent sequencing of parental samples was performed as per the workflow above for samples in which the variant call was confirmed in the index.

Data access

All raw and processed sequencing data generated in this study have been submitted to the European Genome-phenome Archive (EGA; https://ega-archive.org/) under accession number EGAD00001008602. Confirmed causative variants were submitted to ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/) under accession numbers SCV005849075–SCV005849079 and SCV005871627–SCV005871633.

Competing interest statement

The authors declare no competing interests.

Acknowledgments

The Solve-RD consortium is grateful to all involved RD patients and their families as well as other contributors to Solve-RD. The Solve-RD project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement no. 779257. This research is supported (not financially) by several ERNs: ERN on Intellectual Disability, TeleHealth, Autism and Congenital Anomalies (ERN ITHACA)—Project ID no. 101085231; ERN on Rare Neurological Diseases (ERN RND)—Project ID no. 101155994; ERN for Neuromuscular Diseases (ERN Euro-NMD)—Project ID no. 101156434; and ERN for Rare and Complex Epilepsies (ERN EpiCARE)—Project ID no. 101156811. The ERNs are co-funded by the European Union within the framework of the Third Health Program. We would also like to thank all other Solve-RD colleagues that were not mentioned by name in the author list, including members of the Solve-RD data interpretation task force (DATF), and other members of ERNs and DITFs. We also thank The Radboud Technology Center Genomics for the library preparation and sequencing of all samples. V.A.Y. and J.G. received funding from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) via the project NFDI 1/1 "GHGA - German Human Genome-Phenome Archive" (no. 441914366). L.Sa. received funding from the Sigrid Jusélius Foundation (Fellowship No 220540). R.S. received funding from the Bundesministerium für Bildung und Forschung (BMBF) through funding for the TreatHSP network (grant 01GM2209A) and the National Institute of Neurological Disorders and Stroke (NINDS) under Award Number R01NS072248. H.H. was supported by the DFG (HE8803/1–1 to H.H.). H.L. receives support from the Canadian Institutes of Health Research (CIHR) for Foundation grant FDN-167281 (Precision Health for Neuromuscular Diseases), Transnational Team grant ERT-174211 (ProDGNE) and Network grant OR2-189333 (NMD4C), from the Canada Foundation for Innovation (CFI-JELF 38412), the Canada Research Chairs program (Canada Research Chair in Neuromuscular Genomics and Health, 950-232279), the European Commission (grant no. 101080249) and the Canada Research Coordinating Committee New Frontiers in Research Fund (NFRFG-2022-00033) for SIMPATHIC, and from the Government of Canada's Canada First Research Excellence Fund (CFREF) for the Brain-Heart Interconnectome (CFREF-2022-00007). A.H. was supported by a ZonMW (The Netherlands

Organization for Health Research and Development) Vici grant (No. 09150182310053).

Author contributions: W.S., L.Sa., G.D., V.A.Y., R.D., M.W., A.d.O., S.v.d.H., H.S., N.Z., M.St., L.O., K.N., J.X., L.V.d.V., H.M., J.M., K.P., E.R., and M.Sa. generated, analyzed, and/or interpreted data. K.E., G.A., L.M., S.L., H.G.B., H.G., S.B., S.O., L.E.L.M.V., C.G., and A.H. coordinated data sharing. J.X., L.V.d.V., H.M., J.M., A.E.-C., J.G., R.S., M.Sy., D.B., H.H., L.Sc., K.G.C., J.B., A.F., R.S., M.S.A.E., V.S., V.P., L.P., A.S., I.Z., F.M., E.B., R.H., H.L., and M.Sp. recruited and enrolled patients and provided clinical assessment. C.G. and A.H. designed the study and oversaw the interpretation of data. W.S., L.Sa., C.G., and A.H. drafted and revised the manuscript. All authors read and approved the final version of the manuscript.

