nature medicine

Article

https://doi.org/10.1038/s41591-024-03420-w

Genomic reanalysis of a pan-European rare-disease resource yields new diagnoses

Received: 1 May 2023

Accepted: 14 November 2024

Published online: 17 January 2025

Check for updates

A list of authors and their affiliations appears at the end of the paper

Genetic diagnosis of rare diseases requires accurate identification and interpretation of genomic variants. Clinical and molecular scientists from 37 expert centers across Europe created the Solve-Rare Diseases Consortium (Solve-RD) resource, encompassing clinical, pedigree and genomic rare-disease data (94.5% exomes, 5.5% genomes), and performed systematic reanalysis for 6,447 individuals (3,592 male, 2,855 female) with previously undiagnosed rare diseases from 6,004 families. We established a collaborative, two-level expert review infrastructure that allowed a genetic diagnosis in 506 (8.4%) families. Of 552 disease-causing variants identified, 464 (84.1%) were single-nucleotide variants or short insertions/deletions. These variants were either located in recently published novel disease genes ($n = 67$), recently reclassified in ClinVar ($n = 187$) or reclassified by consensus expert decision within Solve-RD ($n = 210$). Bespoke bioinformatics analyses identified the remaining 15.9% of causative variants ($n = 88$). Ad hoc expert review, parallel to the systematic reanalysis, diagnosed 249 (4.1%) additional families for an overall diagnostic yield of 12.6%. The infrastructure and collaborative networks set up by Solve-RD can serve as a blueprint for future further scalable international efforts. The resource is open to the global rare-disease community, allowing phenotype, variant and gene queries, as well as genome-wide discoveries.

While the definition of what constitutes a rare disease is arbitrary, and thus varies by jurisdiction, the European Union has adopted a definition of a rare disease as being an ailment that affects <50 individuals per 100,000. More than 70% of the >6,000 unique rare diseases are genetic and, collectively, they constitute a major health issue, with 3.5–6.0% of individuals affected by a rare disease over their lifetime[1].

Despite improvements in diagnostics and research options for rare diseases, many individuals remain without a molecularly proven genetic diagnosis. In healthcare systems, where exome or genome sequencing is becoming the standard of care, diagnostic yield varies between 20 and 70% depending on the type of rare disease, inclusion criteria, sequencing strategy and analysis standards, as highlighted by projects such as The 100,000 Genomes Project via Genomics England, and the Deciphering Developmental Disorders Study[2–4].

As reviewed in Dai et al.[5], it has been shown that reanalysis of existing genomic data can lead to novel diagnoses, both as a result of newly described disease genes and due to improvements in the identification, annotation and interpretation of genomic variants. However, reanalysis of such data is not routinely undertaken due to the time and multidisciplinary expertise required, and associated costs.

In 2017, the European Union brought together expertise on rare diseases into 24 thematic European Reference Networks (ERNs). Each ERN has multiple national centers across the 27 member states, all of which have been vetted for their clinical, diagnostic and research expertise. These collaborations provide a pan-European framework to improve care for individuals with rare diseases.

Solve-RD is a pan-European omics project that brings together (1) clinicians, geneticists and translational researchers from four ERNs, including rare neurological diseases (RND, https://www.ern-rnd.eu/),

e-mail: sergi.beltran@cnag.eu; alexander.hoischen@radboudumc.nl

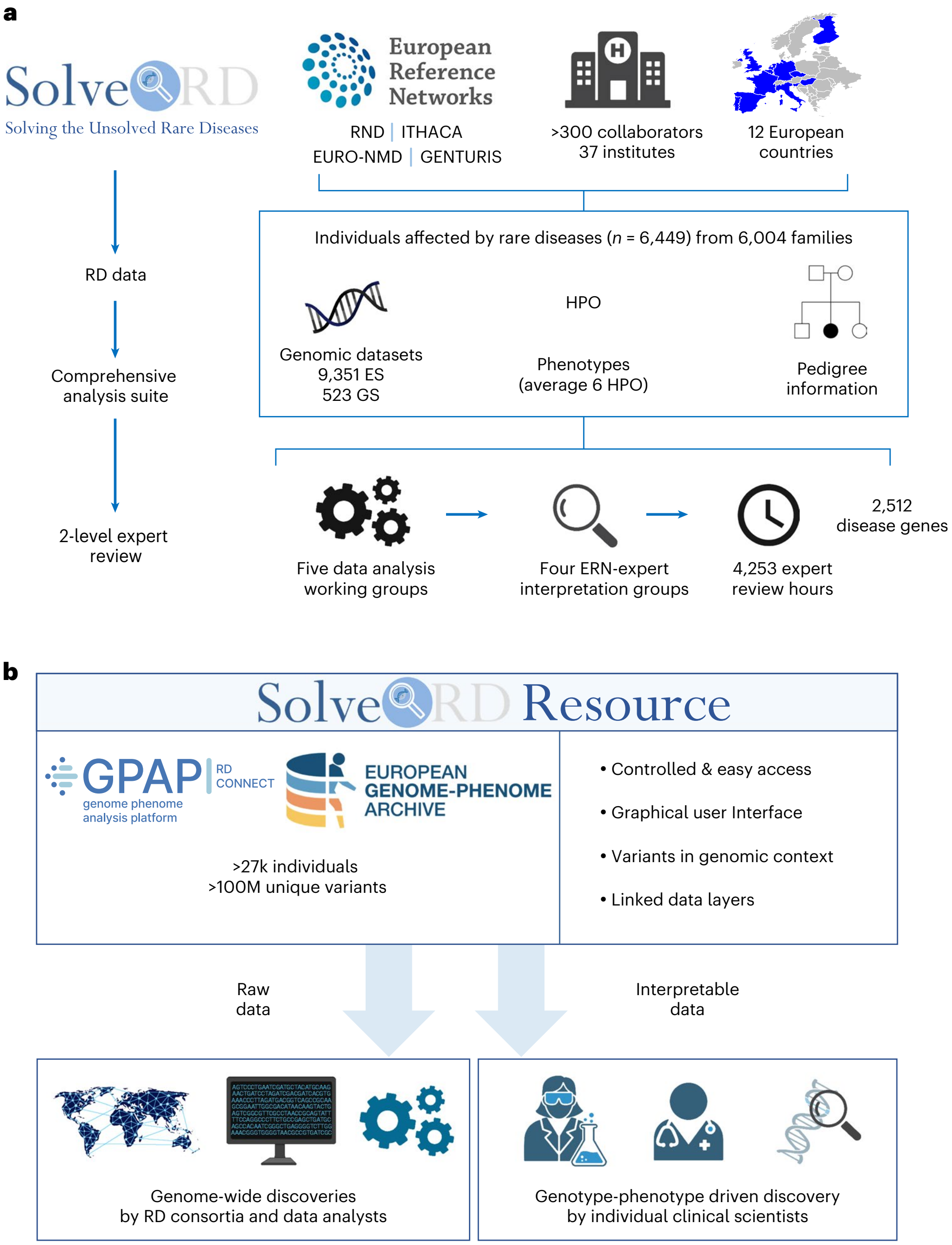

a
Solve RD
Solving the Unsolved Rare Diseases
European Reference Networks
RND | ITHACA
EURO-NMD | GENTURIS
>300 collaborators
37 institutes
12 European countries
RD data
Comprehensive analysis suite
2-level expert review
Individuals affected by rare diseases (n = 6,449) from 6,004 families
Genomic datasets
9,351 ES
523 GS
HPO
Phenotypes (average 6 HPO)
Pedigree information
Five data analysis working groups
Four ERN-expert interpretation groups
4,253 expert review hours
2,512 disease genes
b
Solve RD Resource
GPAP RD CONNECT
genome phenome analysis platform
EUROPEAN GENOME-PHENOME ARCHIVE
>27k individuals
>100M unique variants
• Controlled & easy access
• Graphical user Interface
• Variants in genomic context
• Linked data layers
Raw data
Interpretable data
Genome-wide discoveries by RD consortia and data analysts
Genotype-phenotype driven discovery by individual clinical scientists

Fig. 1 | Overview of the Solve-RD analysis and interpretation framework and community resource established. a, Solve-RD brought together rare-disease data and expertise. Central to Solve-RD are four core ERNs relating to rare diseases; via these expert disease networks, patients with rare diseases were recruited from 43 research groups from 37 institutes in 12 European countries (Belgium, Czech Republic, Finland, France, Germany, Hungary, Italy, the Netherlands, Portugal, Slovenia, Spain and the United Kingdom) and Canada. The work involved >300 collaborators in the submission, analysis and interpretation of rare-disease data. The RD-REAL framework allows sharing of data and expertise on a continental scale, consisting of (1) expert curated data, (2) a comprehensive analysis suite and (3) a two-level (that is, molecular and clinical) expert review. The complete dataset comprises 9,645 individuals from 6,004 families and includes phenotypes in Phenopacket format (average of six HPO terms per affected individual), pedigrees and genomic data (genomes and exomes). **b**, Illustration of the utility of this resource to the global rare-disease community. In total, RD-REAL data of >23,000 individuals with >100 million unique genomic variants are available via RD-Connect GPAP and EGA. This represents a growing resource containing data that have been submitted since the start of Solve-RD. Interpretable data (genetic variants, phenotypes and pedigrees) are standardized and annotated, and are made available for querying, analysis and interpretation in RD-Connect GPAP for authorized users. In addition, all raw and processed data are available for download at EGA under a controlled-access model. All icons, except logos of services (GPAP; EGA) and consortia/networks (Solve-RD; European Reference Networks) that are contributors of this publication, created with Biorender.com.

intellectual disability, telehealth and congenital anomalies (ITHACA, https://ern-ithaca.eu/), neuromuscular diseases (EURO-NMD, https://ern-euro-nmd.eu/) and genetic tumor risk syndromes (GENTURIS, https://www.genturis.eu), as well as the Spanish undiagnosed disease program[6]; (2) patient organizations represented by EURORDIS[7] (https://www.eurordis.org/); (3) genomic data-sharing and -analysis resources, such as the RD-Connect Genome-Phenome Analysis Platform[8] (RD-Connect GPAP, https://platform.rd-connect.eu/) and the European Genome-Phenome Archive[9] (EGA, https://ega-archive.org/); (4) European networks aiming to improve and harmonize the quality of

Table 1 | Solve-RD reanalysis data

Solve-RD RD-REAL data	ERN RND	ERN ITHACA	ERN EURO-NMD	ERN GENTURIS	Sum across ERNs
Experiments (exomes/genomes)	2,852 (2,692/160)	4,470 (4,231/239)	2,162 (2,059/103)	390 (369/21)	9,874
Participants (affected individuals)	2,799 (2,453)	4,331 (1,933)	2,125 (1,685)	390 (378)	9,645 (6,449)
Families	2,271	1,857	1,517	359	6,004
Diagnosed probands (systematic reanalysis) (%)	242 (10.7)	158 (8.5)	96 (6.3)	10 (2.8)	506 (8.4)
Diagnosed probands (ad hoc expert review) (%)	61 (2.7)	145 (7.8)	42 (2.8)	1 (0.3)	249 (4.1)
Probands with 'candidate diagnoses' (%)	119 (5.2)	139 (7.5)	41 (2.7)	45 (12.5)	344 (5.7)

Number of datasets following quality control filtering (Methods), representing the number of previously undiagnosed families/probands. Numbers are given for the entire project and for each ERN separately. We provide the overall yield of newly diagnosed rare-disease cases for both the multicenter systematic reanalysis and the parallel ad hoc expert review. The table also indicates the number of (likely) pathogenic variants that led to candidate diagnoses.

genetic testing services, such as EuroGentest (http://www.eurogentest.org/); and (5) experts in the field of omics technologies, bioinformatics, knowledge management and rare-disease ontology, such as Orphanet Rare Disease Ontology (ORDO, https://www.orphadata.com/ontologies/) and Human Phenotype Ontology (HPO)[10].

One of the core aims of Solve-RD is to improve the rate of genetic diagnosis for individuals affected by a rare disease. A specific objective of Solve-RD is to systematically collate and reanalyze existing exome/genome datasets and corresponding structured ontology-based phenotype and pedigree information across the disease areas of its ERN partners (Fig. 1). Previous pilot studies analyzed only subcohorts and focussed on established pathogenic (ClinVar) variants, whereas the work presented here is the primary large-scale and systematic reanalysis across all diseases of Solve-RD[7,11–13]. Here we report the results from the systematic reanalysis of data from 6,004 undiagnosed rare-disease families recruited from across Europe by Solve-RD. The entire dataset is available as a resource for the global rare-disease research community.

Results

Pan-European rare-disease data collection

Solve-RD involves over 300 clinicians, laboratory geneticists and translational researchers from 43 research groups associated with 37 institutes located in 12 European countries and Canada. In total, we collected 10,276 genomic datasets, as well as phenotypic descriptions and pedigrees, from 10,039 individuals, all previously analyzed through local diagnostic or research efforts. The collection includes 554 genomes and 9,722 exomes enriched using 28 different exome-enrichment kits and generated on several short-read sequencing platforms. Following quality control (Methods), 9,874 datasets (523 genomes and 9,351 exomes) from 9,645 individuals remained. These represent 6,449 individuals affected by rare diseases, and 3,196 unaffected relatives, from 6,004 families (Fig. 1, Table 1 and Supplementary Table 1). Disease categories comprise rare neurological diseases (RND, $n = 2,271$ families), (multiple) malformation syndromes, intellectual disability and other neurodevelopmental disorders (ITHACA and SpainUDP, $n = 1,857$), rare neuromuscular diseases (EURO-NMD, $n = 1,517$) and suspected hereditary gastric and bowel cancer (GENTURIS, $n = 359$).

Phenotypic information was collected using standardized HPO terms, consistent with the GA4GH Phenopacket schema[14], with a median of six terms (range 0–74) assigned per affected individual (Extended Data Fig. 1), varying from a median of four terms for GENTURIS to ten for ITHACA, reflecting the phenotypic complexity of probands affected by the respective rare disease. In addition, for 2,126 (35.4%) probands, a clinical diagnosis was encoded using an ORDO ORPHA code[15], of which 338 were unique.

New genetic diagnoses following systematic reanalysis

A two-level expert analysis strategy (data-expert and clinical-expert levels) was applied, as detailed in Methods. All datasets were reanalyzed for a broad range of genomic variants, including SNVs and short insertions–deletions (InDels), noncanonical splice variants predicted in silico, homoplasmic and heteroplasmic mitochondrial DNA variants, copy number variants (CNVs), structural variants (SVs), mobile element insertions (MEIs) and short tandem repeat expansions (STRs) (Extended Data Fig. 2). Each ERN generated a list of established disease genes for their respective conditions, resulting in gene lists ranging from 230 genes for GENTURIS to 1,820 for RND (Methods and Supplementary Table 2). Systematic reanalyses resulted in 506 genetic diagnoses, by (probable) pathogenic variants that explained the phenotype, representing 8.4% of probands. The amount of time that was invested in expert reanalysis was manageable at 4.8 min per variant, or 42.8 min on average per proband.

New molecular diagnoses. SNV/InDel reanalysis revealed 461 (probable) pathogenic variants, enabling a diagnosis in 419 families. To retrieve the 461 (likely) pathogenic SNV/InDel variants from the >50,000 prioritized variants, an average of nine variants underwent molecular and clinical expert review (Supplementary Table 3).

The 461 SNV/InDel variants identified, in 419 probands, consisted of 282 heterozygous variants with dominant effect, 85 homozygous and 76 compound heterozygous variants with recessive effect and 18 hemizygous variants. Functionally, these represented 187 nonsense/frameshift variants, 249 missense variants, 11 in-frame deletions, ten splicing variants (eight intronic and two synonymous), two 5′ UTR variants, one promoter region variant and one complex InDel variant (Fig. 2 and Supplementary Table 4). Forty-one of the 461 (9.1%) variants could be confirmed as de novo mutations, due to the availability of proband–parent trios for 1,320 (22%) families, primarily from ERN ITHACA (1,081).

We evaluated why the 461 SNV/InDel variants had not been classified as disease causing in previous analyses. We found that 67 affect genes which were established as a novel disease gene following data submission to Solve-RD (that is, appeared in Online Mendelian Inheritance in Man (OMIM) after 1 January 2018; Extended Data Fig. 3 and Supplementary Table 4), while the remaining 394 were among established disease genes at the time of data submission. Of these, 117 variants have been reclassified in the interim (that is, novel or modified ClinVar[16] entry since 2018) and 70 had initially been deemed not fully explaining disease, despite the variant being classified as pathogenic in ClinVar as a result of perceived insufficient clinical concordance at the time. The remaining 207 variants were not included in ClinVar and were classified only as (probable) pathogenic by the experts involved in this project.

We applied a suite of analysis tools for calling and annotating variants. These included queries for noncanonical splice variants, mtDNA variants, CNVs, SVs, MEIs and STRs. These additional analyses yielded a diagnosis in 87 rare-disease families among a total of 88 variants, with CNVs in 44 probands (45 variants) being the most prevalent variant type (Fig. 3). This included three cases where biallelic pairings of an

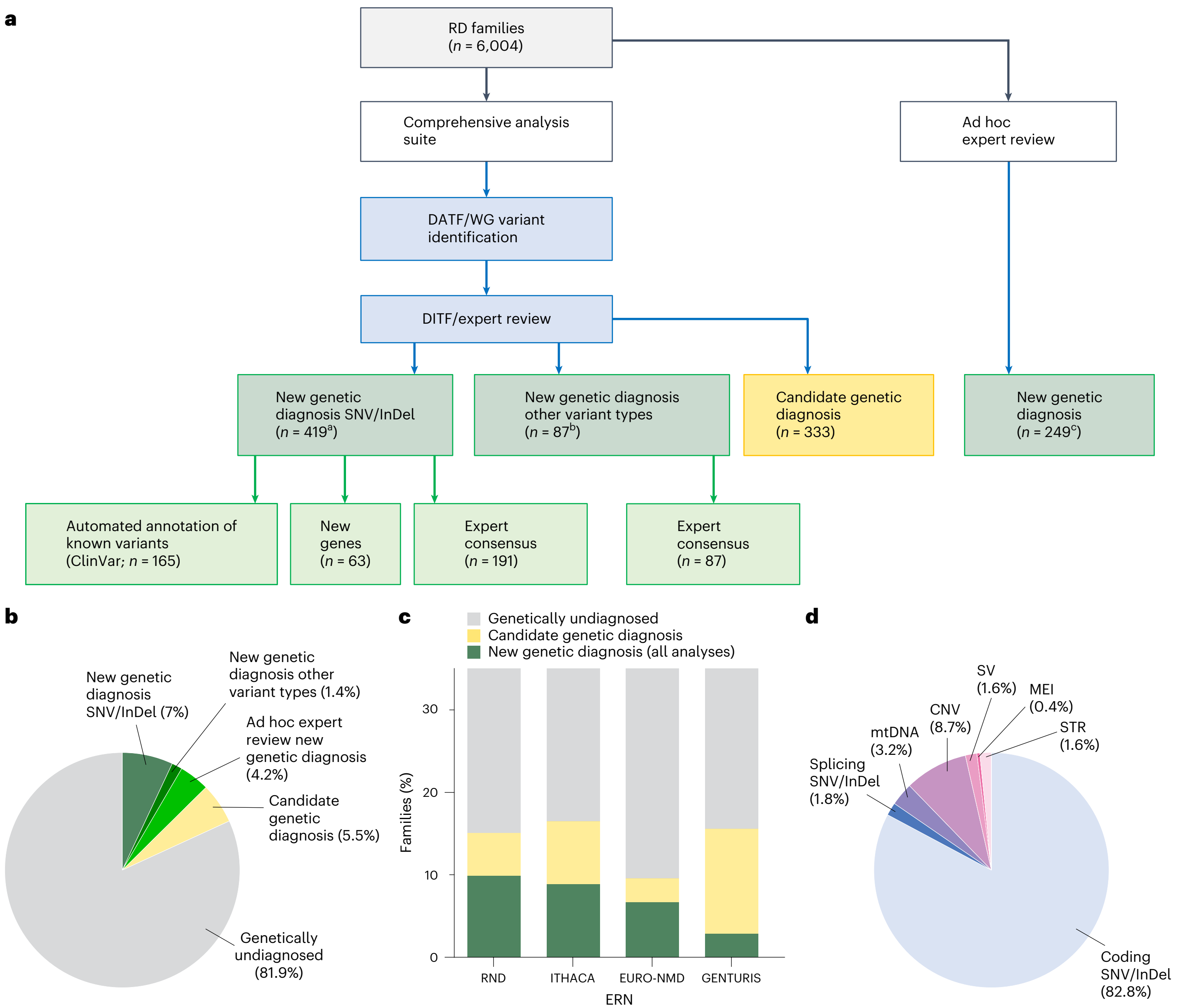
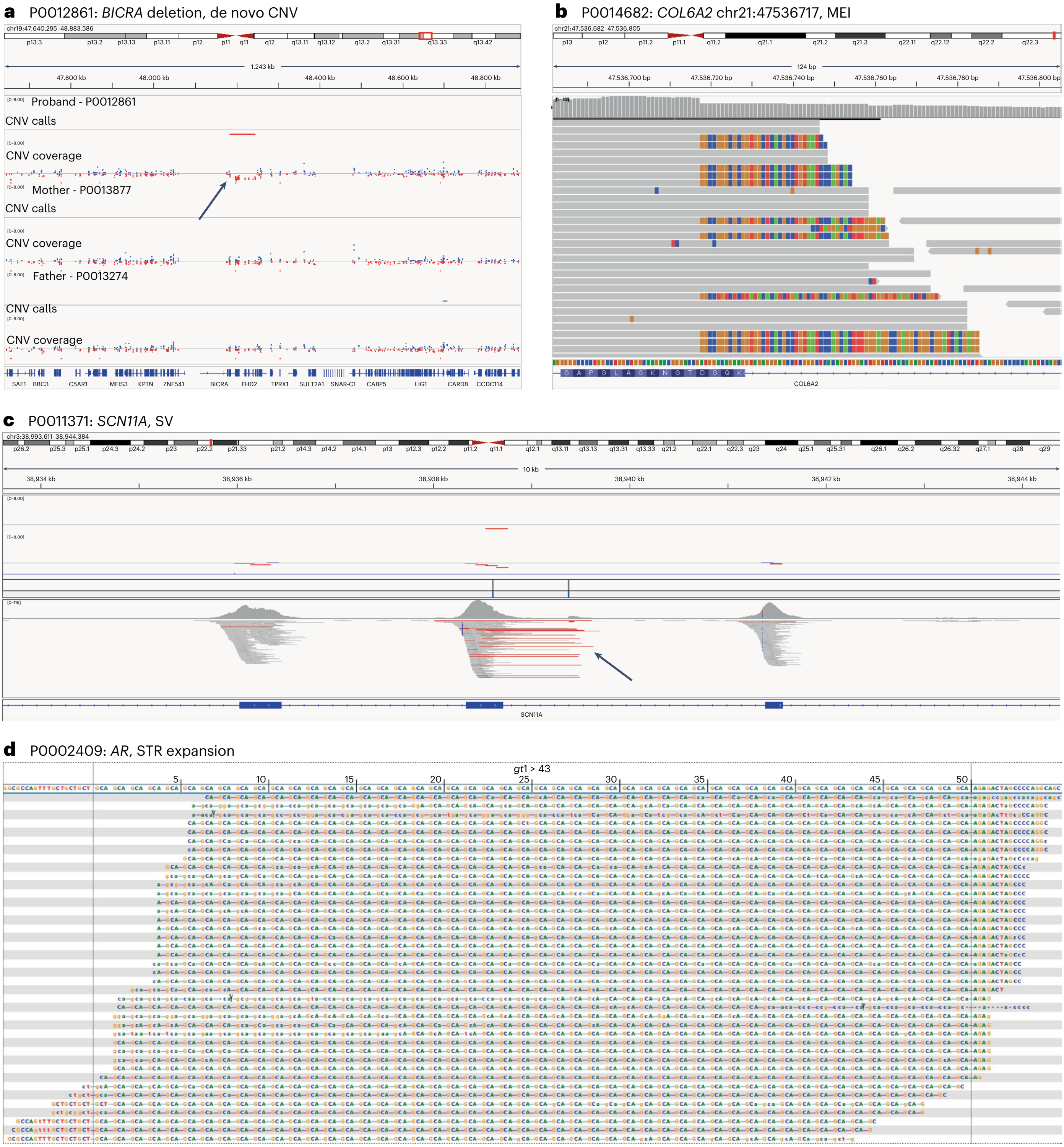