References

Bahi-Buisson N, Poirier K, Fourniol F, Saillour Y, Valence S, Lebrun N, Hully M, Fallet Bianco C, Boddaert N, Elie C, et al. 2014. The wide spectrum of tubulinopathies: what are the key features for the diagnosis? *Brain* **137:** 1676–1700. doi:10.1093/brain/awu082

Beetz C, Schüle R, Deconinck T, Tran-Viet K-N, Zhu H, Kremer BPH, Frints SGM, van Zelst-Stams WAG, Byrne P, Otto S, et al. 2008. REEP1 mutation spectrum and genotype/phenotype correlation in hereditary spastic paraplegia type 31. *Brain* **131:** 1078–1086. doi:10.1093/brain/awn026

Beyter D, Ingimundardottir H, Oddsson A, Eggertsson HP, Bjornsson E, Jonsson H, Atlason BA, Kristmundsdottir S, Mehringer S, Hardarson MT, et al. 2021. Long-read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits. *Nat Genet* **53:** 779–786. doi:10.1038/s41588-021-00865-4

Biesecker LG, Green RC. 2014. Diagnostic clinical genome and exome sequencing. *N Engl J Med* **370:** 2418–2425. doi:10.1056/NEJMra1312543

Biran A, Myers N, Steinberger S, Adler J, Riutin M, Broennimann K, Reuven N, Shaul Y. 2022. The C-terminus of the PSMA3 proteasome subunit preferentially traps intrinsically disordered proteins for degradation. *Cells* **11:** 3231. doi:10.3390/cells11203231

Boycott KM, Dyment DA, Innes AM. 2018. Unsolved recognizable patterns of human malformation: challenges and opportunities. *Am J Med Genet C Semin Med Genet* **178:** 382–386. doi:10.1002/ajmg.c.31665

Chaisson MJP, Sanders AD, Zhao X, Malhotra A, Porubsky D, Rausch T, Gardner EJ, Rodriguez OL, Guo L, Collins RL, et al. 2019. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat Commun* **10:** 1784. doi:10.1038/s41467-018-08148-z

Chintalaphani SR, Pineda SS, Deveson IW, Kumar KR. 2021. An update on the neurological short tandem repeat expansion disorders and the emergence of long-read sequencing diagnostics. *Acta Neuropathol Commun* **9:** 98. doi:10.1186/s40478-021-01201-x

Dolzhenko E, English A, Dashnow H, De Sena Brandine G, Mokveld T, Rowell WJ, Karniski C, Kronenberg Z, Danzi MC, Cheung WA, et al. 2024. Characterization and visualization of tandem repeats at genome scale. *Nat Biotechnol* **42:** 1606–1614. doi:10.1038/s41587-023-02057-3

Fadaie Z, Neveling K, Mantere T, Derks R, Haer-Wigman L, den Ouden A, Kwint M, O'Gorman L, Valkenburg D, Hoyng CB, et al. 2021. Long-read technologies identify a hidden inverted duplication in a family with choroideremia. *HGG Adv* **2:** 100046. doi:10.1016/j.xhgg.2021.100046

Fry AE, Marra C, Derrick AV, Pickrell WO, Higgins AT, te Water Naude J, McClatchey MA, Davies SJ, Metcalfe KA, Tan HJ, et al. 2021. Missense variants in the N-terminal domain of the A isoform of FHF2/FGF13 cause an X-linked developmental and epileptic encephalopathy. *Am J Hum Genet* **108:** 176–185. doi:10.1016/j.ajhg.2020.10.017

Geoffroy V, Herenger Y, Kress A, Stoetzel C, Piton A, Dollfus H, Muller J. 2018. AnnotSV: an integrated tool for structural variations annotation. *Bioinformatics* **34:** 3572–3574. doi:10.1093/bioinformatics/bty304

Goizet C, Depienne C, Benard G, Boukhris A, Mundwiller E, Solé G, Coupry I, Pilliod J, Martin-Négrier M-L, Fedirko E, et al. 2011. REEP1 mutations in SPG31: frequency, mutational spectrum, and potential association with mitochondrial morpho-functional dysfunction. *Hum Mutat* **32:** 1118–1127. doi:10.1002/humu.21542