Fig. 2 | Systematic reanalysis of genomic datasets for the genetic diagnosis of rare diseases. **a**, Flowgram of systematic analysis of 6,004 families. Yield per analysis type (genetic diagnoses by SNV/InDel and other variant types; candidate genetic diagnoses and genetic diagnoses by ad hoc expert review) are shown. For SNV/InDels, we evaluated why the 464 variants previously identified in 419 families had not been classified as disease causing. **b**, Chart summarizing diagnostic yield across 6,004 families in Solve-RD. **c**, Chart summarizing yield per disease category (ERN); the denominator is 6,004 families. **d**, Chart summarizing the different variant types that led to a molecular diagnosis in 506 of 6,004 families as part of the systematic reanalysis effort of Solve-RD. [a]Disease-causing SNVs or short insertions/deletions were identified in 419 families. [b]Disease-causing non-SNV variants identified in 87 families, including three cases of compound heterozygosity involving an SNV and a CNV/SV, identified through the 'other variant type' analyses, and are counted only under 'New genetic diagnosis other variant types'. [c]In 114 of 147 cases where we could confirm the variant identified in the ad hoc analysis, we established that it would also have been found by the standard analysis. RD, rare disease; splicing SNV/InDel, noncanonical splicing sites; WG, work group.

SNV with a CNV/SV formed a compound heterozygous variant, and one case where two CNVs affecting different genes led to a digenic diagnosis (Extended Data Fig. 2 and Supplementary Table 4).

The diagnostic yield across disease groups (that is, ERNs) ranged from 2.8% (genetic tumor risk syndromes, GENTURIS) to 10.6% (rare neurological disorders, RND), correlating with the number of established disease genes provided by the ERNs (Fig. 2 and Supplementary Table 2). Overall, for the 506 newly diagnosed probands, the inheritance pattern was autosomal dominant for 306, autosomal recessive for 137, X-linked for 42, mitochondrial for 16, dual diagnoses in four individuals and digenic inheritance in one individual (Supplementary Table 4).

Next to the overall yield across the cohort, the importance of new diagnoses can be illustrated by individual rare-disease case reports, each benefitting from technical and interpretational improvements, leading to the closure of diagnostic odysseys. For example, we highlight a 58-year-old male from the RND cohort who developed a rare neurological disorder at 42 years of age, including sensory neuronopathy or sensory polyneuropathy, which was later specified as spastic ataxic gait and confirmed the presence of signs of peripheral neuropathy. Our reanalysis revealed a large intragenic deletion in combination with a missense variant in *B4GALTNT1*, which were both proven to be pathogenic (P0015028; Extended Data Fig. 4). Functional confirmation was obtained via glycomics


a P0012861: *BICRA* deletion, de novo CNV

b P0014682: *COL6A2* chr21:47536717, MEI

c P0011371: *SCN11A*, SV

d P0002409: *AR*, STR expansion

Fig. 3 | Examples of 'beyond standard' variant types by Solve-RD. a–**d**, Illustrative examples of previously unsolved rare-disease probands for which a new variant other than a coding SNV/InDel resulted in a new diagnosis. **a**, De novo CNV affecting *BICRA* (P0012861). **b**, MEI variant in *COL6A2* (P0014682). **c**, SV in *SCN11A* (P0011371). **d**, STR expansion affecting *AR* (P0002409).

analysis of plasma glycolipids, indicating reduced levels of *B4GALNT1* glycolipid products.

An example of a previously missed CNV was a small. single-exon deletion of *APC* identified in an individual (P0009136, Extended Data Fig. 5) from the GENTURIS cohort presenting with suggestive familial adenomatous polyposis. Although the clinical course, family history and haplotype analysis had already pointed to an underlying *APC* variant, the diagnostic deletion was not detected in routine diagnostics due to a lack of multiplex ligation-dependent probe amplification probes covering the specific region affected.

In the ITHACA cohort we highlight two individuals, one with a mosaic de novo mutation in *PIK3CA* (Chr3(GRCh37):g.178916876G>A;

NM_006218.4:c.263G>A; p.(Arg88Gln), present in 13% of the reads) in an individual with complex partial seizures and asymmetry of the legs and face (P0012716; Extended Data Fig. 6a). This individual had been clinically suspected of having underdevelopment of the left side of the body, rather than overgrowth of the right side of the body, which meant that an overgrowth syndrome had not previously been considered. Furthermore, probably due to mosaicism, the proband presented with a relatively mild phenotype when considering the spectrum of *PIK3CA*-related overgrowth, which made accurate clinical diagnosis challenging.

The second ITHACA example involves an individual (P0013065; Extended Data Fig. 6b) with severe developmental delay and multiple syndromic features, including delayed motor, communicative and social milestones: crawling at 15 months, walking at 30 months, first words at 7 years of age and speech characterized by severe verbal dyspraxia. Additional medical problems comprised divergent strabismus, muscle tone dysregulation with contractures and inattentive and hyperactive behavior with aggressive tantrums. Physical examination revealed a slender body and microcephaly (height 184 cm (s.d. = 0); weight 51.5 kg; body mass index 15.2; head circumference 54.5 cm, s.d. −2). He had a small, asymmetric thorax of unusual shape (the midthoracic region being broader in the frontal plane and flattened in the sagittal plane compared with the high thoracic region), high thoracic kyphosis and scapular winging. His hands and feet were slender, with long fingers and toes, camptodactyly of the 2nd, 3rd and 4th fingers of the right hand and he exhibited elbow and knee contractures. Facial dysmorphisms included a long and narrow facial shape, full eyebrows with synophrys, downslant of the palpebral fissures, prominent eyelids with ptosis, divergent strabismus, low-set ears with a square-shaped and flattened upper helix, and a short nose. Here, the identification of a de novo variant in *MN1* ended a 20-year diagnostic odyssey. The disease–gene relationship for *MN1* was established following initial routine analysis, but now finally enables the diagnosis of CEBALID syndrome.

In the NMD cohort, we highlight a 14-year-old boy with an initial diagnosis of congenital myasthenic syndrome (CMS) and his mildly affected mother. Systematic reanalysis led to the identification of a mitochondrial variant, m.3243A>G in *MT-TL1*, with an observed heteroplasmy of 0.27 in the proband and 0.14 in his mother (Extended Data Fig. 7). The difference in heteroplasmy probably correlates with the mild phenotype observed in the proband, and with the absence of mitochondrial myopathy features in his mother. While the initial clinical suspicion in the proband was CMS due to the notable fatigability, the fact that mitochondrial disease can be highly variable in presentation means that mild forms of mitochondrial myopathy can be difficult to diagnose clinically.

An example on how variant annotation pipelines can aid in variant interpretation is provided through the diagnostic path of a girl (P0012491) who was clinically suspected to have Rett syndrome (MIM#312750). Exome sequencing performed in 2014 did not yield a diagnosis, despite specific attention being applied to variants affecting *MECP2*, the gene associated with Rett syndrome. Almost 8 years later, the reanalysis presented here uncovered a pathogenic de novo MECP2 variant from the same data. Retrospective analysis of previous interpretation steps revealed that the variant was initially annotated to a less relevant isoform of *MECP2* (MECP2-e2; ENST00000303391.11), in which the variant located to an intron. However, reannotation here revealed that the variant truncates the brain-specific isoform of *MECP2* (MECP2-e1; ENST00000453960.7), and hence is indeed explanatory for the Rett syndrome in this girl.

Cases diagnosed by ad hoc expert review. During the course of Solve-RD, many contributing partners continued to perform analysis on specific families of interest, both locally and using RD-Connect GPAP. This ad hoc expert review provided 249 additional diagnoses (4.1%), some of which have been included in individual reports[13,17–22], and novel disease gene discovery efforts[23,24] published previously (Supplementary Table 5). Cases solved through ad hoc expert review were reported to Solve-RD and not interpreted further as part of the systematic reanalysis. For 197 (79%) of these ad hoc diagnoses, the causative variants were SNVs. For 147 (75%) of these SNVs we could assess post hoc whether the variants would also have been identified by the systematic reanalyses performed. We found that in 114 of 147 (78%) cases the SNVs would have been identified, while the remaining cases were diagnosed due to the discovery of variants located in novel disease genes not included in ERN gene lists, or initially discounted for technical reasons (for example, having insufficient coverage (fewer than ten reads) or being deep intronic variants).

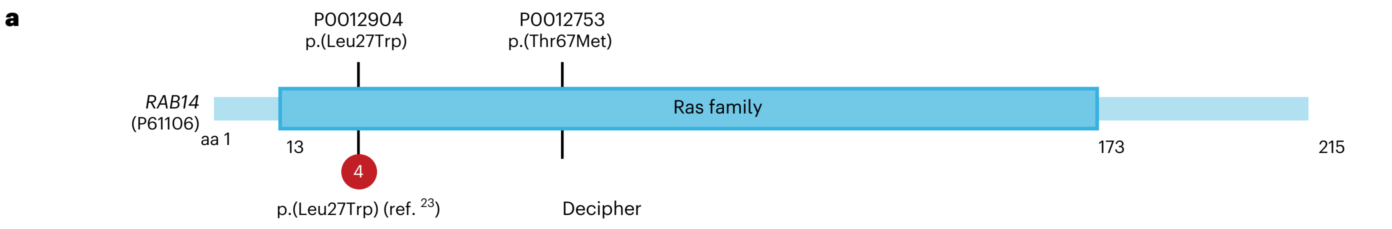
Candidate disease-causing variants. In addition to variants that were deemed causative for disease, we identified a further 378 variants (in 333 affected individuals) in established disease genes that have not yet been confirmed as causative, either because the variant does not fully explain the individual's phenotype or because the variant's pathogenicity cannot yet be conclusively determined (Fig. 2 and Supplementary Table 4).

Cross-ERN analysis, recurrences and clinical actionability

Cross-ERN de novo mutation analysis. Systematic reanalyses were performed by each of the four ERNs, thus maximizing disease-specific expertise. Because the clinical spectrum may occasionally cross ERN boundaries, we assessed all de novo mutations across all genes included in any of the ERN gene lists (2,512 unique genes), irrespective of which ERN originally submitted the case. This led to a molecular diagnosis in an additional three probands through the identification of (probable) pathogenic de novo variants in *CSDE1* (ref. 25), *EP300* and *SYT1* in individuals P0012248, P0014714 and P0018474, respectively (Supplementary Table 6), which would have been missed without this cross-ERN analysis. This included a young girl (P0014714) presenting with microcephaly, face abnormality, muscle hypotonia and neurodevelopmental delay, leading to a clinical suspicion of Cornelia de Lange syndrome (MIM#122470; https://www.omim.org/entry/122470). Solve-RD's efforts led to the identification of a de novo frameshift variant in the histone acetyltransferase p300 gene: *EP300*(NM_001429.4):c.1152_1153del; p.(Gly385GlnfsTer25), suggesting a clinical diagnosis of Rubinstein–Taybi syndrome (MIM#180849). This prompted clinical re-evaluation of the proband's phenotype, at which point the clinical diagnosis was confirmed. Another example (P0012248) concerned a young male with severe neurodevelopmental delay, microcephaly, absent speech, generalized hypotonia, nystagmus and inability to walk. Here, the systematic reanalysis of the proband's ES data within Solve-RD led to the identification of a de novo missense variant in synaptotagmin 1, *SYT1* (NM_001135806.2):c.1103T>C; p.(Ile368Thr), leading to a molecular diagnosis of Baker–Gordon syndrome (MIM#618218). Retrospective analysis of the original ES data of both cases revealed that the variants had not been identified by the corresponding in-house pipeline.

Recurrent variants. We observed recurrence for 21 (probable) pathogenic variants, together accounting for 41 diagnoses (Supplementary Table 7). These 21 variants occurred in 18 genes, with three genes (*SPG7*, *KCNA2* and *SPAST*) harboring two different recurring variants.

One of the recurring variants was identified across three ERNs: an identical *MT-ATP6* missense variant (chrM:9185T>C (ENST00000361899:c.659T>C (p.(Leu220Pro))) was observed in five affected individuals (P0010243, P0009606, P0009608, P0004265 and P0004266) from three unrelated families submitted by ERNs EURO-NMD, RND and ITHACA. The variant was observed with a heteroplasmy of 77 and 90% in the EURO-NMD and RND probands, respectively, while it was homoplasmic in the



b *RAB14* phenotypes

	P0012753 (male) p.(Thr67Met)	P0012904 (male) p.(Leu27Trp)	rumc_580 (female) p.(Leu27Trp) (ref. 23)	Decipher 305550 (female) p.(Thr67Met)
Abnormality of the nervous system	Global developmental delay (HP:0001263); delayed speech and language development (HP:0000750); motor delay (HP:0001270)	Mild intellectual disability (HP:0001256); delayed speech and language development (HP:0000750)	Moderate intellectual disability (HP:0002342)	Moderate intellectual disability (HP:0002342)
Head circumference	–	–	Macrocephaly (HP:0000256)	–
Facial dysmorphisms	–	High, square-shaped forehead (HP:0000244); downslanted palpebral fissures (HP:0000494); low-hanging columella (HP:0009765)	–	Protruding tongue (HP:0010808); smooth philtrum (HP:0000319); wide mouth (HP:0000154); large earlobe (HP:0009748)
Behavioural problems	Sleep disturbance (HP:0002360)		Autism spectrum disorder (HP:0000729)	–
Abnormality of the limbs	–	–	–	Congenital talipes calcaneovalgus (HP:0005850); short foot (HP:0001773)
Abnormality of the musculoskeletal system	–	Muscular hypotonia (HP:0001252);	–	–
Abnormality of the abdominal wall	–	Inguinal hernia (HP:0000023)	–	Umbilical hernia (HP:0001537)
Other	–	Recurrent upper respiratory tract infections (HP:0002788)	–	Thick hair (HP:0100874)

Fig. 4 | Example of a new discovery by Solve-RD. a,**b**, An example of discoveries enabled by the Solve-RD resource. **a**, *RAB14* de novo variants in two cases from this project contribute to the establishment of a new genotype–phenotype relationship. The first individual (P0012753) presents with mild global developmental delay in the absence of any facial dysmorphism or congenital anomalies, and carries a de novo variant in *RAB14* (chr9:123952916G>A; NM_016322.3:c.200C>T; p.(Thr67Met)), which is rare (not observed in gnomAD v.2.1.1), likely to be deleterious (CADD score of 29) and has been observed de novo in at least four additional individuals with developmental disorders in the literature[23]. The second individual (P0012904) presents with mild ID, subtle facial dysmorphisms comprising a high, square-shaped forehead, downslant of palpebral fissures and a low-hanging columella, in the absence of congenital anomalies. The de novo variant found in this individual (chr9:123954475A>C; NM_016322.3:c.80T>G; (p.(Leu27Trp)) is also absent from gnomAD, predicted to be deleterious (CADD score of 28) and has been observed de novo in at least one additional individual with a neurodevelopmental disorder in DECIPHER (https://www.deciphergenomics.org/patient/305550/phenotypes/person/62257). The female individual reported in Decipher presents with moderate ID, facial dysmorphism consisting of large earlobes, smooth philtrum, a wide mouth and protruding tongue, short feet with congenital talipes calcaneovalgus, thick hair and an umbilical hernia. **b**, Salent features of the two cases in **a**. aa, Amino acid.

ITHACA proband, in line with the variable phenotypic presentation (Supplementary Table 8).

Beyond diagnosis to clinical actionability. We investigated the number of diagnosed individuals that would potentially benefit from therapy or other actionability, by considering medications or interventions included in three databases: IEMbase[26], Treatabolome[27] and ClinGen[28], and in international cancer guidelines.

We identified 73 affected individuals (14.4% of diagnosed individuals) that harbored variants in a potentially actionable gene (Extended Data Fig. 8).

Implementation, and feedback to referring clinicians and eventually to families and patients, is following local guidelines that differ between centers. Actual actionability has already happened and is continuously ongoing. To date we have received feedback for a subset of the aforementioned cases, with details of 16 examples summarized in Supplementary Table 9.

An example from ERN EURO-NMD is provided by the case of two young-adult patients from different families who had presented with limb-girdle muscle weakness and fatigability from 2 years of age, and subsequently developed ptosis and difficulty in swallowing, leading to a suspected diagnosis of limb-girdle myasthenic syndrome (P0020778). While previous ES analyses were negative, reanalysis within Solve-RD using SpliceAI[29] led to the identification of a homozygous intronic variant with a potential splice donor effect, c.1023+5G>A proximal to the exon 5–intron 5 junction of *DES* in both patients. In parallel, but outwith Solve-RD, a female with a similar phenotype, among a cohort of patients suspected of having CMS being treated in the same hospital, was also found to be homozygous for this mutation. Subsequent laboratory analyses indicated reduced production of normal desmin transcript and protein. Administration of the standard CMS treatment of pyridostigmine and salbutamol was initiated and, while one of the two patients showed no improvement after 3 months, the other exhibited 50% improvement in measures of fatigable weakness.

Discussion

Genomic data from rare-disease cases that have been extensively analyzed by experts in the past can still yield a large number of new diagnoses, with previous studies reporting success rates commonly in the range of 6–13% (ref. 5). We previously reported on preliminary ClinVar-focussed reanalyses undertaken within Solve-RD, which resulted in molecular diagnoses being provided for 111 families[12,13]. The value of an in-depth systematic reanalysis is supported by our success in diagnosing 8.4% of affected individuals through our systematic reanalysis, and the further 4.1% diagnosed in parallel by local reanalysis in individual centers through ad hoc expert review. In total, we have successfully diagnosed 12.6% of families to date. While a few recent studies have reported higher diagnostic rates following reanalysis, ranging from 15–21% (refs. 30–33), it should be noted that those datasets were more homogeneous in nature, usually originated from a single country and were of substantially smaller scale and breadth. Nevertheless, our diagnostic yield is at the top end of the typical range[5].

The proposed framework, rare disease–reanalysis logistics (RD-REAL), with its two-level expert review (Methods), represents a practical blueprint for reanalysis efforts. Here we limited our analysis to four of 24 ERN rare-disease domains and, although it remains to be established whether similar results can be obtained in the other domains, the approach applied in Solve-RD is generic and can easily be implemented across the full gamut of rare diseases and at global scale.

Such collaborative reanalysis efforts can, for the present, exist in parallel with local or national reanalysis efforts, ideally embedded within the healthcare system and allowing for prompt return of results with immediate actionability in some individual cases. Ultimately, reanalysis efforts should be automated.