Hazan J, Fonknechten N, Mavel D, Paternotte C, Samson D, Artiguenave F, Davoine C-S, Cruaud C, Dürr A, Wincker P, et al. 1999. Spastin, a new AAA protein, is altered in the most frequent form of autosomal dominant spastic paraplegia. *Nat Genet* **23:** 296–303. doi:10.1038/15472

Hiatt SM, Lawlor JMJ, Handley LH, Latner DR, Bonnstetter ZT, Finnila CR, Thompson ML, Boston LB, Williams M, Rodriguez Nunez I, et al. 2024. Long-read genome sequencing and variant reanalysis increase diagnostic yield in neurodevelopmental disorders. *Genome Res* **34:** 1747–1762. doi:10.1101/gr.279227.124

Höps W, Weiss MM, Derks R, Galbany JC, den Ouden A, van den Heuvel S, Timmermans R, Smits J, Mokveld T, Dolzhenko E, et al. 2025. Hifi long-read genomes for difficult-to-detect, clinically relevant variants. *Am J Hum Genet* **112:** 450–456. doi:10.1016/j.ajhg.2024.12.013

Ishiura H, Takahashi Y, Hayashi T, Saito K, Furuya H, Watanabe M, Murata M, Suzuki M, Sugiura A, Sawai S, et al. 2014. Molecular epidemiology and clinical spectrum of hereditary spastic paraplegia in the Japanese population based on comprehensive mutational analyses. *J Hum Genet* **59:** 163–172. doi:10.1038/jhg.2013.139

Kirsche M, Prabhu G, Sherman R, Ni B, Battle A, Aganezov S, Schatz MC. 2023. Jasmine and Iris: population-scale structural variant comparison and analysis. *Nat Methods* **20:** 408–417. doi:10.1038/s41592-022-01753-3

Kobayashi H, Abe K, Matsuura T, Ikeda Y, Hitomi T, Akechi Y, Habu T, Liu W, Okuda H, Koizumi A. 2011. Expansion of intronic GGCCTG hexanucleotide repeat in *NOP56* causes SCA36, a type of spinocerebellar ataxia accompanied by motor neuron involvement. *Am J Hum Genet* **89:** 121–130. doi:10.1016/j.ajhg.2011.05.015

Kucuk E, van der Sanden BPGH, O'Gorman L, Kwint M, Derks R, Wenger AM, Lambert C, Chakraborty S, Baybayan P, Rowell WJ, et al. 2023. Comprehensive de novo mutation discovery with HiFi long-read sequencing. *Genome Med* **15:** 34. doi:10.1186/s13073-023-01183-6

Latour BL, Van De Weghe JC, Rusterholz TDS, Letteboer SJF, Gomez A, Shaheen R, Gesemann M, Karamzade A, Asadollahi M, Barroso-Gil M, et al. 2020. Dysfunction of the ciliary ARMC9/TOGARAM1 protein module causes Joubert syndrome. *J Clin Invest* **130:** 4423–4439. doi:10.1172/JCI131656

Laurie S, Piscia D, Matalonga L, Corvó A, Fernández-Callejo M, Garcia-Linares C, Hernandez-Ferrer C, Luengo C, Martínez I, Papakonstantinou A, et al. 2022. The RD-connect genome-phenome analysis platform: accelerating diagnosis, research, and gene discovery for rare diseases. *Hum Mutat* **43:** 717–733. doi:10.1002/humu.24353

Laurie S, Steyaert W, de Boer E, Polavarapu K, Schuermans N, Sommer AK, Demidov G, Ellwanger K, Paramonov I, Thomas C, et al. 2025. Genomic reanalysis of a pan-European rare-disease resource yields new diagnoses. *Nat Med* **31:** 478–489. doi:10.1038/s41591-024-03420-w

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34:** 3094–3100. doi:10.1093/bioinformatics/bty191

Li H. 2021. New strategies to improve minimap2 alignment accuracy. *Bioinformatics* **37:** 4572–4574. doi:10.1093/bioinformatics/btab705

Martin M, Ebert P, Marschall T. 2023. Read-based phasing and analysis of phased variants with WhatsHap. *Methods Mol Biol* **2590:** 127–138. doi:10.1007/978-1-0716-2819-5_8

Matilla-Dueñas A, Volpini V. 1993. *Spinocerebellar Ataxia type 37*. University of Washington, Seattle, WA. http://europepmc.org/abstract/MED/31145571.