Further, the previously generated exome and genome sequencing data were highly heterogeneous because this is a pan-European project aiming to provide diagnoses for individuals across Europe. This heterogeneity, both in terms of the quality of the historic ES data and the breadth of phenotypic descriptions, impacted upon our ability to confidently identify potentially pathogenic variants. The limited number of genomes, and the focus on well-established disease genes used in this study, were not sufficient to support a systematic advantage of genome over exome sequencing in rare-disease studies (Supplementary Table 10). Another limitation was that, for two-thirds of the families analyzed (4,103 of 6,004), we had sequencing data only from the affected proband, thus limiting supporting segregation information during downstream variant interpretation, especially with respect to the identification of pathogenic de novo variants.

This study provides several key insights. After more than a decade of diagnostic exome sequencing[34,35], our knowledge of the spectrum of genes and variants causing monogenic rare disease, and of the bioinformatic pipelines used to detect them, is still increasing. This is exemplified not only by the large number of SNV/InDel variants that can now be correctly interpreted, leading to 84.1% of all novel diagnoses (n = 419), based on the availability of new gene- or variant-level information, but also by the substantial proportion (15.9%, n = 87) of novel diagnoses that were a result of individually rare variant types not previously detectable by standard diagnostic bioinformatics pipelines.

With the growing size of rare-disease datasets, we shall identify an increasing number of identical variants in multiple individuals, improving the odds of arriving at the correct variant interpretation for multiple cases. This is evident here, because we identified 21 (probable) pathogenic variants that occurred two or three times across a total of 41 unrelated probands from the 6,004 families analyzed, sometimes straddling different clinical disease categories.

We examined clinical actionability for the diagnoses in the series, using a definition that considered only approved medication or (preventive) interventions. This is a more restrictive definition than that applied in a previous study[3]. Even without considering reproductive choice and surveillance of family members, there was potential for medical actionability in 14.4% of those receiving a diagnosis in our series, with ongoing implementation and the first concrete examples shown in Supplementary Table 9.

In Solve-RD, we developed several practical recommendations for large-scale distributed genomic reanalysis initiatives.

Because data submitted are likely to be heterogeneous, it is essential to standardize phenoclinical data and metadata, and to start genomic reanalysis using raw sequencing reads: define strict inclusion criteria, including checking and verifying biological relationships; and define a minimum on-target coverage of 80-fold for exome sequencing and 30-fold for genome sequencing. Multiple variant-calling pipelines should be used for each variant type, as highlighted by the results of our CNV analysis. Regular updates to bioinformatic workflows are essential for integration of new tools and the latest versions of databases such as gnomAD and ClinVar. When variants are found in genes linked to the individual's phenotype, consider reducing stringency in alternative allele frequency and/or read-depth to detect mosaicism or true heterozygotes with poor allele balance.

When prioritizing cases for reanalysis, focus on those analyzed further in the past, and prioritize variants based on their presence in clinical interpretation databases such as ClinVar, HGMD and similar resources. Favor specificity over sensitivity when sharing short lists of variants, and ensure they are shared only once per individual. Record feedback from variant interpretation—whether confirming disease-causing variants, identifying potential candidates or discarding them—in an accessible database to prevent duplicated efforts. Finally, reverse phenotyping is crucial for re-evaluation of clinical diagnoses, particularly in syndromic cases.

We already have the first insights into the future value of the Solve-RD resource and infrastructure. Our current effort focussed on diagnoses in established rare-disease genes. However, this resource and the datasets in Solve-RD should be well suited for the generation of continued insights. Since the systematic analysis presented here was completed, we have already promoted two SVs and seven CNVs from candidate to disease causing[36,37], and likewise for an additional ten SNV/InDel variants (Supplementary Table 11). This resource shall also allow the discovery of novel disease genes or loci, and the discovery of new disease mechanisms and causes is an ongoing part of Solve-RD[7,11]. The recent association of the noncoding RNA gene *RNU4-2* with a complex NDD phenotype[38,39] led to one further solved case in Solve-RD (P001996), in addition to the Solve-RD case (P0007197) that contributed to the original discovery (Extended Data Fig. 9c). As a further example we highlight *RAB14*, which had been suggested to play a role in neurodevelopmental disorders by a statistically significant enrichment of de novo variants in a developmental disorder cohort in 2020 (ref. 23). The Solve-RD dataset includes data from two male individuals with neurodevelopmental phenotypes harboring de novo variants in *RAB14*, now enabling genotype–phenotype characterization as a result of the comprehensive HPO description collected here (Fig. 4a,b). Similarly, many additional genotype–phenotype and/or mechanistic studies have been initiated from the Solve-RD datasets and are currently followed up within the Solve-RD RDMM-Europe initiative[40].

Global data sharing is essential for discoveries in rare-disease diagnostic research[41], and has been enabled here. Authorized users can use either RD-Connect GPAP to search and analyze integrated phenotype (HPO and ORPHA codes) and gene- and variant-level data, or EGA to download all data. The worldwide detection of gene-level recurrence in other individuals affected by a rare condition is further facilitated through connection to the MatchMaker Exchange network[42]. To benefit the rare-disease community, our framework will involve expansion to other types of rare diseases through their respective ERNs, the incorporation of novel omics datasets[43–45]—including those obtained from long-read technologies[46–51]—and the inclusion of artificial intelligence-based methodology[52]. The tools and infrastructure developed within Solve-RD have been adopted as the core framework for undiagnosed rare-disease case reanalysis within the ERDERA project, which aims to extend out to all 24 ERNs and reanalyze >100,000 datasets from rare-disease families across all disease types (https://erdera.org/).

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41591-024-03420-w.

References

1. Nguengang Wakap, S. et al. Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. *Eur. J. Hum. Genet.* **28**, 165–173 (2020).
2. Turro, E. et al. Whole-genome sequencing of patients with rare diseases in a national health system. *Nature* **583**, 96–102 (2020).
3. Smedley, D. et al. 100,000 Genomes Pilot on Rare-Disease Diagnosis in Health Care – Preliminary Report. *N. Engl. J. Med.* **385**, 1868–1880 (2021).
4. Wright, C. F. et al. Genomic diagnosis of rare pediatric disease in the United Kingdom and Ireland. *N. Engl. J. Med.* **388**, 1559–1571 (2023).
5. Dai, P. et al. Recommendations for next generation sequencing data reanalysis of unsolved cases with suspected Mendelian disorders: a systematic review and meta-analysis. *Genet. Med.* **24**, 1618–1629 (2022).
6. López-Martín, E., Martínez-Delgado, B., Bermejo-Sánchez, E. & Alonso, J. SpainUDP: the Spanish Undiagnosed Rare Diseases Program. *Int. J. Environ. Res. Public Health* **15**, 1746 (2018).
7. Zurek, B. et al. Solve-RD: systematic pan-European data sharing and collaborative analysis to solve rare diseases. *Eur. J. Hum. Genet.* **29**, 1325–1331 (2021).
8. Laurie, S. et al. The RD-Connect Genome-Phenome Analysis Platform: accelerating diagnosis, research, and gene discovery for rare diseases. *Hum. Mutat.* **43**, 717–733 (2022).
9. Freeberg, M. A. et al. The European Genome-phenome Archive in 2021. *Nucleic Acids Res.* **50**, D980–D987 (2022).
10. Köhler, S. et al. The human phenotype ontology in 2021. *Nucleic Acids Res.* **49**, D1207–D1217 (2021).
11. Graessner, H., Zurek, B., Hoischen, A. & Beltran, S. Solving the unsolved rare diseases in Europe. *Eur. J. Hum. Genet* **29**, 1319–1320 (2021).
12. Denommé-Pichon, A. S. et al. A Solve-RD ClinVar-based reanalysis of 1522 index cases from ERN-ITHACA reveals common pitfalls and misinterpretations in exome sequencing. *Genet. Med.* **25**, 100018 (2023).
13. Matalonga, L. et al. Solving patients with rare diseases through programmatic reanalysis of genome-phenome data. *Eur. J. Hum. Genet.* **29**, 1337–1347 (2021).
14. Jacobsen, J. O. B. et al. The GA4GH Phenopacket schema defines a computable representation of clinical data. *Nat. Biotechnol.* **40**, 817–820 (2022).
15. Lagorce, D. et al. Phenotypic similarity-based approach for variant prioritization for unsolved rare disease: a preliminary methodological report. *Eur. J. Hum. Genet.* **32**, 182–189 (2024).
16. Landrum, M. J. et al. ClinVar: Improvements to accessing data. *Nucleic Acids Res.* **48**, D835–D844 (2020).
17. Schüle, R. et al. Solving unsolved rare neurological diseases—a Solve-RD viewpoint. *Eur. J. Hum. Genet.* **29**, 1332–1336 (2021).
18. de Boer, E. et al. A MT-TL1 variant identified by whole exome sequencing in an individual with intellectual disability, epilepsy, and spastic tetraparesis. *Eur. J. Hum. Genet.* **29**, 1359–1368 (2021).
19. Töpf, A. et al. Exome reanalysis and proteomic profiling identified TRIP4 as a novel cause of cerebellar hypoplasia and spinal muscular atrophy (PCH1). *Eur. J. Hum. Genet.* **29**, 1348–1353 (2021).
20. te Paske, I. B. A. W. et al. A mosaic PIK3CA variant in a young adult with diffuse gastric cancer: case report. *Eur. J. Hum. Genet.* **29**, 1354–1358 (2021).
21. Pauly, M. G. et al. Not to miss: intronic variants, treatment, and review of the phenotypic spectrum in VPS13D-related disorder. *Int. J. Mol. Sci.* **24**, 1874 (2023).
22. Pauly, M. G. et al. The expanding phenotypical spectrum of WARS2-related disorder: four novel cases with a common recurrent variant. *Genes (Basel)* **14**, 822 (2023).
23. Kaplanis, J. et al. Evidence for 28 genetic disorders discovered by combining healthcare and research data. *Nature* **586**, 757–762 (2020).
24. Weihl, C. C. et al. Loss of function variants in DNAJB4 cause a myopathy with early respiratory failure. *Acta Neuropathol.* **145**, 127–143 (2023).
25. Gangfuß, A. et al. A de novo CSDE1 variant causing neurodevelopmental delay, intellectual disability, neurologic and psychiatric symptoms in a child of consanguineous parents. *Am. J. Med. Genet. A* **188**, 283–291 (2022).
26. Ferreira, C. R., van Karnebeek, C. D. M., Vockley, J. & Blau, N. A proposed nosology of inborn errors of metabolism. *Genet. Med.* **21**, 102–106 (2019).
27. Atalaia, A. et al. A guide to writing systematic reviews of rare disease treatments to generate FAIR-compliant datasets: building a Treatabolome. *Orphanet J. Rare Dis.* **15**, 206 (2020).
28. Rehm, H. L. et al. ClinGen–the Clinical Genome Resource. *N. Engl. J. Med.* **372**, 2235–2242 (2015).
29. Jaganathan, K. et al. Predicting splicing from primary sequence with deep learning. *Cell* **176**, 535–548 (2019).
30. Wright, C. F. et al. Making new genetic diagnoses with old data: iterative reanalysis and reporting from genome-wide data in 1,133 families with developmental disorders. *Genet. Med.* **20**, 1216–1223 (2018).
31. Baker, S. W. et al. Automated clinical exome reanalysis reveals novel diagnoses. *J. Mol. Diagn.* **21**, 38–48 (2019).
32. Liu, P. et al. Reanalysis of clinical exome sequencing data. *N. Engl. J. Med.* **380**, 2478–2480 (2019).
33. Bullich, G. et al. Systematic collaborative reanalysis of genomic data improves diagnostic yield in neurologic rare diseases. *J. Mol. Diagn.* **24**, 529–542 (2022).
34. de Ligt, J. et al. Diagnostic exome sequencing in persons with severe intellectual disability. *N. Engl. J. Med.* **367**, 1921–1929 (2012).
35. Rauch, A. et al. Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. *Lancet* **380**, 1674–1682 (2012).
36. Demidov, G. et al. Comprehensive reanalysis for CNVs in ES data from unsolved rare disease cases results in new diagnoses. *NPJ Genom. Med.* **9**, 49 (2024).
37. Demidov, G. et al. Structural variant calling and clinical interpretation in 6224 unsolved rare disease exomes. *Eur. J. Hum. Genet.* **32**, 998–1004 (2024).
38. Greene, D. et al. Mutations in the U4 snRNA gene RNU4-2 cause one of the most prevalent monogenic neurodevelopmental disorders. *Nat. Med.* **30**, 2165–2169 (2024).
39. Chen, Y. et al. De novo variants in the RNU4-2 snRNA cause a frequent neurodevelopmental syndrome. *Nature* **632**, 832–840 (2024).
40. Ellwanger, K. et al. Model matchmaking via the Solve-RD Rare Disease Models & Mechanisms Network (RDMM-Europe). *Lab. Anim.* **53**, 161–165 (2024).
41. Rehm, H. L. Time to make rare disease diagnosis accessible to all. *Nat. Med.* **28**, 241–242 (2022).
42. Boycott, K. M., Azzariti, D. R., Hamosh, A. & Rehm, H. L. Seven years since the launch of the Matchmaker Exchange: the evolution of genomic matchmaking. *Hum. Mutat.* **43**, 659–667 (2022).
43. Cummings, B. B. et al. Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. *Sci. Transl. Med.* **9**, eaal5209 (2017).

44. Wortmann, S. B. et al. How to proceed after 'negative' exome: a review on genetic diagnostics, limitations, challenges, and emerging new multiomics techniques. *J. Inherit. Metab. Dis.* **45**, 663–681 (2022).
45. Yépez, V. A. et al. Clinical implementation of RNA sequencing for Mendelian disease diagnostics. *Genome Med.* **14**, 38 (2022).
46. Mantere, T., Kersten, S. & Hoischen, A. Long-read sequencing emerging in medical genetics. *Front. Genet.* **10**, 426 (2019).
47. Beyter, D. et al. 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 (2021).
48. Merker, J. D. et al. Long-read genome sequencing identifies causal structural variation in a Mendelian disease. *Genet. Med.* **20**, 159–163 (2018).
49. Sabatella, M. et al. Optical genome mapping identifies a germline retrotransposon insertion in SMARCB1 in two siblings with atypical teratoid rhabdoid tumors. *J. Pathol.* **255**, 202–211 (2021).
50. Cohen, A. S. A. et al. Genomic answers for children: dynamic analyses of >1000 pediatric rare disease genomes. *Genet. Med.* **24**, 1336–1348 (2022).
51. Te Paske, I. B. A. W. et al. Noncoding aberrations in mismatch repair genes underlie a substantial part of the missing heritability in Lynch syndrome. *Gastroenterology* **163**, 1691–1694 (2022).
52. Cheng, J. et al. Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* **381**, eadg7492 (2023).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Steven Laurie[1,2,83], Wouter Steyaert[3,4,83], Elke de Boer[3,5,83], Kiran Polavarapu[6,83], Nika Schuermans[7,8,9,83], Anna K. Sommer[10,83], German Demidov[11], Kornelia Ellwanger[11], Ida Paramonov[1,2], Coline Thomas[12], Stefan Aretz[10,13], Jonathan Baets[14,15,16], Elisa Benetti[17,18], Gemma Bullich[1,2], Patrick F. Chinnery[19,20], Jill Clayton-Smith[21,22], Enzo Cohen[23], Daniel Danis[24], Jean-Madeleine de Sainte Agathe[25], Anne-Sophie Denommé-Pichon[26,27], Jordi Diaz-Manera[28], Stephanie Efthymiou[29], Laurence Faivre[26,30,31,32,33], Marcos Fernandez-Callejo[1,2], Mallory Freeberg[12], José Garcia-Pelaez[34,35,36], Lena Guillot-Noel[37], Tobias B. Haack[11], Mike Hanna[38], Holger Hengel[39,40], Rita Horvath[19], Henry Houlden[29], Adam Jackson[21,22], Lennart Johansson[41], Mridul Johari[42], Erik-Jan Kamsteeg[3], Melanie Kellner[39,40], Tjitske Kleefstra[3,5,43,44], Didier Lacombe[45,46], Hanns Lochmüller[1,6,47,48,49], Estrella López-Martín[50], Alfons Macaya[51], Anna Marcé-Grau[51], Aleš Maver[52], Heba Morsy[29,53], Francesco Muntoni[54,55], Francesco Musacchia[56,57], Isabelle Nelson[23], Vincenzo Nigro[56,57], Catarina Olimpio[19,58], Carla Oliveira[35,36,37], Jaroslava Paulasová Schwabová[59], Martje G. Pauly[60,61,62], Borut Peterlin[52], Sophia Peters[10], Rolph Pfundt[3,5], Giulio Piluso[56], Davide Piscia[1,2], Manuel Posada[50], Selina Reich[39,40], Alessandra Renieri[17,18,63], Lukas Ryba[64], Karolis Šablauskas[3,65], Marco Savarese[42], Ludger Schöls[39,40], Leon Schütz[11], Verena Steinke-Lange[66,67], Giovanni Stevanin[37], Volker Straub[28], Marc Sturm[11], Morris A. Swertz[41], Marco Tartaglia[68], Iris B. A. W. te Paske[3,4], Rachel Thompson[6], Annalaura Torella[56,57], Christina Trainor[28], Bjarne Udd[42,69,70], Liedewei Van de Vondel[14,15,71], Bart van de Warrenburg[5,72], Jeroen van Reeuwijk[3,5], Jana Vandrovcova[29], Antonio Vitobello[26,27], Janet Vos[3,4], Emílie Vyhnálková[64], Robin Wijngaard[3,4], Carlo Wilke[39,40], Doreen William[73,74], Jishu Xu[11,39,40], Burcu Yaldiz[3], Luca Zalatnai[1,2], Birte Zurek[11], Solve-RD DITF-GENTURIS*, Solve-RD DITF-ITHACA*, Solve-RD DITF-EURO-NMD*, Solve-RD DITF-RND*, Solve-RD consortium*, Anthony J. Brookes[75], Teresinha Evangelista[23], Christian Gilissen[3,4], Holm Graessner[11,76], Nicoline Hoogerbrugge[3,4], Stephan Ossowski[11,77], Olaf Riess[11,76], Rebecca Schüle[39,40], Matthis Synofzik[39,40], Alain Verloes[78,79], Leslie Matalonga[1,2], Han G. Brunner[3,5,80], Katja Lohmann[61,84], Richarda M. de Voer[3,4,84], Ana Töpf[28,84], Lisenka E.L.M. Vissers[3,5,84], Sergi Beltran[1,81,84] & Alexander Hoischen[3,4,82,84]

[1]Centro Nacional de Análisis Genómico (CNAG), Barcelona, Spain. [2]Universitat de Barcelona (UB), Barcelona, Spain. [3]Department of Human Genetics, Radboud University Medical Center, Nijmegen, the Netherlands. [4]Radboud Institute for Medical Innovation, Nijmegen, the Netherlands. [5]Donders Institute for Brain, Cognition and Behaviour, Radboud University Medical Center, Nijmegen, the Netherlands. [6]Children's Hospital of Eastern Ontario Research Institute, University of Ottawa, Ottawa, Ontario, Canada. [7]Program for Undiagnosed Rare Diseases (UD-PrOZA), Ghent University Hospital, Ghent, Belgium. [8]Department of Biomolecular Medicine, Faculty of Medicine and Health Sciences, Ghent University, Ghent, Belgium. [9]Center for Medical Genetics, Ghent University Hospital, Ghent, Belgium. [10]Institute of Human Genetics, Medical Faculty, University of Bonn, Bonn, Germany. [11]Institute of Medical Genetics and Applied Genomics, University of Tübingen, Tübingen, Germany. [12]European Bioinformatics Institute, European Molecular Biology Laboratory, Cambridge, UK. [13]Center for Hereditary Tumor Syndromes, University Hospital Bonn, Bonn, Germany. [14]Translational Neurosciences, Faculty of Medicine and Health Sciences, University of Antwerp, Antwerp, Belgium. [15]Laboratory of Neuromuscular Pathology, Institute Born-Bunge, University of Antwerp, Antwerp, Belgium. [16]Neuromuscular Reference Centre, Department of Neurology, Antwerp University Hospital, Antwerp, Belgium. [17]Med Biotech Hub and Competence Center, Department of Medical Biotechnologies, University of Siena, Siena, Italy. [18]Medical Genetics, University of Siena,

Siena, Italy. [19]Department of Clinical Neurosciences, University of Cambridge, Cambridge, UK. [20]Medical Research Council Mitochondrial Biology Unit, University of Cambridge, Cambridge, UK. [21]Division of Evolution, Infection and Genomics, School of Biological Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Manchester, UK. [22]Manchester Centre for Genomic Medicine, St Mary's Hospital, Manchester University Hospitals NHS Foundation Trust, Health Innovation Manchester, Manchester, UK. [23]Centre de Recherche en Myologie, Sorbonne Université, Inserm, Institut de Myologie, Paris, France. [24]Jackson Laboratory for Genomic Medicine, Farmington, CT, USA. [25]Department of Genetics, Assistance Publique-Hôpitaux de Paris, Sorbonne Université, Pitié-Salpêtrière University Hospital, Paris, France. [26]University of Burgundy, Dijon, France. [27]Functional Unit for Diagnostic Innovation in Rare Diseases, Dijon Bourgogne University Hospital, Dijon, France. [28]John Walton Muscular Dystrophy Research Centre, Translational and Clinical Research Institute, Newcastle University and Newcastle Hospitals NHS Foundation Trust, Newcastle upon Tyne, UK. [29]Department of Neuromuscular Diseases, UCL Queen Square Institute of Neurology, London, UK. [30]Genetics Department, Dijon University Hospital, Dijon, France. [31]Centre of Reference for Rare Diseases: Development Disorders and Malformation Syndromes, Dijon University Hospital, Dijon, France. [32]University of Burgundy-Franche Comté, Dijon, France. [33]GIMI institute, Dijon University Hospital, Dijon, France. [34]Instituto de Investigação e Inovação em Saúde, Universidade do Porto, Porto, Portugal. [35]IPATIMUP – Institute of Molecular Pathology and Immunology of the University of Porto, Porto, Portugal. [36]Faculty of Medicine, University of Porto, Porto, Portugal. [37]Institut du Cerveau, Sorbonne University, Paris, France. [38]MRC Centre for Neuromuscular Diseases and National Hospital for Neurology and Neurosurgery, UCL Queen Square Institute of Neurology, London, UK. [39]Department of Neurodegeneration, Hertie Institute for Clinical Brain Research (HIH), University of Tübingen, Tübingen, Germany. [40]German Center for Neurodegenerative Diseases (DZNE), Tübingen, Germany. [41]Department of Genetics, Genomics Coordination Center, University Medical Center Groningen, University of Groningen, Groningen, the Netherlands. [42]Folkhälsan Research Centre and Medicum, University of Helsinki, Helsinki, Finland. [43]Department of Clinical Genetics, Erasmus MC, Rotterdam, the Netherlands. [44]Center of Excellence for Neuropsychiatry, Vincent van Gogh Institute for Psychiatry, Venray, the Netherlands. [45]MRGM, Maladies Rares: Génétique et Métabolisme, INSERM U1211, Université de Bordeaux, Bordeaux, France. [46]Service de Génétique Médicale, Centre Hospitalier Universitaire de Bordeaux, Bordeaux, France. [47]Department of Neuropediatrics and Muscle Disorders, Medical Center, Faculty of Medicine, University of Freiburg, Freiburg, Germany. [48]Division of Neurology, Department of Medicine, The Ottawa Hospital, Ottawa, Ontario, Canada. [49]Brain and Mind Research Institute, University of Ottawa, Ottawa, Ontario, Canada. [50]Institute of Rare Diseases Research, Spanish Undiagnosed Rare Diseases Cases Program (SpainUDP) & Undiagnosed Diseases Network International (UDNI), Instituto de Salud Carlos III, Madrid, Spain. [51]Pediatric Neurology Research Group, Vall d'Hebron Research Institute, Universitat Autònoma de Barcelona, Barcelona, Spain. [52]Clinical Institute of Genomic Medicine, University Medical Centre Ljubljana, Ljubljana, Slovenia. [53]Department of Human Genetics, Medical Research Institute, Alexandria University, Alexandria, Egypt. [54]Dubowitz Neuromuscular Centre, UCL Great Ormond Street Hospital, London, UK. [55]NIHR Great Ormond Street Hospital Biomedical Research Centre, London, UK. [56]Dipartimento di Medicina di Precisione, Università degli Studi della Campania "Luigi Vanvitelli", Naples, Italy. [57]Telethon Institute of Genetics and Medicine, Pozzuoli, Italy. [58]East Anglian Medical Genetics Service, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK. [59]Centre of Hereditary Ataxia, Department of Neurology, Charles University Prague–2nd Faculty of Medicine and University Hospital Motol, Prague, Czech Republic. [60]Institute of Systems Motor Science, University of Lübeck, Lübeck, Germany. [61]Institute of Neurogenetics, University of Lübeck, Lübeck, Germany. [62]Department of Neurology, University Hospital Schleswig Holstein, Lübeck, Germany. [63]Genetica Medica, Azienda Ospedaliero-Universitaria Senese, Siena, Italy. [64]Department of Biology and Medical Genetics, Second Faculty of Medicine, Charles University and Motol University Hospital, Prague, Czech Republic. [65]Institute of Data Science and Digital Technologies, Vilnius University, Vilnius, Lithuania. [66]Medizinische Klinik und Poliklinik IV – Campus Innenstadt, Klinikum der Universität München, Munich, Germany. [67]MGZ – Medical Genetics Center, Munich, Germany. [68]Molecular Genetics and Functional Genomics, Ospedale Pediatrico Bambino Gesù, IRCCS, Rome, Italy. [69]Tampere Neuromuscular Center, Tampere, Finland. [70]Vasa Central Hospital, Vaasa, Finland. [71]Peripheral Neuropathy Research Group, University of Antwerp, Antwerp, Belgium. [72]Department of Neurology, Radboud University Medical Center, Nijmegen, the Netherlands. [73]Institute of Clinical Genetics, University Hospital Carl Gustav Carus, Technical University Dresden, Dresden, Germany. [74]National Center for Tumor Diseases (NCT), Dresden, Germany. [75]Department of Genetics and Genome Biology, University of Leicester, Leicester, UK. [76]Centre for Rare Diseases, University of Tübingen, Tübingen, Germany. [77]NGS Competence Center Tübingen (NCCT), University of Tübingen, Tübingen, Germany. [78]Dept of Genetics, Assistance Publique-Hôpitaux de Paris, Université de Paris, Robert DEBRE University Hospital, Paris, France. [79]INSERM UMR 1141 "NeuroDiderot", Hôpital Robert DEBRE, Paris, France. [80]Department of Clinical Genetics, Maastricht University Medical Centre and GROW School for Development and Oncology, University of Maastricht, Maastricht, the Netherlands. [81]Departament de Genètica, Microbiologia i Estadística, Facultat de Biologia, Universitat de Barcelona (UB), Barcelona, Spain. [82]Department of Internal Medicine and Radboud Center for Infectious Diseases (RCI), Radboud University Medical Center, Nijmegen, the Netherlands. [83]These authors contributed equally: Steven Laurie, Wouter Steyaert, Elke de Boer, Kiran Polavarapu, Nika Schuermans, Anna K. Sommer. [84]These authors jointly supervised this work: Katja Lohmann, Richarda M. de Voer, Ana Töpf, Lisenka E. L. M. Vissers, Sergi Beltran, Alexander Hoischen. *Lists of authors and their affiliations appear at the end of the paper. ✉e-mail: sergi.beltran@cnag.eu; alexander.hoischen@radboudumc.nl

Solve-RD DITF-GENTURIS

Stefan Aretz[10,13], Richarda M. de Voer[3,4], José Garcia-Pelaez[34,35,36], Nicoline Hoogerbrugge[3,4], Carla Oliveira[35,36,37], Sophia Peters[10], Anna K. Sommer[10,83], Verena Steinke-Lange[66,67], Iris B. A. W. te Paske[3,4] & Doreen William[73,74]

Solve-RD DITF-ITHACA

Elke de Boer[3,5], Jill Clayton-Smith[21,22], Jean-Madeleine de Sainte Agathe[25], Anne-Sophie Denommé-Pichon[26,27], Laurence Faivre[26,30,31,32,33], Tobias B. Haack[11], Adam Jackson[21,22], Tjitske Kleefstra[3,5,43,44], Didier Lacombe[45,46], Estrella López-Martín[50], Vincenzo Nigro[56,57], Manuel Posada[50], Alessandra Renieri[17,18,63], Olaf Riess[11,76], Lukas Ryba[64], Annalaura Torella[56,57], Alain Verloes[78,79], Lisenka E.L.M. Vissers[3,5,84] & Antonio Vitobello[26,27]

Solve-RD DITF-EURO-NMD

Jonathan Baets[14,15,16], Patrick F. Chinnery[19,20], Enzo Cohen[23], Teresinha Evangelista[23], Rita Horvath[19], Henry Houlden[29], Mridul Johari[42], Hanns Lochmüller[1,6,47,48,49], Francesco Muntoni[54,55], Francesco Musacchia[56,57], Isabelle Nelson[23], Vincenzo Nigro[56,57], Catarina Olimpio[19,58], Giulio Piluso[56], Kiran Polavarapu[6,83], Marco Savarese[42], Rachel Thompson[6], Ana Töpf[28,84], Annalaura Torella[56,57], Bjarne Udd[42,69,70], Liedewei Van de Vondel[14,15,71] & Jana Vandrovcova[29]

Solve-RD DITF-RND

Jonathan Baets[14,15,16], Patrick F. Chinnery[19,20], Stephanie Efthymiou[29], Holm Graessner[11,76], Lena Guillot-Noel[37], Tobias B. Haack[11], Mike Hanna[38], Holger Hengel[39,40], Rita Horvath[19], Henry Houlden[29], Erik-Jan Kamsteeg[3], Melanie Kellner[39,40], Katja Lohmann[61,84], Alfons Macaya[51], Anna Marcé-Grau[51], Aleš Maver[52], Heba Morsy[29,53], Martje G. Pauly[60,61,62], Borut Peterlin[52], Selina Reich[39,40], Olaf Riess[11,76], Ludger Schöls[39,40], Rebecca Schüle[39,40], Nika Schuermans[7,8,9,83], Giovanni Stevanin[37], Matthis Synofzik[39,40], Liedewei Van de Vondel[14,15,71], Bart van de Warrenburg[5,72], Jana Vandrovcova[29], Carlo Wilke[39,40] & Jishu Xu[11,39,40]

Solve-RD consortium

Stefan Aretz[10,13], Jonathan Baets[14,15,16], Sergi Beltran[1,81,84], Elisa Benetti[17,18], Elke de Boer[3,5], Anthony J. Brookes[75], Han G. Brunner[3,5,80], Gemma Bullich[1,2], Patrick F. Chinnery[19,20], Jill Clayton-Smith[21,22], Enzo Cohen[23], Daniel Danis[24], Jean-Madeleine de Sainte Agathe[25], German Demidov[11], Anne-Sophie Denommé-Pichon[26,27], Jordi Diaz-Manera[28], Stephanie Efthymiou[29], Kornelia Ellwanger[11], Teresinha Evangelista[23], Laurence Faivre[26,30,31,32,33], Marcos Fernandez-Callejo[1,2], Mallory Freeberg[12], José Garcia-Pelaez[34,35,36], Christian Gilissen[3,4], Holm Graessner[11,76], Lena Guillot-Noel[37], Tobias B. Haack[11], Mike Hanna[38], Holger Hengel[39,40], Alexander Hoischen[3,4,82,84], Nicoline Hoogerbrugge[3,4], Rita Horvath[19], Henry Houlden[29], Adam Jackson[21,22], Lennart Johansson[41], Mridul Johari[42], Erik-Jan Kamsteeg[3], Melanie Kellner[39,40], Tjitske Kleefstra[3,5,43,44], Didier Lacombe[45,46], Steven Laurie[1,2,83], Hanns Lochmüller[1,6,47,48,49], Katja Lohmann[61,84], Estrella López-Martín[50], Alfons Macaya[51], Anna Marcé-Grau[51], Leslie Matalonga[1,2], Aleš Maver[52], Heba Morsy[29,53], Francesco Muntoni[54,55], Francesco Musacchia[56,57], Isabelle Nelson[23], Vincenzo Nigro[56,57], Carla Oliveira[35,36,37], Stephan Ossowski[11,77], Ida Paramonov[1,2], Martje G. Pauly[60,61,62], Borut Peterlin[52], Sophia Peters[10], Giulio Piluso[56], Davide Piscia[1,2], Kiran Polavarapu[6,83], Manuel Posada[50], Alessandra Renieri[17,18,63], Olaf Riess[11,76], Karolis Šablauskas[3,65], Marco Savarese[42], Ludger Schöls[39,40], Rebecca Schüle[39,40], Nika Schuermans[7,8,9,83], Anna K. Sommer[10,83], Verena Steinke-Lange[66,67], Giovanni Stevanin[37], Wouter Steyaert[3,4,83], Volker Straub[28], Marc Sturm[11], Morris A. Swertz[41], Matthis Synofzik[39,40], Marco Tartaglia[68], Iris B. A. W. te Paske[3,4], Coline Thomas[12], Rachel Thompson[6], Ana Töpf[28,84], Annalaura Torella[56,57], Bjarne Udd[42,69,70], Liedewei Van de Vondel[14,15,71], Bart van de Warrenburg[5,72], Jana Vandrovcova[29], Alain Verloes[78,79], Lisenka E. L. M. Vissers[3,5], Antonio Vitobello[26,27], Richarda M. de Voer[3,4], Carlo Wilke[39,40], Jishu Xu[11,39,40], Burcu Yaldiz[3] & Birte Zurek[11]

Full lists of members and their affiliations appear in the Supplementary Information.

Methods

Ethics oversight and enrollment

The ethics committee/IRB of University of Tübingen gave ethical approval for this work (ClinicalTrials.gov no. NCT03491280). Informed consent for data sharing, including indirect identifiers within Europe for the purpose of research, was obtained from all recruited individuals, and all data submitters signed the Adherence Agreement and Code of Conduct of RD-Connect GPAP. This covers the use of P-numbers that link to sample IDs only in an arbitrary fashion and have the function to allow traceability of results throughout the manuscript.

All individuals were recruited via four ERNs. Inclusion criteria were a clinical rare-disease diagnosis in at least one family member by one of the associated expert centers and an inconclusive exome or genome analysis at the time of submission. We did not exclude anyone based on sex, gender, ethnicity, race, age or any other socially relevant groupings.

Each patient entry was associated with its submitting investigator or clinician and linked to its corresponding ERN or UDP. The responsibility of checking that the data were suitable for submission to RD-Connect GPAP and Solve-RD lay with the data submitter, as required by their Code of Conduct (current institution: Consorcio para la Explotación del Centro Nacional de Análisis Genómico) and Data-sharing Policy (institution: Solve-RD general assembly), respectively. In some cases, individuals had to be reconsented before data submission. The individuals described in Extended Data Fig. 6 gave permission for their photographs to be used in this publication, for which we thank them and their families. This study adheres to the principles set out in the Declaration of Helsinki.

Family recruitment

Any undiagnosed individual with an apparent genetic rare disease that falls under the umbrella of conditions in which one of the four partner ERNs specialize, and for whom a previous ES analysis had been undertaken and proven inconclusive, was a candidate for inclusion in this study. The pan-European recruitment effort involved over 300 clinicians with expertise in rare-disease working in 43 research groups across 37 institutions located in 13 countries. To facilitate data submission and sharing, we implemented a pragmatic approach to collecting datasets to allow efficient reanalysis across centers. We refer to these datasets as RD-REAL, which must include genomic data, family information and phenotypic descriptions. The RD-REAL framework facilitates sharing of data and expertise at a continental scale, consisting of (1) expert curated data, (2) a comprehensive analysis suite and (3) two-level (that is, molecular and clinical) expert review (Fig. 1).

Data pertaining to 10,039 individuals from 6,246 undiagnosed families were initially assembled, which were then reduced to 9,645 individuals (6,447 affected) in 6,004 families following application of quality control measures, as described below. Of the 6,447 affected individuals, 3,592 (56%) were male and 2,855 (44%) female; 6,215 (96.4%) were alive at the start of the study, 84 (1.3%) were deceased and for 148 (2.3%) their vital status was unknown.

Pseudonymized phenotypic data collation for all individuals was facilitated using the PhenoStore module of RD-Connect GPAP. PhenoStore promotes deep phenotyping of affected individuals using HPO terms, and disease classification using Orphanet Rare Disease Ontology (ORDO) ORPHA codes (http://www.orphadata.org/cgi-bin/index.php) and/or OMIM identifiers (https://www.omim.org/) as appropriate, and can import/export this information using the GA4GH Phenopackets format[14].

ERN cohort descriptions

For all families recruited to Solve-RD, local standard-of-care genetic diagnostic work-up and/or research-based analyses had failed to identify any molecular genetic cause underlying the proband's rare condition.

ERN RND

The ERN RND cohort consists of 2,799 individuals from 2,271 families with previously unsolved rare neurological diseases. Genomic and phenotypic data for all affected individuals, and for family members where available (~20% of families), were submitted for reanalysis by nine ERN RND partner institutions located in eight European countries: Belgium, France, Germany, Hungary, the Netherlands, Slovenia, Spain and the UK. Individuals had been recruited and sequenced either as part of standard diagnostic care or through participation in large European rare-neurological disease research projects such as NeurOmics (https://rd-neuromics.eu/) and Treat-HSP (https://www.treathsp.net/). The 2,271 families comprised 1,924 singletons, 168 duos, 141 triples (103 of which were parent–child trios) and 38 families with four or more members, giving a total of 2,453 affected individuals. The HPO terms most frequently used to describe phenotypes were ataxia, gait disturbance, dysarthria and spastic paraplegia (Supplementary Table 12).

ERN ITHACA

The ERN ITHACA cohort consists of 4,405 individuals from 1,836 families, submitted for reanalysis by 12 partner institutions located in six countries: the Czech Republic, France, Germany, Italy, the Netherlands and the UK. A further 65 individuals from 21 families from the Spanish Undiagnosed Disease Program (SpainUDP)[6] were included in this cohort for analysis, due to the similarity of the underlying phenotypes. The clinical spectrum of the ERN ITHACA cohort consisted of individuals with intellectual disability (ID) with or without additional phenotypic features, and individuals with (multiple) congenital anomalies without ID. Given the importance of de novo mutations underlying the rare conditions within ERN ITHACA[34,53], unaffected parents and/or unaffected siblings were also included, wherever possible, to allow for direct segregation of variants. The 1,857 families comprised 632 singletons, 38 duos, 1,138 triples (1,081 parent–child trios) and 49 families with four or more members, giving a total of 1,933 affected individuals. The HPO terms most frequently used to describe affected individuals related to global developmental decay, intellectual disability and autism (Supplementary Table 12).

ERN EURO-NMD

The ERN EURO-NMD cohort consists of 2,125 individuals from 1,517 families, submitted for reanalysis by 16 partner institutions located in eight countries: Belgium, Canada, Finland, France, Germany, Italy, Spain and the UK. Previously unsolved datasets submitted to Solve-RD had either been recruited and sequenced as part of large international neuromuscular research projects, such as NeurOmics (https://rd-neuromics.eu/), SeqNMD, Myocapture (https://www.france-genomique.org/projet/myocapture-novel-for-genes-myopathies/?lang=en), MYO-SEQ[54], UK10K (https://www.uk10k.org/), Unravel-CMS, BBMRI-LPC (https://cordis.europa.eu/project/id/313010), CMS CMG (https://cmg.broadinstitute.org/) or Consequitur[55], or through participating centers' own diagnostic or research pipelines. Samples incorporated from the MYO-SEQ project were recruited from 50 specialized neuromuscular disease centers across Europe and the Middle East, and some datasets incorporated from the Unravel-CMS, BBMRI-LPC and CMS CMG projects were from privately sequenced undiagnosed individuals followed at Nimhans, India (https://nimhans.ac.in/). The 1,517 families comprised 1,202 singletons, 90 duos, 156 triples (135 parent–child trios) and 69 families with four or more members, giving a total of 1,685 affected individuals. The HPO terms most frequently used to describe affected individuals related to muscle weakness, myopathy and abnormal muscle morphology (Supplementary Table 12).

ERN GENTURIS

The ERN GENTURIS cohort consists of 390 individuals, from 359 families, with a suspected genetic tumor risk syndrome, submitted for reanalysis by seven partner institutions located in four countries:

Germany, the Netherlands, Portugal and Spain. All individuals were either recruited and sequenced as part of daily diagnostic care, or as part of research projects. The 359 families comprised 345 singletons, six duos, four triples (one parent–child trio) and four families with four or more members, giving a total of 378 affected individuals. The terms most frequently used to describe affected individuals related to colorectal cancer, followed by gastric cancer and pheochromocytoma (Supplementary Table 12).

Phenotype and clinical diagnosis

A median of six HPO terms (range 0–74) were used to describe each affected individual across this Solve-RD cohort. This drops to five HPO terms (range 0–45) following removal of HPO redundancies. To remove annotation redundancy, only the most specific HPO terms were considered by counting terms from leaf nodes, or nodes without selected parent or child entities. Overall quality of phenotypic descriptions was assessed using the Monarch Initiative annotation sufficiency score (maximum possible value of 5.0). The median annotation sufficiency value across the Solve-RD cohort was 3.61 (Extended Data Fig. 1). Clinical diagnosis was reported using ORDO codes for 2,126 affected individuals.

Generation of ERN-specific candidate gene lists

To facilitate the potential for clinicians to confirm a diagnosis based on identified variants, findings returned to the ERN data interpretation task forces (DITFs) for interpretation were restricted to those in disease genes of interest to the specific ERN, apart from any potentially pathogenic variants encountered in the mitochondrial genome, all of which were returned. Each of the four ERNs generated a curated list of genes implicated in diseases studied, exploiting their pan-European disease expertise. The RND list was primarily based on genes associated with neurological disease with green review status in Genomics England PanelApp[56], with the addition of a further 25 genes based on recommendations by clinical experts (n = 1,821 genes). For ITHACA, a consolidation of gene lists pertaining to ID from a variety of resources was undertaken, followed by evaluation based on occurrence in multiple resources and the quality of curation of said resources, resulting in a list of diagnostically relevant genes (n = 1,645). In the case of GENTURIS, the list included all genes routinely screened in the partners' diagnostic laboratories (n = 230). For EURO-NMD, the manually curated and annually updated Gene Table of Muscular Disorders[57] was used (n = 615 in 2021). These ERN gene lists were used as a primary filter in the identification of potentially pathogenic variants of any type in affected individuals submitted to Solve-RD by collaborators from the corresponding ERN, irrespective of the individual's phenotype. This resulted in a list of 2,512 distinct genes implicated in rare diseases of interest to the four ERNs, many of which were identified by more than one ERN (Supplementary Table 2).