Merker JD, Wenger AM, Sneddon T, Grove M, Zappala Z, Fresard L, Waggott D, Utiramerur S, Hou Y, Smith KS, et al. 2018. Long-read genome sequencing identifies causal structural variation in a Mendelian disease. *Genet Med* **20:** 159–163. doi:10.1038/gim.2017.86

Mizuguchi T, Suzuki T, Abe C, Umemura A, Tokunaga K, Kawai Y, Nakamura M, Nagasaki M, Kinoshita K, Okamura Y, et al. 2019a. A 12-kb structural variation in progressive myoclonic epilepsy was newly identified by long-read whole-genome sequencing. *J Hum Genet* **64:** 359–368. doi:10.1038/s10038-019-0569-5

Mizuguchi T, Toyota T, Adachi H, Miyake N, Matsumoto N, Miyatake S. 2019b. Detecting a long insertion variant in SAMD12 by SMRT sequencing: implications of long-read whole-genome sequencing for repeat expansion diseases. *J Hum Genet* **64:** 191–197. doi:10.1038/s10038-018-0551-7

Molinard-Chenu A, Fluss J, Laurent S, Laurent M, Guipponi M, Dayer AG. 2020. *MCF2* is linked to a complex perisylvian syndrome and affects cortical lamination. *Ann Clin Transl Neurol* **7:** 121–125. doi:10.1002/acn3.50949

Nguengang Wakap S, Lambert DM, Olry A, Rodwell C, Gueydan C, Lanneau V, Murphy D, Le Cam Y, Rath A. 2020. Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. *Eur J Hum Genet* **28:** 165–173. doi:10.1038/s41431-019-0508-0

Obayashi M, Stevanin G, Synofzik M, Monin M-L, Duyckaerts C, Sato N, Streichenberger N, Vighetto A, Desestret V, Tesson C, et al. 2015. Spinocerebellar ataxia type 36 exists in diverse populations and can be caused by a short hexanucleotide GGCCTG repeat expansion. *J Neurol Neurosurg Psychiatry* **86:** 986–995. doi:10.1136/jnnp-2014-309153

Okubo M, Noguchi S, Hayashi S, Nakamura H, Komaki H, Matsuo M, Nishino I. 2020. Exon skipping induced by nonsense/frameshift mutations in DMD gene results in becker muscular dystrophy. *Hum Genet* **139:** 247–255. doi:10.1007/s00439-019-02107-4

Pauper M, Kucuk E, Wenger AM, Chakraborty S, Baybayan P, Kwint M, van der Sanden B, Nelen MR, Derks R, Brunner HG, et al. 2021. Long-read trio sequencing of individuals with unsolved intellectual disability. *Eur J Hum Genet* **29:** 637–648. doi:10.1038/s41431-020-00770-0

Philippakis AA, Azzariti DR, Beltran S, Brookes AJ, Brownstein CA, Brudno M, Brunner HG, Buske OJ, Carey K, Doll C, et al. 2015. The matchmaker exchange: a platform for rare disease gene discovery. *Hum Mutat* **36:** 915–921. doi:10.1002/HUMU.22858

Pogue RE, Cavalcanti DP, Shanker S, Andrade RV, Aguiar LR, de Carvalho JL, Costa FF. 2018. Rare genetic diseases: update on diagnosis, treatment and online resources. *Drug Discov Today* **23:** 187–195. doi:10.1016/j.drudis.2017.11.002

Poplin R, Chang P-C, Alexander D, Schwartz S, Colthurst T, Ku A, Newburger D, Dijamco J, Nguyen N, Afshar PT, et al. 2018. A universal SNP and small-indel variant caller using deep neural networks. *Nat Biotechnol* **36:** 983–987. doi:10.1038/nbt.4235

Porubsky D, Vollger MR, Harvey WT, Rozanski AN, Ebert P, Hickey G, Hasenfeld P, Sanders AD, Stober C, Consortium HPR, et al. 2023. Gaps and complex structurally variant loci in phased genome assemblies. *Genome Res* **33:** 496–510. doi:10.1101/gr.277334.122

R Core Team. 2022. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/.