Identification of clinically actionable genes

Potentially clinically actionable genes in affected individuals were identified from three independent initiatives: ClinGen[28] (n = 77), IEMbase[58] (n = 214) and Treatabolome[59] (n = 154; https://treatabolome.cnag.crg.eu). This provided a total of 392 unique genes, of which 311 (79%) were included in at least one of the curated gene lists from the ERNs. For the assessment of clinically actionable genes in individuals affected by a hereditary cancer disposition, we searched GeneReviews and the National Comprehensive Cancer Network Clinical Practice Guidelines in Oncology (https://www.nccn.org/guidelines/category_1) for actionability based on surveillance for cancer advice.

Data submission and analysis workflow

Raw sequencing data, associated metadata and phenotypic and pedigree descriptions were collated from 43 research groups across Europe using RD-Connect GPAP[8]. To ensure secure, rapid and robust transfer of the large quantity of raw genomic data (FASTQ, BAM or CRAM) for reanalysis (approximately 100 TB in total), each research group was provided with access to a dedicated private space in which to upload their sequencing data, on an Aspera server hosted by RedIRIS, the Spanish national research and education network (https://www.rediris.es/). From here the sequencing data were downloaded to the Centro Nacional de Análisis Genómico in Barcelona, which develops and hosts RD-Connect GPAP.

All genomic data submitted to Solve-RD were analyzed in identical fashion to minimize any batch effects, using the RD-Connect GPAP standard analysis pipeline[60]. Briefly, reads were aligned to the decoy version of GRCh37 (hs37d5) using BWA-MEM. Short variants (that is, SNVs) and insertions and deletions <50 nt in length (referred to here as InDels*)* were identified across the genome, independent of the target capture region of interest, using the GATK HaplotypeCaller in accord with the GATK Best Practices workflow. The output of the pipeline for each experiment is an aligned, base quality score recalibrated BAM, and a genetic variant call format (gVCF) per chromosome and for the mitochondrion. All variant positions covered by at least eight reads, and a GATK-assigned genotype quality of at least 20, are uploaded to RD-Connect GPAP, as are any nonvariant positions for which at least one other experiment in the uploaded batch has a variant position at the same genomic location. SNVs, InDels and mitochondrial variants received detailed annotations provided by Ensembl Variant Effect Predictor[61], gnomAD[62] and ClinVar[16], among other resources.

In addition to the above described annotations available through RD-Connect GPAP, all gVCFs derived from affected individuals were converted to VCFs and annotated by a custom annotation pipeline at RadboudUMC, as described previously[63]. This comprises variant-based annotations, including nucleotide conservation scores (phyloP and CADD), RadboudUMC in-house database allele frequencies and gene-based annotations including, for example, mouse knockout model phenotypes and pLI/LOEUF scores, among others. These annotated VCF files were made available to the Solve-RD consortium through the Solve-RD Sandbox, a cloud environment used by project partners to conduct bespoke analyses and thereby to securely share analysis and interpretation results, hosted by UMC Groningen, the Netherlands. A more detailed description of the Solve-RD data infrastructure has been published previously[64].

Raw sequencing data (FASTQ), and newly generated alignment (BAM or CRAM) and variant call (gVCF) files for each experiment, accompanied by the corresponding phenotypic description in Phenopackets and pedigree descriptions in PLINK PED format, were submitted to EGA[9] in Hinxton, UK for long-term archival and to allow controlled access by the wider human genomics community.

Quality control

A total of 10,276 ES and GS RD-REAL datasets from 10,039 individuals were initially submitted to Solve-RD for reanalysis. Preliminary quality control of sequencing data required a median coverage of at least ten reads over at least 70% of the defined target region of interest for the corresponding enrichment kit, or across the entire genome in the case of GS data. Furthermore, with respect to phenotypic data, each submitted family was required to have an affected proband with associated HPO terms. Misassigned relationships were identified, and subsequently corrected where possible, using KING (https://www.kingrelatedness.com/). Following application of these quality control measures, the final number of datasets taken forward for reanalysis comprised data from 9,645 individuals from 6,004 families, of which 6,447 (66.9%) were affected by a rare disease. Of these, ES data were available for 9,124 (94.6%) individuals, GS data for 333 (3.5%) and both ES and GS data for the remaining 190 (2.0%).

Variant identification and prioritization

RD-REAL data analysis and interpretation. We applied two-level expert analysis and interpretation to the RD-REAL datasets, comprising firstly bioinformatic and molecular genetics experts working together in dedicated working groups within a data analysis task force, and secondly, clinicians and rare-disease experts from each ERN who jointly prioritized and interpreted all variants returned by the data analysis task force, working in four distinct DITFs. To maximize the generalizability of this effort, the entire dataset of 6,004 families was included in a comprehensive analysis suite comprising an initial centralized analysis of each different variant type: short SNVs and InDels; de novo mutations; and mitochondrial variants, noncanonical splice variants, CNVs, SVs, STRs and MEIs. Subsequently, filters were applied with respect to variant quality, control population allele frequencies and predicted consequence, followed by further ERN- and disease-specific filters including the application of the ERN-specific gene lists described above. Details of all tools applied in these analyses are provided in Supplementary Table 13.

Because Solve-RD processed data in multiple data freezes, subsets of experiments continued to undergo analyses in parallel, some of which resulted in diagnoses before the results of the centralized systematic analyses were returned to submitters. This includes the preliminary analysis of a smaller dataset[12,13]. Furthermore, many datasets underwent parallel or additional analyses in the laboratories of the respective submitters, resulting in the identification of (probable) pathogenic, or candidate disease-causing, variants in established or novel genes. These results are labeled as ad hoc expert review (Fig. 2 and Supplementary Table 5), although the majority of these variants were also prioritized in the systematic analyses.

Taken together, this resulted in either diagnosed individuals (that is, those harboring (probable) pathogenic variants that fully explain the proband's phenotype, unequivocally allowing a molecular diagnosis of a rare condition) or affected individuals with candidate variants worthy of further follow-up and/or functional studies, which may prove to be diagnostic in the future, as adjudged by the referring clinicians and/or expert ERN partners.

SNVs/InDels. Programmatic reanalysis was undertaken on annotated variants from RD-Connect GPAP using application programming interface endpoints, enabling complex queries with different combinations of filters across specific datasets[13]. Two different sets of parameters were used: first, a low-hanging fruit analysis to identify (probable) pathogenic variants already listed in ClinVar; second, identification of rare variants of high or moderate impact in ERN genes of interest, matching the expected mode(s) of inheritance.

(1) Low-hanging fruit analysis: depth of coverage (DP) >7; GATK genotype quality (GQ) >19; minor allele frequency (MAF) <0.01 in gnomAD; observed allele frequency <0.02 in the internal RD-Connect GPAP database; affecting a gene in the corresponding ERN gene list, and annotated as pathogenic (class 5) or probably pathogenic (class 4) for any disorder in ClinVar as of May 2021.

(2) High–moderate-impact variant analysis: DP >7; GQ >19; MAF <0.01 in gnomAD; observed allele frequency <0.02 in the internal RD-Connect GPAP database; affecting a gene in the corresponding ERN gene list, predicted to have a high or moderate consequence at the protein level according to Ensembl VEP and matching the expected inheritance pattern (that is, autosomal dominant, autosomal recessive or X-linked).

Variants passing the above filtering criteria were returned in a single table to the respective DITF for each ERN, to facilitate evaluation and provision of feedback. Across the Solve-RD cohort we identified a mean of eight SVs per affected individual for interpretation, ranging from one to 13 across ERNs, this difference largely reflecting differences in the number of genes included in the corresponding ERN gene lists (Supplementary Tables 2 and 14).

De novo mutations. For all families for which parent-child trios were available (n=1,320; 22% overall), de novo mutation calling was undertaken using both HaplotypeCaller and DeNovoCNN[65]. De novo mutation calls from DeNovoCNN with probability >0.85 of being a bona fide event, and any apparent de novo mutations identified by HaplotypeCaller which were located in a gene on the respective ERN gene list, were returned to DITFs for variant interpretation.

Mitochondrial genome variants. Mitochondrial DNA variants were identified using MToolBox.The workflow includes mapping reads to the revised Cambridge Reference Sequence mitochondrial genome and annotation using the MITOMAP database (https://www.mitomap.org/MITOMAP, accessed 28 June 2021). Both homoplasmic and heteroplasmic variants were identified (Supplementary Table 15).

Identification of noncanonical SVs. For identification of variants potentially affecting splicing at sites other than canonical splice sites, two novel tools were applied, SpliceAI[29] and SQUIRLS[66]. Rare variants receiving a strong splice-altering prediction from both tools (that is, both a delta-score >0.8 in SpliceAI and a pathogenic classification by SQUIRLS, which would potentially alter splicing of any gene in the corresponding ERN gene list) were returned to DITFs for interpretation.

Large CNVs and SVs. Three different tools were used to maximize the likelihood of identifying pathogenic CNVs, as described in Demidov et al.[36]: ClinCNV[67], Conifer and ExomeDepth. Variants observed to have a frequency >0.01 across the cohort were discarded, and the remaining rare CNVs were intersected with the corresponding ERN gene list and annotated using AnnotSV[68] before being returned to DITFs for interpretation. In parallel, Manta[37] was run in exome mode to search for signatures of split reads, which might indicate the presence of balanced SV such as inversions. To facilitate interpretation, Integrative Genomics Viewer (IGV) tracks were generated for all large variants, indicating the exons, the position and type of call produced by the tools and beta-allele frequency. See Supplementary Table 13 for details regarding sources and exact versions of tools applied.

STR expansions. The identification of potentially pathogenic STR expansions was largely based on the work of van der Sanden et al.[69]. In brief, ExpansionHunter[70] was used to screen 21 genomic loci previously described as harboring pathogenic repeat expansions in both ES and GS data (Supplementary Tables 16 and 17), from a total of 5,983 families. Following retrieval of predicted pathogenic genotypes across all samples, any frequently observed events were discarded and the remaining variants affecting genes on the corresponding ERN gene list were manually curated by visual inspection, before being returned to DITFs for interpretation.

MEIs. To identify any MEIs potentially affecting ERN genes of interest, the methods described by Wijngaard et al.[71] were followed. In brief, MEI identification was undertaken using both MELT and SCRAMble. MEIs of potential interest were limited to those that fell within a window of ±50 base pairs (bp) of ES target areas. All MEIs observed in nonaffected cases were removed, followed by the exclusion of those present in the Database of Retrotransposon Insertion Polymorphisms in Humans. MEI frequency was calculated by counting all overlapping (±50 bp) MEIs in the cohort, and only rare events—defined as having a frequency <0.03% in their respective cohorts—were retained. We further filtered to MEIs found in clinically relevant genes based on the patient's phenotype as defined by the ERN. The remaining MEIs were visually inspected in IGV to discard low-quality calls. Finally, MEIs were selected for confirmation by ERN members, taking into consideration the phenotype–genotype match, inheritance pattern and presence of a second variant in the case of an autosomal recessive disorder.

An overview of the accurate number of families analyzed for each variant type is provided in Supplementary Table 18.

Statistics and reproducibility

This study includes only observational statistics, primarily counts. We report means and medians where appropriate, and applied two-tailed Fisher's exact tests to compare differences between groups. Each family was analyzed independently, in order of submission, and no statistical method was required to predetermine sample size. No data were excluded, with the exception of cases that failed quality control as described above. Sex is not a relevant variant, because both sexes are essentially equally likely to be affected by a rare disease. The investigators were not blinded to allocation during outcome assessment.

Reproduction of results was not applicable. However, follow-up and validation of identified variants by orthologous means and/or using other bioinformatic tools were undertaken in the vast majority of cases, to ensure that the variants identified were biologically real and relevant. As commonly found in the rare-disease field, replication of previously variant observations has happened, or will happen, via databases (for example, ClinVar) or the scientific literature.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Access to pseudonymized phenotypic information for all individuals and their genetic variants is possible through RD-Connect GPAP (https://platform.rd-connect.eu/), on completion of registration and approval by the independent RD-Connect Data Access Committee (Code of Conduct and registration details can be found at https://platform.rd-connect.eu/userregistration/). All raw and processed data files (FASTQs, BAM/CRAMs, gVCFs, PED and Phenopackets) are available at the European Genome-Phenome Archive (https://ega-archive.org/datasets/; datasets EGAD00001009767, EGAD00001009768, EGAD00001009769 and EGAD00001009770, under the Solve-RD study EGAS00001003851), following approval from the Solve-RD Data Access Committee. Confirmed causative variants were submitted to ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/) under the following accession nos: SCV005091231–SCV005091564, SCV005199960–SCV005200075 and SCV005200692–SCV005200738.

Code availability

All analysis was undertaken using previously published tools and resources. No custom code was used. Details of all tools applied in these analyses, and relevant repositories, are provided in Supplementary Table 13.

References

53. Gilissen, C. et al. Genome sequencing identifies major causes of severe intellectual disability. *Nature* **511**, 344–347 (2014).
54. Töpf, A. et al. Sequential targeted exome sequencing of 1001 patients affected by unexplained limb-girdle weakness. *Genet. Med.* **22**, 1478–1488 (2020).
55. Hiz Kurul, S. et al. High diagnostic rate of trio exome sequencing in consanguineous families with neurogenetic diseases. *Brain* **145**, 1507–1518 (2022).
56. Martin, A. R. et al. PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels. *Nat. Genet.* **51**, 1560–1565 (2019).
57. Benarroch, L., Bonne, G., Rivier, F. & Hamroun, D. The 2023 version of the gene table of neuromuscular disorders (nuclear genome). *Neuromuscul. Disord.* **33**, 76–117 (2023).
58. Lee, J. J. Y., Wasserman, W. W., Hoffmann, G. F., Van Karnebeek, C. D. M. & Blau, N. Knowledge base and mini-expert platform for the diagnosis of inborn errors of metabolism. *Genet. Med.* **20**, 151–158 (2018).
59. Bonne, G. The Treatabolome, an emerging concept. *J. Neuromuscul. Dis.* **8**, 337–339 (2021).
60. Laurie, S. et al. From wet-lab to variations: concordance and speed of bioinformatics pipelines for whole genome and whole exome sequencing. *Hum. Mutat.* **37**, 1263–1271 (2016).
61. McLaren, W. et al. The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).
62. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
63. Lelieveld, S. H. et al. Meta-analysis of 2,104 trios provides support for 10 new genes for intellectual disability. *Nat. Neurosci.* **19**, 1194–1196 (2016).
64. Johansson, L. F. et al. An interconnected data infrastructure to support large-scale rare disease research. *Gigascience* **13**, giae058 (2024).
65. Khazeeva, G. et al. DeNovoCNN: a deep learning approach to de novo variant calling in next generation sequencing data. *Nucleic Acids Res.* **50**, e97 (2022).
66. Danis, D. et al. Interpretable prioritization of splice variants in diagnostic next-generation sequencing. *Am. J. Hum. Genet.* **108**, 1564–1577 (2021).
67. Demidov, G., Sturm, M. & Ossowski, S. ClinCNV: multi-sample germline CNV detection in NGS data. Preprint at *bioRxiv* https://doi.org/10.1101/2022.06.10.495642 (2022).
68. Geoffroy, V. et al. AnnotSV: an integrated tool for structural variations annotation. *Bioinformatics* **34**, 3572–3574 (2018).
69. van der Sanden, B. P. G. H. et al. Systematic analysis of short tandem repeats in 38,095 exomes provides an additional diagnostic yield. *Genet. Med.* **23**, 1569–1573 (2021).
70. Dolzhenko, E. et al. Detection of long repeat expansions from PCR-free whole-genome sequence data. *Genome Res.* **27**, 1895–1903 (2017).
71. Wijngaard, R. et al. Mobile element insertions in rare diseases: a comparative benchmark and reanalysis of 60,000 exome samples. *Eur. J. Hum. Genet.* **32**, 200–208 (2024).

Acknowledgements

The Solve-RD consortium thanks all involved rare-disease 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 (to all authors). This research is supported (not financially) by four ERNs: ERN ITHACA (project ID no. 101085231), ERN RND (project ID no. 101155994), ERN EURO-NMD (project ID no. 101156434) and ERN GENTURIS (project ID no. 101155809). All ERNs are cofunded by the European Union within the framework of the Third Health Programme. The RD-Connect GPAP was developed under funded project FP7/2007–2013 (grant agreement no. 305444) and ongoing funding from EJP RD (grant nos. H2020 779257 and H2020 825575), further supported by Instituto de Salud Carlos III (grant nos. PT13/0001/0044 and PT17/0009/0019), Instituto Nacional de Bioinformática, ELIXIR-EXCELERATE (grant no. EU H2020 676559); and ELIXIR Implementation Studies (remote real-time visualization of human rare disease genomics data, RD-Connect, stored at EGA ELIXIR 2017–2018, ELIXIR IT-2017-INTEGRATION, Rare Disease Infrastructure ELIXIR 2019–2020 and Beacon ELIXIR 2019–2021). The RD-Connect GPAP has leveraged developments funded through the project VEIS (no. 001-P-001647, cofinanced by the European Regional Development Fund of the European Union within the framework of the Operational Program FEDER of Catalonia 2014–2020, and with the support of

Secretaria d'Universitats i Recerca del Departament d'Empresa i Coneixement de la Generalitat de Catalunya) and URD-Cat (no. PERIS SLT002/16/00174, Departament de Salut, Generalitat de Catalunya). The Spanish academic and research network RedIris (https://www.rediris.es/) provided the Aspera service used for uploading raw data for processing to the RD-Connect GPAP, and for transfer of data between centers. We further acknowledge support from the Netherlands Science Organizations (grant no. NWO VIDI 917.164.55 to C.G.), Ministero della Salute (Genoma mEdiciNa pERsonalizzatA, grant no. T3-AN-04, to V.N., A.R. and M.T.), the Network for Italian Genomes, Cell lines and DNA bank of Rett Syndrome, X-linked mental retardation and other genetic diseases, member of the Telethon Network of Genetic Biobanks (project no. GTB12001) and the EuroBioBank network. H.L. receives support from the Canadian Institutes of Health Research (CIHR) for Foundation Grant no. FDN-167281 (Precision Health for Neuromuscular Diseases), Transnational Team Grant no. ERT-174211 (ProDGNE) and Network Grant no. OR2-189333 (NMD4C), the Canada Foundation for Innovation (no. CFI-JELF 38412), the Canada Research Chairs program (Canada Research Chair in Neuromuscular Genomics and Health, no. 950-232279), the European Commission (grant no. 101080249), the Canada Research Coordinating Committee New Frontiers in Research Fund (no. NFRFG-2022-00033) for SIMPATHIC and from the Government of Canada First Research Excellence Fund (CFREF) for the Brain–Heart Interconnectome (no. CFREF-2022-00007). K.P. is a recipient of a CIHR postdoctoral fellowship under award no. MFE-491707. This work was also supported by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) no. 441409627, as part of the PROSPAX consortium under the framework of the European Joint Programme on Rare Diseases, under EJP RD COFUND-EJP no. 825575 (to M. Synofzik, R.S. and R.H.) and the Clinician Scientist programme PRECISE.net, funded by Else Kröner-Fresenius-Stiftung (to C.W., M.K., R.S. and M. Synofzik). J.P.S. was financed by Programme EXCELES (ID project no. LX22NPO5107), funded by the European Union – Next Generation EU. B.v.d.W. is supported by ZonMW, the Gossweiler Foundation and Hersenstichting. The work of F. Muntoni was also supported by Muscular Dystrophy UK and Muscular Dystrophy USA. H.G. and T.B.H. are supported by the European Union's Horizon 2020 research and innovation program, project Recon4IMD (grant no. 101080997). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

S.L., W.S., G.D., I.P., R.H., C.G., S.O., L.M., S.B. and A.H. led data analysis. E.d.B., K.P., N.S., A.K.S., S.A., J.B., E.B., G.B., P.F.C., J.C.-S., E.C., D.D., J.-M.d.S.A., A.-S.D.-P., J.D.-M., S.E., L.F., J.G.-P., L.G.-N., T.B.H., M.H., H. Hengel, R.H., H. Houlden, A.J., M.J., E.-J.K., M.K., T.K., D.L., H.L., E.L.-M., A. Macaya, A.M.-G., A. Maver, H.M., F. Muntoni, F. Musacchia, I.N., V.N., C. Olimpio, C. Oliveira, J.P.S., M.G.P., B.P., S.P., R.P., G.P., M.P., S.R., A.R., L.R., K.S., M. Savarese, L. Schöls, L. Schütz, V.S.-L., G.S., V.S., M. Sturm, M.T., I.B.A.W.t.P., R.T., A. Torella, C. Trainor, B.U., L.V.d.V., B.v.d.W., J.v.R., J. Vandrovcova, A. Vitobello, J. Vos, E.V., R.W., C.W., D.W., J.X., B.Y., T.E., H.G., N.H., O.R., R.S., M. Synofzik, A. Verloes, K.L., R.M.d.V., A. Topf, L.E.L.M.V., S.B. and A.H. contributed to data analysis and/or clinical and genomic interpretation. S.L., W.S., C. Thomas, M.F.-C., M.F., L.J., D.P., M.A.S., L.Z., A.J.B., L.M. and S.B. contributed to data processing and infrastructure. S.L., W.S., E.d.B., K.P., N.S., A.K.S., K.E., K.L., R.M.d.V., A. Topf, L.E.L.M.V., S.B. and A.H. contributed to data collation. S.L., W.S., E.d.B., K.P., N.S., A.K.S., H.G.B., K.L., R.M.d.V., A. Topf, L.E.L.M.V., S.B. and A.H. contributed to writing of the manuscript. R.M.d.V. designed all figures. B.Z., A.J.B., T.E., C.G., H.G., N.H., S.O., O.R., R.S., M. Synofzik, A. Verloes, L.M., H.G.B., K.L., R.M.d.V., A. Topf, L.E.L.M.V., S.B. and A.H. were responsible for supervision and design of the study.

Competing interests

M. Synofzik has received consultancy honoraria from Janssen, Ionis, Orphazyme, Servier, Reata, GenOrph and AviadoBio, all unrelated to the present manuscript. B.v.d.W. has received consultancy honoraria from, and/or has served on advisory boards for, Servier, Biohaven Pharmaceuticals, Vico Therapeutics and Biogen, all unrelated to the present manuscript. The other authors declare no competing interests.

Additional information

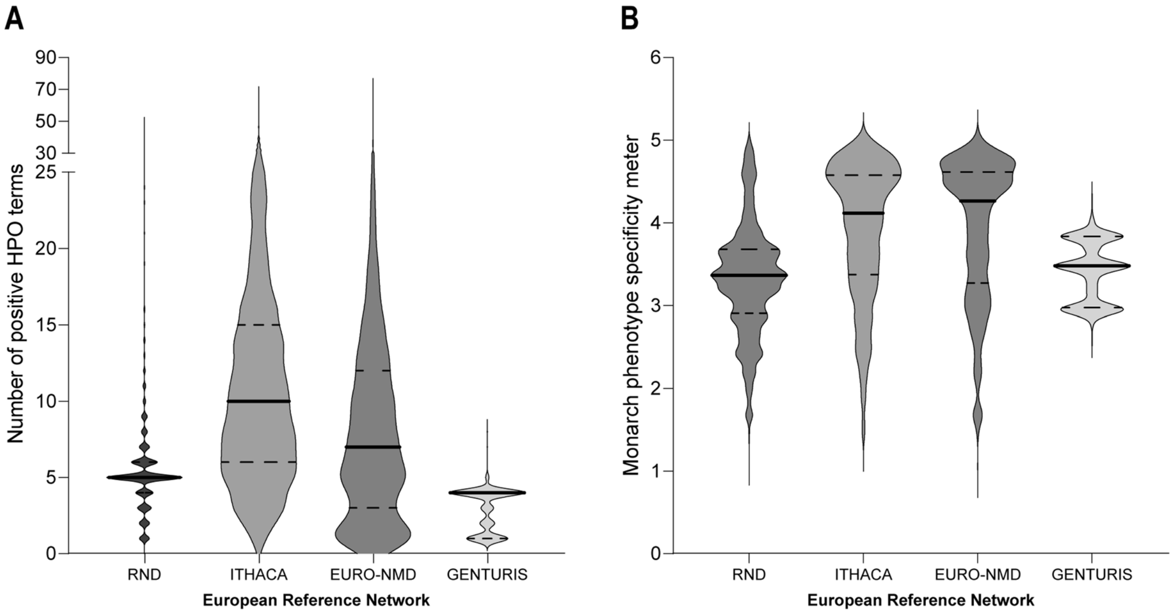
Extended data is available for this paper at https://doi.org/10.1038/s41591-024-03420-w.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41591-024-03420-w.

Correspondence and requests for materials should be addressed to Sergi Beltran or Alexander Hoischen.

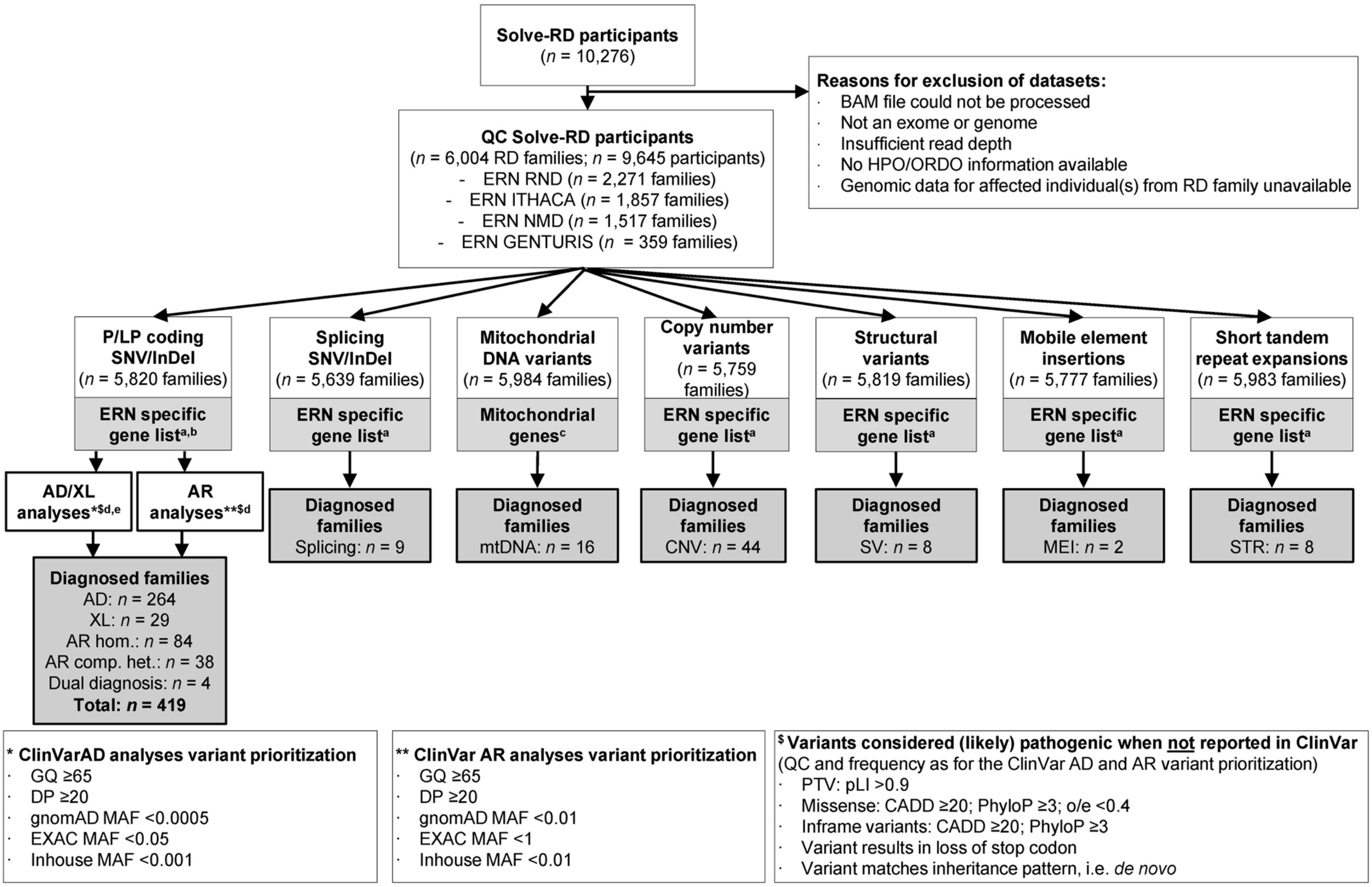
Peer review information *Nature Medicine* thanks Mark Cowley, Zornitza Stark and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: Anna Maria Ranzoni, in collaboration with the *Nature Medicine* team.

Reprints and permissions information is available at www.nature.com/reprints.

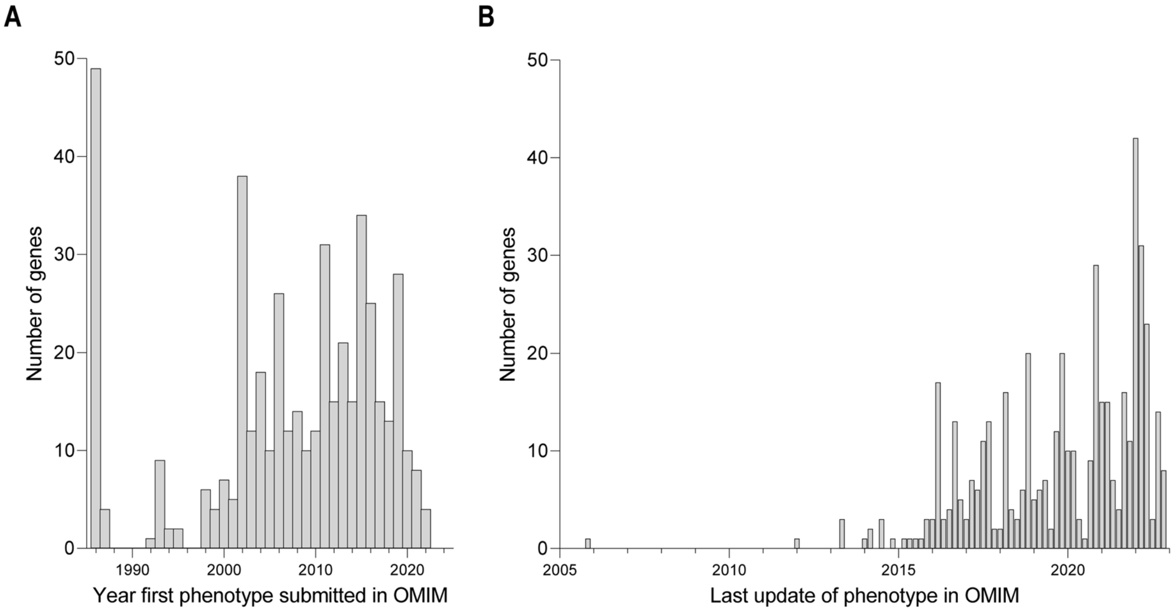


Extended Data Fig. 1 | HPO terms (A) and Monarch phenotype specificity meter (B). Violin plots illustrating (**a**) the number of Human Phenotype Ontology terms associated to each proband across ERN and (**b**) the Monarch specificity score (range 0–5, higher better) which provides an indication of how comprehensive the phenotypic description of the affected individual is. The solid line indicates the median, and the dashed line the 25th and 75th centiles.

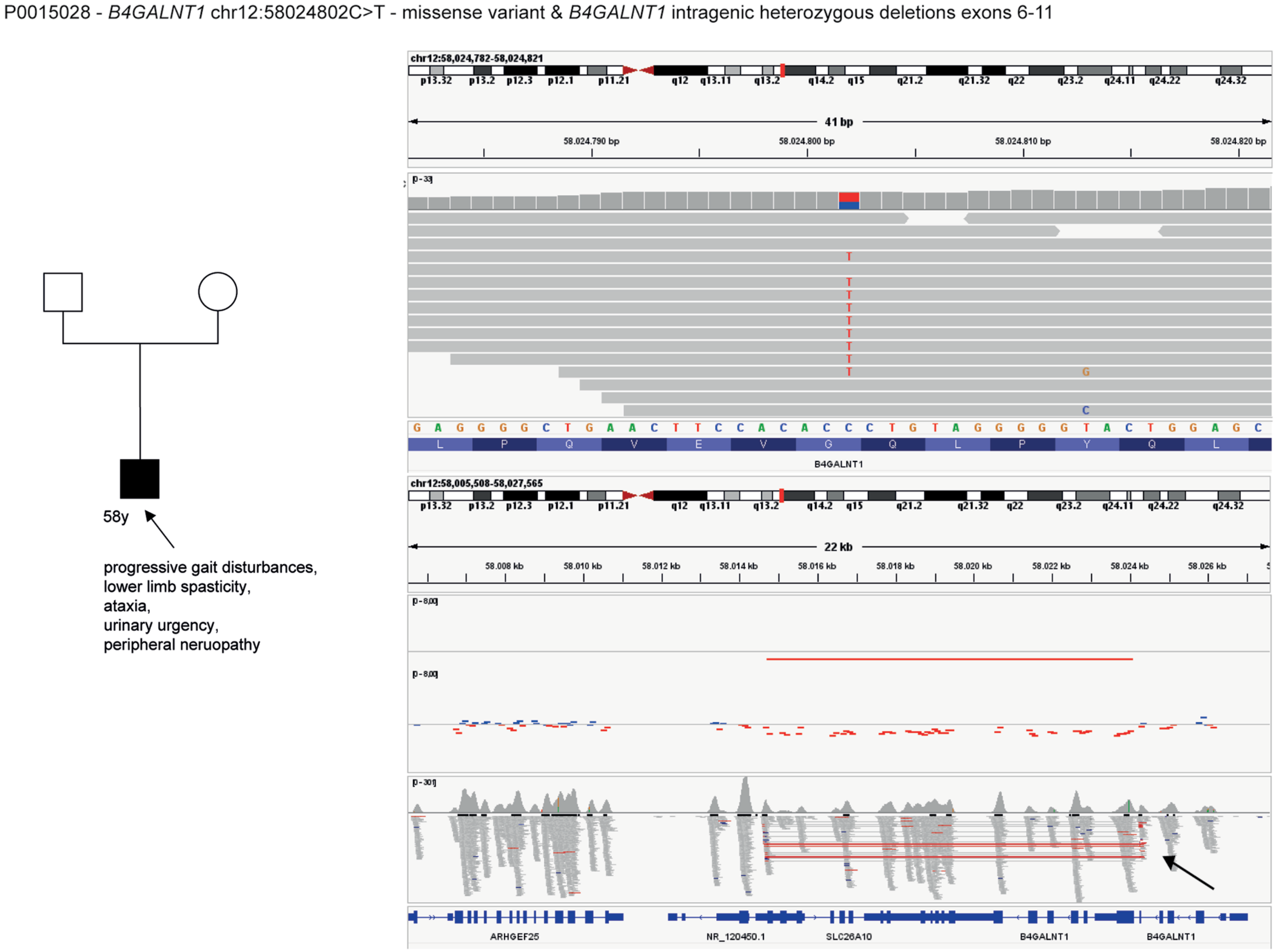
https://doi.org/10.1038/s41591-024-03420-w


Solve-RD participants
(n = 10,276)
Reasons for exclusion of datasets:
BAM file could not be processed
Not an exome or genome
Insufficient read depth
No HPO/ORDO information available
Genomic data for affected individual(s) from RD family unavailable
QC Solve-RD participants
(n = 6,004 RD families; n = 9,645 participants)
- ERN RND (n = 2,271 families)
- ERN ITHACA (n = 1,857 families)
- ERN NMD (n = 1,517 families)
- ERN GENTURIS (n = 359 families)
P/LP coding SNV/InDel
(n = 5,820 families)
ERN specific gene list[a,b]
Splicing SNV/InDel
(n = 5,639 families)
ERN specific gene list[a]
Mitochondrial DNA variants
(n = 5,984 families)
Mitochondrial genes[c]
Copy number variants
(n = 5,759 families)
ERN specific gene list[a]
Structural variants
(n = 5,819 families)
ERN specific gene list[a]
Mobile element insertions
(n = 5,777 families)
ERN specific gene list[a]
Short tandem repeat expansions
(n = 5,983 families)
ERN specific gene list[a]
AD/XL analyses[*$d,e]
AR analyses[**$d]
Diagnosed families
Splicing: n = 9
Diagnosed families
mtDNA: n = 16
Diagnosed families
CNV: n = 44
Diagnosed families
SV: n = 8
Diagnosed families
MEI: n = 2
Diagnosed families
STR: n = 8
Diagnosed families
AD: n = 264
XL: n = 29
AR hom.: n = 84
AR comp. het.: n = 38
Dual diagnosis: n = 4
Total: n = 419
* ClinVarAD analyses variant prioritization
GQ ≥65
DP ≥20
gnomAD MAF <0.0005
EXAC MAF <0.05
Inhouse MAF <0.001
** ClinVar AR analyses variant prioritization
GQ ≥65
DP ≥20
gnomAD MAF <0.01
EXAC MAF <1
Inhouse MAF <0.01
$ Variants considered (likely) pathogenic when not reported in ClinVar
(QC and frequency as for the ClinVar AD and AR variant prioritization)
PTV: pLI >0.9
Missense: CADD ≥20; PhyloP ≥3; o/e <0.4
Inframe variants: CADD ≥20; PhyloP ≥3
Variant results in loss of stop codon
Variant matches inheritance pattern, i.e. de novo

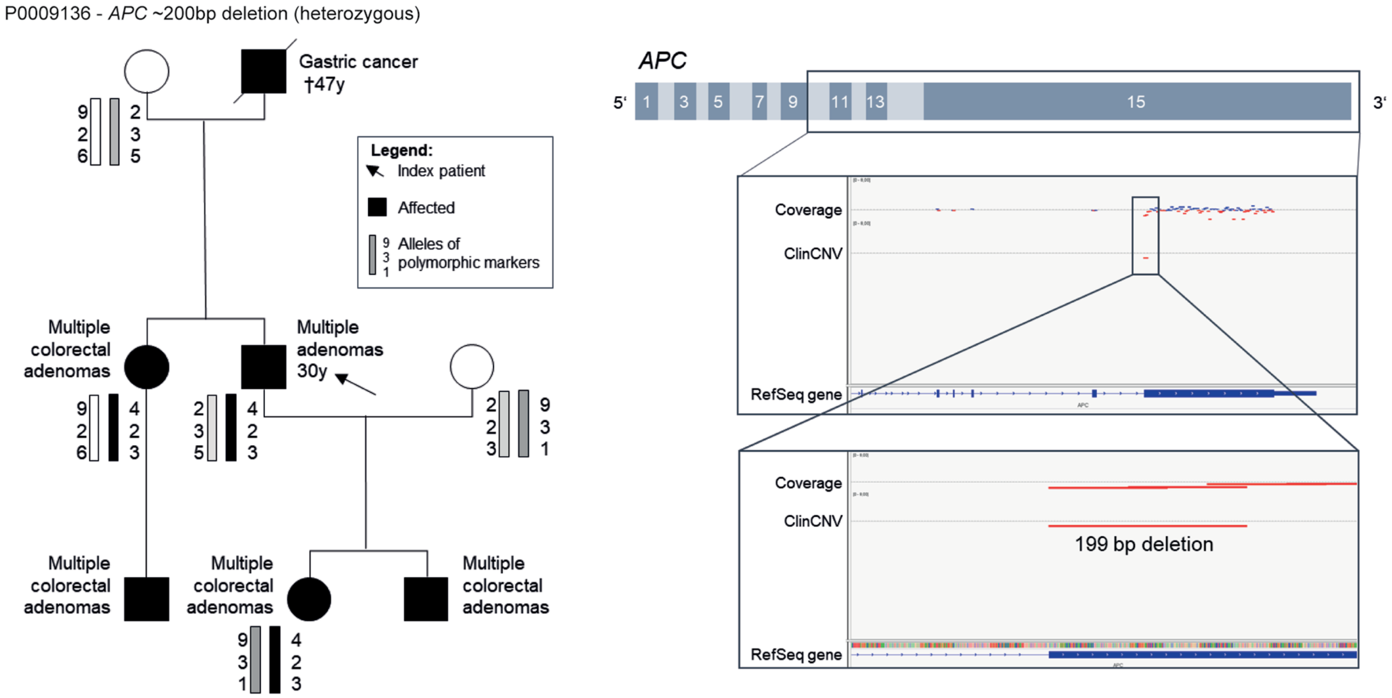
Extended Data Fig. 2 | Flowgram of all analyses performed within the Solve-RD systematic reanalysis. [a]See Supplementary Table 2 for ERN specific gene lists; [b]De novo analysis was performed genome-wide, irrespective of previously identified disease genes; [c]SNV/InDels were investigated within the mitochondrial DNA; [d] Small exceptions in the prioritisation were made between ERNs for certain genes. See Online Methods, and Supplementary Tables 15–18 for further details.