Sanchis-Juan A, Stephens J, French CE, Gleadall N, Mégy K, Penkett C, Shamardina O, Stirrups K, Delon I, Dewhurst E, et al. 2018. Complex structural variants in Mendelian disorders: identification and breakpoint resolution using short- and long-read genome sequencing. *Genome Med* **10:** 95. doi:10.1186/s13073-018-0606-6

Savarese M, Jonson PH, Huovinen S, Paulin L, Auvinen P, Udd B, Hackman P. 2018. The complexity of titin splicing pattern in human adult skeletal muscles. *Skelet Muscle* **8:** 11. doi:10.1186/s13395-018-0156-z

Sernadela P, González-Castro L, Carta C, van der Horst E, Lopes P, Kaliyaperumal R, Thompson M, Thompson R, Queralt-Rosinach N, Lopez E, et al. 2017. Linked registries: connecting rare diseases patient registries through a Semantic Web layer. *Biomed Res Int* **2017:** 8327980. doi:10.1155/2017/8327980

Spinazzi M, Savarese M, Letournel F, Sagath L, Manero F, Guichet A, Hoischen A, Metay C, Gouju J, Udd B. 2025. Myotilin gene duplication causing late-onset myotilinopathy. *Eur J Neurol* **32:** e70029. doi:10.1111/ene.70029

Van De Weghe JC, Rusterholz TDS, Latour B, Grout ME, Aldinger KA, Shaheen R, Dempsey JC, Maddirevula S, Cheng Y-HH, Phelps IG, et al. 2017. Mutations in *ARMC9*, which encodes a basal body protein, cause Joubert syndrome in humans and ciliopathy phenotypes in zebrafish. *Am J Hum Genet* **101:** 23–36. doi:10.1016/j.ajhg.2017.05.010

Wright CF, FitzPatrick DR, Firth HV. 2018. Paediatric genomics: diagnosing rare disease in children. *Nat Rev Genet* **19:** 253–268. doi:10.1038/nrg.2017.116

Yun T, Li H, Chang P-C, Lin MF, Carroll A, McLean CY. 2021. Accurate, scalable cohort variant calls using DeepVariant and GLnexus. *Bioinformatics* **36:** 5582–5589. doi:10.1093/bioinformatics/btaa1081

Zeng S, Zhang M-Y, Wang X-J, Hu Z-M, Li J-C, Li N, Wang J-L, Liang F, Yang Q, Liu Q, et al. 2019. Long-read sequencing identified intronic repeat expansions in *SAMD12* from Chinese pedigrees affected with familial cortical myoclonic tremor with epilepsy. *J Med Genet* **56:** 265–270. doi:10.1136/jmedgenet-2018-105484

Zook JM, Hansen NF, Olson ND, Chapman L, Mullikin JC, Xiao C, Sherry S, Koren S, Phillippy AM, Boutros PC, et al. 2020. A robust benchmark for detection of germline large deletions and insertions. *Nat Biotechnol* **38:** 1347–1355. doi:10.1038/s41587-020-0538-8

Züchner S, Wang G, Tran-Viet K-N, Nance MA, Gaskell PC, Vance JM, Ashley-Koch AE, Pericak-Vance MA. 2006. Mutations in the novel mitochondrial protein REEP1 cause hereditary spastic paraplegia type 31. *Am J Hum Genet* **79:** 365–369. doi:10.1086/505361

Received March 28, 2024; accepted in revised form February 21, 2025.

Unraveling undiagnosed rare disease cases by HiFi long-read genome sequencing

Wouter Steyaert, Lydia Sagath, German Demidov, et al.

Genome Res. 2025 35: 755-768 originally published online March 26, 2025
Access the most recent version at doi:10.1101/gr.279414.124