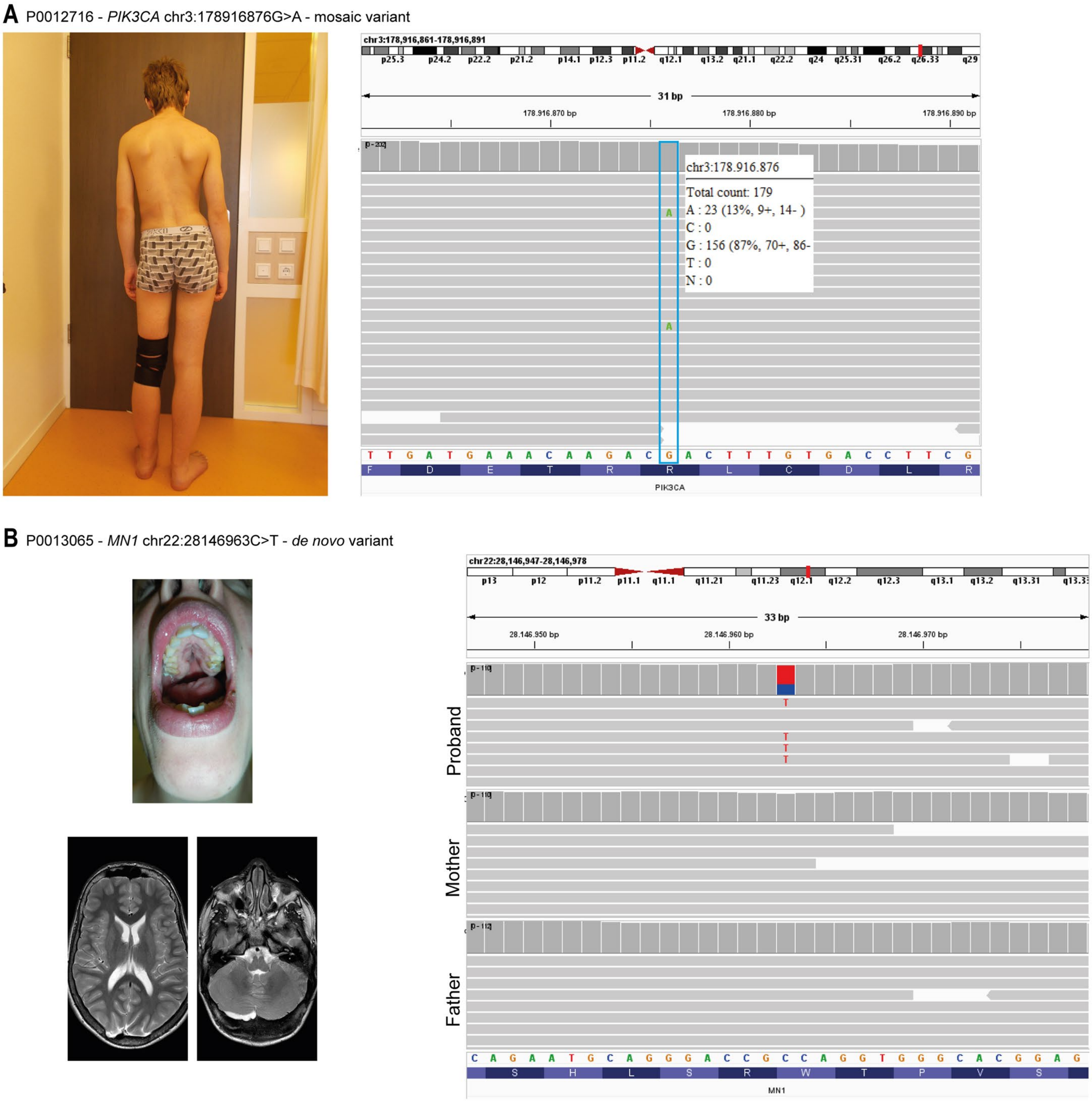
Extended Data Fig. 3 | Date of initial creation, and of last update of OMIM records for genes shown to be disease-causing in this study. This figure shows (**a**) the date of creation of the first OMIM entry for a particular gene determined to be explanatory for the condition in a Solve-RD proband-phenotype association, and (**b**) the date of the last update of the relevant entry. The OMIM entry for 67 genes was only created after 01/01/2018, when Solve-RD started, and many genes of interest have had their records updated since then. This explains why a number of these genes were only confirmed as being disease-causing in affected individuals here as a result of reanalysis in Solve-RD.


P0015028 - *B4GALNT1* chr12:58024802C>T - missense variant & *B4GALNT1* intragenic heterozygous deletions exons 6-11
58y
progressive gait disturbances,
lower limb spasticity,
ataxia,
urinary urgency,
peripheral neruopathy
chr12:58,024,782-58,024,821
p13.32 p13.2 p12.3 p12.1 p11.21 q12 q13.11 q13.2 q14.2 q15 q21.2 q21.32 q22 q23.2 q24.11 q24.22 q24.32
41 bp
58,024,790 bp
58,024,800 bp
58,024,810 bp
58,024,820 bp
G A G G G G C T G A A C T T C C A C A C C C T G T A G G G G T A C T G G A G C
L P Q V E V G Q L P Y Q L
B4GALNT1
chr12:58,005,508-58,027,565
22 kb
58,008 kb
58,010 kb
58,012 kb
58,014 kb
58,016 kb
58,018 kb
58,020 kb
58,022 kb
58,024 kb
58,026 kb
ARHGEF25
NR_120450.1
SLC26A10
B4GALNT1
B4GALNT1

Extended Data Fig. 4 | Example of an individual diagnosed with a rare disease from ERN RND. The left panel shows the pedigree of a 58-year old individual first diagnosed at 42 years of age with progressive gait disturbance and urinary urgency, in the absence of family history of these symptoms (P0015028). The right panel shows two IGV screenshots indicating a heterozygous missense SNV (c.451G>A (p.(Gly151Ser)) in *B4GALNT1* (top) and a heterozygous, approximately 10kb in length, deletion on the other allele (bottom), resulting in complete deletion of exons 6–11 (commencing in exon 5, removing exons 6–11 (NM_001478.5), and ending in the 3'UTR (Chr12(GRch37): g.58014705-58024263del). Location of the deletion is indicated by the red line in the top track, supported by the reduced beta-allele frequency of variants in this region as shown in the centre track, and further supported by read pairs spanning the full 10kb (in red) observed in the lower track.

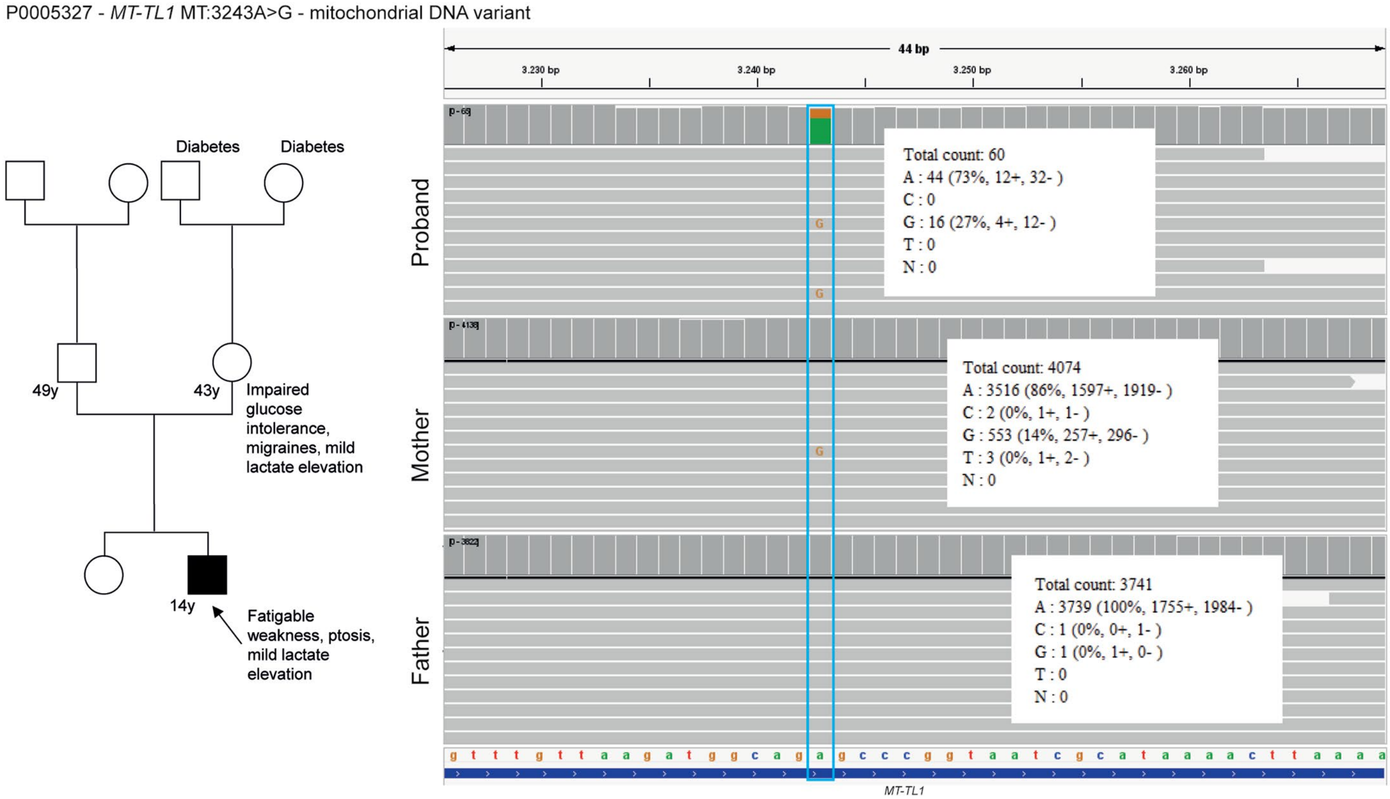


Extended Data Fig. 5 | Example of an individual diagnosed with a rare disease from ERN GENTURIS. Left panel: pedigree of proband P0009136 (indicated by the arrow). Haplotype analysis demonstrated that all affected individuals carry the same allele at the *APC* locus, inherited from the paternal branch of the family. Right panel: comprehensive CNV analysis uncovered a heterozygous germline deletion, approximately 200bp in length, at the beginning of coding exon 15 of the *APC* gene which could not be identified by routine diagnostics using just the sequencing and MLPA methods.


A P0012716 - PIK3CA chr3:178916876G>A - mosaic variant
chr3:178,916,861-178,916,891
p25.3 p24.2 p22.2 p21.2 p14.1 p12.3 p11.2 q12.1 q13.2 q21.1 q22.2 q24 q25.31 q26.2 q26.33 q29
31 bp
178.916.870 bp
178.916.880 bp
178.916.890 bp
[0 - 204]
chr3:178.916.876
Total count: 179
A : 23 (13%, 9+, 14-)
C : 0
G : 156 (87%, 70+, 86-
T : 0
N : 0
A
A
T T G A T G A A A C A A G A C G A C T T T G T G A C C T T C G
F D E T R R L C D L R
PIK3CA
B P0013065 - MN1 chr22:28146963C>T - de novo variant
chr22:28,146,947-28,146,978
p13 p12 p11.2 p11.1 q11.1 q11.21 q11.23 q12.1 q12.2 q12.3 q13.1 q13.2 q13.31 q13.33
33 bp
28.146.950 bp
28.146.960 bp
28.146.970 bp
[0 - 110]
T
T
T
T
Proband
[0 - 110]
Mother
[0 - 112]
Father
C A G A A T G C A G G G A C C G C C A G G T G G G C A C G A G
S H L S R W T P V S
MN1

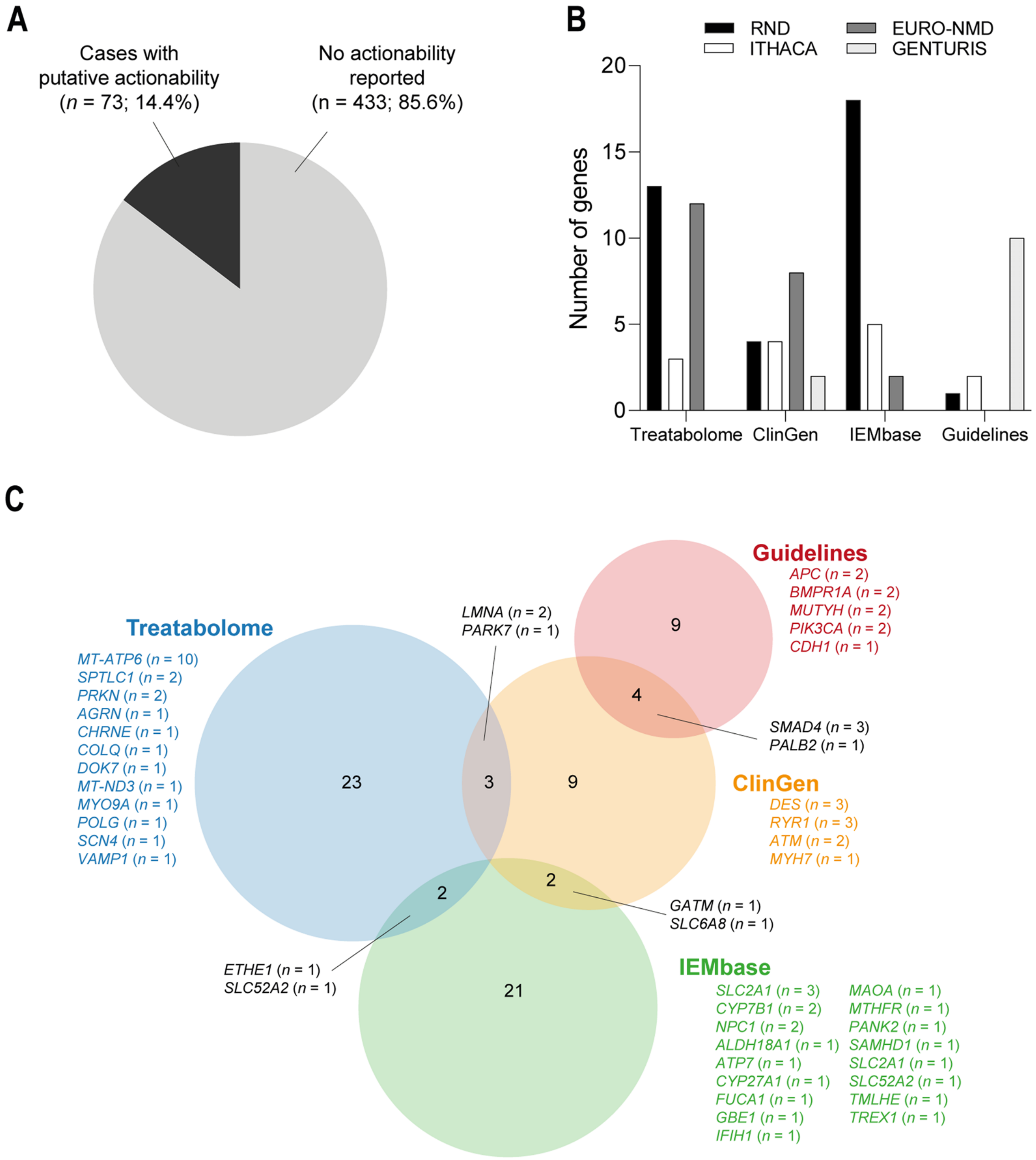
Extended Data Fig. 6 | Examples of two individuals diagnosed with a rare disease from ERN ITHACA. a) The left panel shows the phenotypic presentation of a 24-year old male diagnosed at fifteen years of age with asymmetry of legs and face, described at that time as underdevelopment of the left side (P0012716, written consent that allows sharing of photographs was given). At birth, asymmetry of the legs and face was evident and there was a postaxial rudimentary digit on the right hand that regressed to a small nodule over time. The asymmetry of the face and legs was reported to be stable over time and his cognitive development was within the normal range (IQ of 89). He was affected by complex partial seizures with continuous spike-and-wave during sleep from childhood, however the seizures had a good clinical progression and medication could be discontinued at eleven years of age. Other medical problems included scoliosis, autism spectrum disorder, clumsy motor skills, and sleeping problems. The IGV screenshot in the right panel confirms the presence of a rare de novo mosaic missense variant (observed in only 13% of reads) in *PIK3CA* (chr3:178916876G>A), validated by Sanger sequencing. This variant had previously been reported elsewhere to cause *PIK3CA*-related overgrowth, leading to a change in the clinical diagnosis for this young man, and the resolution of his diagnostic odyssey. **b**) The left panel shows the phenotypic presentation of an undiagnosed 22-year old male who had experienced severe developmental delay, and presented with a variety of physical anomalies, including an open mouth with full lip vermillion, a high and narrow palate with gum hypertrophy and irregular dentition. A brain MRI was initially reported to be uninformative (P0013065, written consent that allows sharing of photographs was given)., The IGV screenshot in the right panel indicates the presence of a rare de novo nonsense variant in *MN1* (Chr22(GRCh37):g.28146963C>T; NM_002430.2:c.3903G>A; p.(Trp1301*)) unobserved in the parents. Retrospective reanalysis of the brain MRI revealed dysplasia of the cerebellar vermis, rhombencephalosynapsis and mild bitemporal narrowing of the skull, consistent with a diagnosis of CEBALID syndrome. The individuals described gave permission for their photos to be used in this publication, for which we thank them and their families.

https://doi.org/10.1038/s41591-024-03420-w

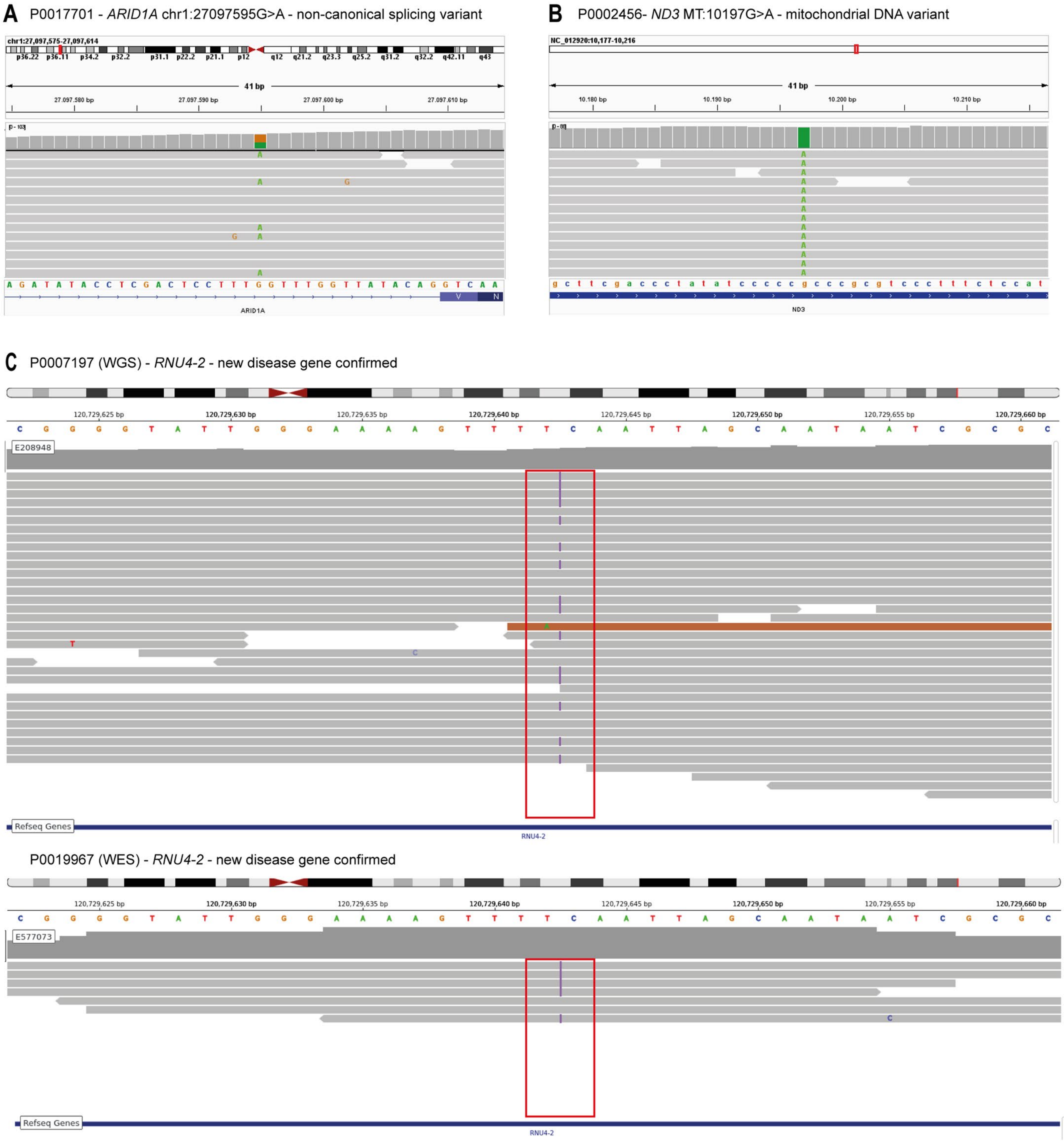

P0005327 - MT-TL1 MT:3243A>G - mitochondrial DNA variant
Diabetes
Diabetes
49y
43y
Impaired glucose intolerance, migraines, mild lactate elevation
14y
Fatigable weakness, ptosis, mild lactate elevation
44 bp
3.230 bp
3.240 bp
3.250 bp
3.260 bp
Proband
[0 - 63]
Total count: 60
A : 44 (73%, 12+, 32-)
C : 0
G : 16 (27%, 4+, 12-)
T : 0
N : 0
Mother
[0 - 4138]
Total count: 4074
A : 3516 (86%, 1597+, 1919-)
C : 2 (0%, 1+, 1-)
G : 553 (14%, 257+, 296-)
T : 3 (0%, 1+, 2-)
N : 0
Father
[0 - 3822]
Total count: 3741
A : 3739 (100%, 1755+, 1984-)
C : 1 (0%, 0+, 1-)
G : 1 (0%, 1+, 0-)
T : 0
N : 0
g t t t g t t a a g a t g g c a g a g c c c g g t a a t c g c a t a a a a c t t a a a a
MT-TL1

Extended Data Fig. 7 | Example of an individual diagnosed with a rare disease from ERN EURO-NMD. The left panel shows the pedigree, and clinical history of proband P0005327 (indicated by the arrow). At eight years of age he began to develop progressive lower limb weakness and fatigability. He started to experience recurrent falls at eight years of age and went on to develop progressive proximal lower limb weakness with prominent fatigability, and a waddling gait. There was no history of bulbar or ocular symptoms. On examination, bilateral asymmetric ptosis with fatigability was observed, as was polyminimyoclonus. Muscle strength was normal in all four limbs, but fatigue occurred upon sustained arm abduction. Deep tendon reflexes were normal, as were serum creatine kinase levels, while repetitive nerve stimulation was inconclusive. Due to a clinical suspicion of Congenital Myasthenic Syndrome (CMS), a trial of pyridostigmine was initiated, but the individual was non-compliant. However, his parents reported spontaneous improvement in baseline limb weakness and falls over the following six years with only episodic worsening due to fever and exertional myalgias. There was a strong family history of diabetes on the maternal side and the mother's fasting glucose levels were suggestive of borderline diabetes, and she also has a long history of migraines. Retrospective serum lactate testing in both proband and mother showed mildly elevated levels (>20 mg/dl). The IGV screenshot in the right panel indicates the presence of a heteroplasmic mitochondrial variant (*MT-TL1*, MT:3243A>G)) observed with a frequency of 27% in the proband, and 14% in his mother. This difference in heteroplasmy may explain the divergence in symptoms between mother and child. While the initial clinical suspicion in the proband was CMS due to the notable fatigability, the fact that mitochondrial disease can be clinically highly variable means that mild forms of mitochondrial myopathy can be difficult to diagnose clinically.

https://doi.org/10.1038/s41591-024-03420-w



Extended Data Fig. 8 | Clinical actionability. **a**) Percentage of solved cases for which the causative gene is reported in one of the three gene-treatment databases included in this study (ClinGen, IEMbase and Treatabolome) and guidelines for surveillance of genetic tumour risk syndromes. **b**) Gene-treatment databases and surveillance guidelines for genes in which (likely) disease-causing variants have been identified per ERN. **c**) List of genes with (likely) disease-causing variants, and number of rare disease probands/families diagnosed in this study in parentheses, identified in each of the three gene-treatment databases as well as surveillance guidelines included in this study.


A P0017701 - ARID1A chr1:27097595G>A - non-canonical splicing variant
chr1:27,097,575-27,097,614
41 bp
27,097,580 bp
27,097,590 bp
27,097,600 bp
27,097,610 bp
A G A T A T A C C T C G A C T C C T T T G G T T T G G T T A T A C A G G T C A A
ARID1A
B P0002456- ND3 MT:10197G>A - mitochondrial DNA variant
NC_012920:10,177-10,216
41 bp
10,180 bp
10,190 bp
10,200 bp
10,210 bp
g c t t c g a c c c t a t a t c c c c c g c c c g c g t c c c t t t c t c c a t
ND3
C P0007197 (WGS) - RNU4-2 - new disease gene confirmed
120,729,625 bp
120,729,630 bp
120,729,635 bp
120,729,640 bp
120,729,645 bp
120,729,650 bp
120,729,655 bp
120,729,660 bp
C G G G G T A T T G G G A A A A G T T T T C A A T T A G C A A T A A T C G C G C
E208948
Refseq Genes
RNU4-2
P0019967 (WES) - RNU4-2 - new disease gene confirmed
E577073
Refseq Genes
RNU4-2

Extended Data Fig. 9 | Examples of 'beyond standard' variant types and discovery by Solve-RD. Panels A&B provide illustrative examples of previously unsolved rare disease probands for which a new variant other than standard coding SNV/InDel resulted in a new diagnosis. **a**) Non-canonical splicing variant in *ARID1A* (individual P0017701); **b**) mtDNA variant in *ND3-MT* (P0002456). **c**) The new discovery of recurrent de novo variants in *RNU4-2* led to likely new diagnoses in two Solve-RD cases. Both variants have been validated, and the phenotypes match the recently published phenotypic descriptions[42,43].

nature portfolio

Corresponding author(s): Alexander Hoischen, Sergi Beltran

Last updated by author(s): Oct 27, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

□ ☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement

☒ □ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly

□ ☒ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.

☒ □ A description of all covariates tested

☒ □ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons

□ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)

☒ □ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted
Give P values as exact values whenever suitable.

☒ □ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings

☒ □ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes

☒ □ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection

Phenotypic data, metadata, and genomic variant data was collated in the RD-Connect Genome-Phenome Analysis Platform (https://platform.rd-connect.eu/#/).

Data analysis

All analysis was undertaken using tools and resources published in peer-reviewed articles, as indicated below.
Tool Version Publication Repository
BWA-MEMv0.7.8 https://arxiv.org/abs/1303.3997 https://github.com/lh3/bwa
GATK HaplotypeCaller v3.6.0 De Pristo et al, 2011 https://github.com/broadinstitute/gatk/releases
DeNovoCNN v1.0 Khazeeva et al, 2022 https://github.com/Genome-Bioinformatics-RadboudUMC/DeNovoCNN
SQUIRLS v1.0 Danis et al, 2021 https://github.com/TheJacksonLaboratory/Squirls
SpliceAI v1.3 Jaganathan et al, 2019 https://github.com/Illumina/SpliceAI
MToolBox v1.2.1 Calabrese et al, 2014 https://github.com/mitoNGS/MToolBox
ClinCNV v1.17.0 Demidov et al, 2022 https://github.com/imgag/ClinCNV
Conifer v0.2.2 Krumm et al, 2012 https://conifer.sourceforge.net/
ExomeDepth v1.1.12 Plagnol et al, 2012 https://github.com/vplagnol/ExomeDepth
Manta v1.6.0 Chen et al, 2016 https://github.com/Illumina/manta
ExpansionHunter v3.1.2 Dolzhenko et al, 2019 https://github.com/Illumina/ExpansionHunter
MELT v2.2.2 Gardner et al, 2017 https://melt.igs.umaryland.edu/
SCRAMble v1.0.2 Torene et al, 2020 https://github.com/GeneDx/scramble
AnnotSV v2.2 Geoffroy et al, 2018 https://lbgi.fr/AnnotSV/

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Access to pseudonymised phenotypic information for all individuals and their genetic variants is possible through the RD-Connect GPAP (https://platform.rd-connect.eu/) upon completion of registration and approval by the independent RD-Connect Data Access Committee (Code of Conduct and registration details can be found at https://platform.rd-connect.eu/userregistration/).
All raw and processed data files (FASTQs, BAM/CRAMs, gVCFs, PED files, Phenopackets) are available at the EGA (Datasets EGAD00001009767, EGAD00001009768, EGAD00001009769, and EGAD00001009770, under Solve-RD study EGAS00001003851), following approval from the Solve-RD Data Access Committee.
Confirmed causative variants have been submitted to ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/) under the following accession numbers: SCV005091231 -- SCV005091564; SCV005199960 -- SCV005200075 ; SCV005200692 -- SCV005200738.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	As rare-diseases affect both sexes more or less equally with the exception of a few conditions that are linked to sex-chromosomes, we did not differentiate between males and females in this work. Gender is not relevant for genetic rare diseases. Breakdown on genetically determined sex is provided in the text: Of 6,447 affected individuals, 3,592 (56%) were male, and 2,855 (44%) female.
Reporting on race, ethnicity, or other socially relevant groupings	We do not report on race/ethnicity/ancestry or any other socially constructed categorization as they are not relevant within the field of rare disease research. We did not exclude anyone based on sex, gender, ethnicity, race, age or any other socially relevant groupings.
Population characteristics	A cohort of rare disease cases covering a range of pathologies falling under the umbrella of 4 European Reference Networks (ERNs) for rare disease (ERN-GENTURIS, ERN-EURO-NMD, ERN-ITHACA, ERN-RND). We did not assess ethnicity/ancestry as a variate, but as patients were recruited across Europe, the majority would be of European ancestry.
Recruitment	All individuals were recruited via the four European Reference Networks. Any undiagnosed individual with an apparent genetic rare disease that falls under the umbrella of conditions in which one of the four partner ERNs specialise, and for whom a prior ES analysis had been undertaken and proven inconclusive, was a candidate for inclusion in this study. We did not exclude anyone based on sex, gender, ethnicity, race, age or any other socially relevant groupings. Informed consent for data sharing, including indirect identifiers within Europe for the purpose of research was obtained from all recruited individuals.
Ethics oversight	The ethics committee/IRB of University of Tübingen gave ethical approval for this work (Clinical Trials.gov Nr.: NCT03491280 (https://clinicaltrials.gov/study/NCT03491280)).

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No formal sample size calculation was performed, or deemed necessary. The 6,004 families analyzed here, represent those which were recruited and which passed quality control requirements,during the first two years of the Solve-RD recruitment process.
Data exclusions	Exclusion criteria were limited to removing families where the data submitted for the proband (genomic or phenotypic) was not of sufficient quality for further analysis to be underaken successfully e.g. a lack of a clear phenotypic description, or where a required median sequencing depth of coverage of at least ten reads over at least 70% of the defined target region of interest for the corresponding exome enrichment kit, or across the entire genome in the case of whole genome sequencing data, was not met. No other exclusion criteria were applied at any point.
Replication	Replication of results was not applicable per se, since each family was analyzed as an individual unit. However, follow-up and validation of identified variants by orthologous means and using other bioinformatic tools was undertaken in the majority of cases to assure that the

variants identified were biologically real and relevant. To a large extent and as common practice in the rare genetic disease field, replication of previously variants has happened or will happen via databases (e.g. ClinVar) or scientific literature.

Randomization

Recruitment was performed concurrently by the four European Reference Networks, thus randomising the order in which data was processed.

Blinding

This is an observational study, not a clinical trial, and as such blinding was unnecessary.

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description

Briefly describe the study type including whether data are quantitative, qualitative, or mixed-methods (e.g. qualitative cross-sectional, quantitative experimental, mixed-methods case study).

Research sample

State the research sample (e.g. Harvard university undergraduates, villagers in rural India) and provide relevant demographic information (e.g. age, sex) and indicate whether the sample is representative. Provide a rationale for the study sample chosen. For studies involving existing datasets, please describe the dataset and source.

Sampling strategy

Describe the sampling procedure (e.g. random, snowball, stratified, convenience). Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient. For qualitative data, please indicate whether data saturation was considered, and what criteria were used to decide that no further sampling was needed.

Data collection

Provide details about the data collection procedure, including the instruments or devices used to record the data (e.g. pen and paper, computer, eye tracker, video or audio equipment) whether anyone was present besides the participant(s) and the researcher, and whether the researcher was blind to experimental condition and/or the study hypothesis during data collection.

Timing

Indicate the start and stop dates of data collection. If there is a gap between collection periods, state the dates for each sample cohort.

Data exclusions

If no data were excluded from the analyses, state so OR if data were excluded, provide the exact number of exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.

Non-participation

State how many participants dropped out/declined participation and the reason(s) given OR provide response rate OR state that no participants dropped out/declined participation.

Randomization

If participants were not allocated into experimental groups, state so OR describe how participants were allocated to groups, and if allocation was not random, describe how covariates were controlled.

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description

Briefly describe the study. For quantitative data include treatment factors and interactions, design structure (e.g. factorial, nested, hierarchical), nature and number of experimental units and replicates.

Research sample

Describe the research sample (e.g. a group of tagged Passer domesticus, all Stenocereus thurberi within Organ Pipe Cactus National Monument), and provide a rationale for the sample choice. When relevant, describe the organism taxa, source, sex, age range and any manipulations. State what population the sample is meant to represent when applicable. For studies involving existing datasets, describe the data and its source.

Sampling strategy

Note the sampling procedure. Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient.

Data collection

Describe the data collection procedure, including who recorded the data and how.

Timing and spatial scale

Indicate the start and stop dates of data collection, noting the frequency and periodicity of sampling and providing a rationale for these choices. If there is a gap between collection periods, state the dates for each sample cohort. Specify the spatial scale from which the data are taken

Data exclusions

If no data were excluded from the analyses, state so OR if data were excluded, describe the exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.

Reproducibility

Describe the measures taken to verify the reproducibility of experimental findings. For each experiment, note whether any attempts to repeat the experiment failed OR state that all attempts to repeat the experiment were successful.

Randomization

Describe how samples/organisms/participants were allocated into groups. If allocation was not random, describe how covariates were controlled. If this is not relevant to your study, explain why.

Blinding

Describe the extent of blinding used during data acquisition and analysis. If blinding was not possible, describe why OR explain why blinding was not relevant to your study.

Did the study involve field work? □ Yes □ No

Field work, collection and transport

Field conditions

Describe the study conditions for field work, providing relevant parameters (e.g. temperature, rainfall).

Location

State the location of the sampling or experiment, providing relevant parameters (e.g. latitude and longitude, elevation, water depth).

Access & import/export

Describe the efforts you have made to access habitats and to collect and import/export your samples in a responsible manner and in compliance with local, national and international laws, noting any permits that were obtained (give the name of the issuing authority, the date of issue, and any identifying information).

Disturbance

Describe any disturbance caused by the study and how it was minimized.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study	
☒	□	Antibodies
☒	□	Eukaryotic cell lines
☒	□	Palaeontology and archaeology
☒	□	Animals and other organisms
□	☒	Clinical data
☒	□	Dual use research of concern
☒	□	Plants

Methods

n/a	Involved in the study	
☒	□	ChIP-seq
☒	□	Flow cytometry
☒	□	MRI-based neuroimaging

Antibodies

Antibodies used

Describe all antibodies used in the study; as applicable, provide supplier name, catalog number, clone name, and lot number.

Validation

Describe the validation of each primary antibody for the species and application, noting any validation statements on the manufacturer's website, relevant citations, antibody profiles in online databases, or data provided in the manuscript.

Eukaryotic cell lines

Policy information about cell lines and Sex and Gender in Research

Cell line source(s)

State the source of each cell line used and the sex of all primary cell lines and cells derived from human participants or vertebrate models.

Authentication

Describe the authentication procedures for each cell line used OR declare that none of the cell lines used were authenticated.

Mycoplasma contamination

Confirm that all cell lines tested negative for mycoplasma contamination OR describe the results of the testing for mycoplasma contamination OR declare that the cell lines were not tested for mycoplasma contamination.

Commonly misidentified lines
(See ICLAC register)

Name any commonly misidentified cell lines used in the study and provide a rationale for their use.

Palaeontology and Archaeology

Specimen provenance

Provide provenance information for specimens and describe permits that were obtained for the work (including the name of the issuing authority, the date of issue, and any identifying information). Permits should encompass collection and, where applicable, export.

Specimen deposition

Indicate where the specimens have been deposited to permit free access by other researchers.

Dating methods

If new dates are provided, describe how they were obtained (e.g. collection, storage, sample pretreatment and measurement), where they were obtained (i.e. lab name), the calibration program and the protocol for quality assurance OR state that no new dates are provided.

☐ Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

Ethics oversight

Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Animals and other research organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research, and Sex and Gender in Research

Laboratory animals

For laboratory animals, report species, strain and age OR state that the study did not involve laboratory animals.

Wild animals

Provide details on animals observed in or captured in the field; report species and age where possible. Describe how animals were caught and transported and what happened to captive animals after the study (if killed, explain why and describe method; if released, say where and when) OR state that the study did not involve wild animals.

Reporting on sex

Indicate if findings apply to only one sex; describe whether sex was considered in study design, methods used for assigning sex. Provide data disaggregated for sex where this information has been collected in the source data as appropriate; provide overall numbers in this Reporting Summary. Please state if this information has not been collected. Report sex-based analyses where performed, justify reasons for lack of sex-based analysis.

Field-collected samples

For laboratory work with field-collected samples, describe all relevant parameters such as housing, maintenance, temperature, photoperiod and end-of-experiment protocol OR state that the study did not involve samples collected from the field.

Ethics oversight

Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration

Not Applicable

Study protocol

Not Applicable

Data collection

Not Applicable

Outcomes

Not Applicable

Dual use research of concern

Policy information about dual use research of concern

Hazards

Could the accidental, deliberate or reckless misuse of agents or technologies generated in the work, or the application of information presented in the manuscript, pose a threat to:

No	Yes	
☒	☐	Public health
☒	☐	National security
☒	☐	Crops and/or livestock
☒	☐	Ecosystems
☒	☐	Any other significant area

Experiments of concern

Does the work involve any of these experiments of concern:

No	Yes	
☒	☐	Demonstrate how to render a vaccine ineffective
☒	☐	Confer resistance to therapeutically useful antibiotics or antiviral agents
☒	☐	Enhance the virulence of a pathogen or render a nonpathogen virulent
☒	☐	Increase transmissibility of a pathogen
☒	☐	Alter the host range of a pathogen
☒	☐	Enable evasion of diagnostic/detection modalities
☒	☐	Enable the weaponization of a biological agent or toxin
☒	☐	Any other potentially harmful combination of experiments and agents

Plants

Seed stocks

Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.

Novel plant genotypes

Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.

Authentication

Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.

ChIP-seq

Data deposition

☐ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☐ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links
May remain private before publication.

For "Initial submission" or "Revised version" documents, provide reviewer access links. For your "Final submission" document, provide a link to the deposited data.

Files in database submission

Provide a list of all files available in the database submission.

Genome browser session
(e.g. UCSC)

Provide a link to an anonymized genome browser session for "Initial submission" and "Revised version" documents only, to enable peer review. Write "no longer applicable" for "Final submission" documents.

Methodology

Replicates

Describe the experimental replicates, specifying number, type and replicate agreement.

Sequencing depth

Describe the sequencing depth for each experiment, providing the total number of reads, uniquely mapped reads, length of reads and whether they were paired- or single-end.

Antibodies

Describe the antibodies used for the ChIP-seq experiments; as applicable, provide supplier name, catalog number, clone name, and lot number.

Peak calling parameters

Specify the command line program and parameters used for read mapping and peak calling, including the ChIP, control and index files used.

Data quality

Describe the methods used to ensure data quality in full detail, including how many peaks are at FDR 5% and above 5-fold enrichment.

Software

Describe the software used to collect and analyze the ChIP-seq data. For custom code that has been deposited into a community repository, provide accession details.

Flow Cytometry

Plots

Confirm that:

☐ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☐ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☐ All plots are contour plots with outliers or pseudocolor plots.

☐ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation: *Describe the sample preparation, detailing the biological source of the cells and any tissue processing steps used.*

Instrument: *Identify the instrument used for data collection, specifying make and model number.*

Software: *Describe the software used to collect and analyze the flow cytometry data. For custom code that has been deposited into a community repository, provide accession details.*

Cell population abundance: *Describe the abundance of the relevant cell populations within post-sort fractions, providing details on the purity of the samples and how it was determined.*

Gating strategy: *Describe the gating strategy used for all relevant experiments, specifying the preliminary FSC/SSC gates of the starting cell population, indicating where boundaries between "positive" and "negative" staining cell populations are defined.*

☐ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Magnetic resonance imaging

Experimental design

Design type: *Indicate task or resting state; event-related or block design.*

Design specifications: *Specify the number of blocks, trials or experimental units per session and/or subject, and specify the length of each trial or block (if trials are blocked) and interval between trials.*

Behavioral performance measures: *State number and/or type of variables recorded (e.g. correct button press, response time) and what statistics were used to establish that the subjects were performing the task as expected (e.g. mean, range, and/or standard deviation across subjects).*

Acquisition

Imaging type(s): *Specify: functional, structural, diffusion, perfusion.*

Field strength: *Specify in Tesla*

Sequence & imaging parameters: *Specify the pulse sequence type (gradient echo, spin echo, etc.), imaging type (EPI, spiral, etc.), field of view, matrix size, slice thickness, orientation and TE/TR/flip angle.*

Area of acquisition: *State whether a whole brain scan was used OR define the area of acquisition, describing how the region was determined.*

Diffusion MRI ☐ Used ☐ Not used

Preprocessing

Preprocessing software: *Provide detail on software version and revision number and on specific parameters (model/functions, brain extraction, segmentation, smoothing kernel size, etc.).*

Normalization: *If data were normalized/standardized, describe the approach(es): specify linear or non-linear and define image types used for transformation OR indicate that data were not normalized and explain rationale for lack of normalization.*

Normalization template: *Describe the template used for normalization/transformation, specifying subject space or group standardized space (e.g. original Talairach, MNI305, ICBM152) OR indicate that the data were not normalized.*

Noise and artifact removal: *Describe your procedure(s) for artifact and structured noise removal, specifying motion parameters, tissue signals and physiological signals (heart rate, respiration).*

Volume censoring	*Define your software and/or method and criteria for volume censoring, and state the extent of such censoring.*

Statistical modeling & inference

Model type and settings	*Specify type (mass univariate, multivariate, RSA, predictive, etc.) and describe essential details of the model at the first and second levels (e.g. fixed, random or mixed effects; drift or auto-correlation).*
Effect(s) tested	*Define precise effect in terms of the task or stimulus conditions instead of psychological concepts and indicate whether ANOVA or factorial designs were used.*

Specify type of analysis: □ Whole brain □ ROI-based □ Both

Statistic type for inference (See Eklund et al. 2016)	*Specify voxel-wise or cluster-wise and report all relevant parameters for cluster-wise methods.*
Correction	*Describe the type of correction and how it is obtained for multiple comparisons (e.g. FWE, FDR, permutation or Monte Carlo).*

Models & analysis

n/a	Involved in the study
□	□ Functional and/or effective connectivity
□	□ Graph analysis
□	□ Multivariate modeling or predictive analysis

Functional and/or effective connectivity	*Report the measures of dependence used and the model details (e.g. Pearson correlation, partial correlation, mutual information).*
Graph analysis	*Report the dependent variable and connectivity measure, specifying weighted graph or binarized graph, subject- or group-level, and the global and/or node summaries used (e.g. clustering coefficient, efficiency, etc.).*
Multivariate modeling and predictive analysis	*Specify independent variables, features extraction and dimension reduction, model, training and evaluation metrics.*