HUMAN GENETICS

Targeted long-read RNA sequencing for rare disease diagnosis and variant interpretation

Robert Wang[1,2]†, Feng Wang[1]†, Nicole DeBruyne[1,3], Xinjun Ji[1,4], Nicole M. Engelhardt[5], Joseph Jee-Hwan Park[1], Amber Notaro[1], Samantha Gaerlan[1], Ryan Park[1], Matthew J. Schultz[6], Sheila Clever[1,5], Elizabeth M. McCormick[7], Kelsey Keith[8], Bobby G. Ng[9], Kathryn E. Kadash-Edmondson[1], Hudson H. Freeze[9], Christina T. Lam[10,11], Eva Morava[12], Ingo Helbig[8,13,14,15], Marni J. Falk[7,16]*, Rebecca D. Ganetzky[1,5,16]*, Andrew C. Edmondson[1,5,16]*, Lan Lin[1,17,18]*, Yi Xing[1,8,17]*

Diagnosing rare genetic diseases remains a major challenge despite widespread clinical testing. Long-read RNA sequencing (RNA-seq) offers a powerful approach to capturing the effects of genetic variants on the transcriptome, yet challenges with sequencing coverage, cost, tissue selection, and scalability have limited its clinical adoption. To address this, we developed STRIPE (Sequencing Targeted RNAs Identifies Pathogenic Events), a targeted long-read RNA-seq–based strategy for rare disease diagnosis and variant interpretation. STRIPE enables deep sequencing of full-length transcripts for any customized disease-specific gene panel such that a wide range of clinically informative readouts, including transcript aberrations and sequence variants, can be detected at haplotype-level resolution. Applying STRIPE to 88 individuals spanning two major rare disease groups, we accurately reidentified known pathogenic variants and revealed their transcript consequences, including many unexpected ones. For 8 of 15 splice site region variants, we observed more complex RNA processing defects beyond single exon skipping or cryptic splice site activation. Notably, we find that donor splice site variants frequently activate cryptic intronic polyadenylation sites, leading to premature transcript termination. Leveraging unique strengths of long-read RNA-seq, STRIPE also resolved variants of uncertain significance and uncovered disease-causing variants in five previously undiagnosed individuals. Overall, STRIPE is a powerful, adaptable, and scalable strategy with broad potential to improve clinical variant interpretation and advance genetic diagnosis of rare diseases.

INTRODUCTION

Exome and genome sequencing are routinely used in clinical and research settings to identify the molecular basis of rare genetic conditions. However, the diagnostic yield of these technologies varies from 20 to 50% (*1*–*4*), highlighting a large number of patients for whom current genomic testing approaches cannot provide a molecular diagnosis. While some of these cases may be explained by polygenic or environmental risk factors, many undiagnosed cases are likely due to genetic changes that cannot be identified or prioritized by current clinical workflows. Over the past decade, RNA sequencing (RNA-seq) has steadily emerged as a complementary assay for rare disease diagnosis, helping clarify variants of uncertain significance (VUS) and uncover disease-causing variants in individuals lacking a molecular diagnosis after standard genetic testing (*5*–*12*). RNA-seq not only captures sequence variation in exonic regions but also reveals changes in gene expression and splicing induced by regulatory genetic variants.

Long-read sequencing technologies, which can generate reads tens to thousands of kilobases in length, have increasingly enhanced our understanding of genetic and transcriptomic variation and their relationship to human traits and diseases (*13*–*15*). In RNA analysis, long reads permit end-to-end sequencing of full-length transcripts and provide phasing information that can link individual transcripts to their haplotype of origin (*16*, *17*). Analysis of long-read RNA-seq data has already proven useful in individual rare disease cases as a follow-up test to observe the functional effects of known disease-causing variants (*18*, *19*). However, despite its potential, long-read RNA-seq is underutilized in rare disease research, in part due to the lower throughput and higher cost of long-read sequencing platforms, as well as the lack of a diagnostic workflow tailored to long-read RNA-seq

[1]Center for Computational and Genomic Medicine, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA. [2]Graduate Group in Genomics and Computational Biology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA. [3]Graduate Group in Cell and Molecular Biology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA. [4]Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA. [5]Division of Human Genetics, Department of Pediatrics, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA. [6]Biochemical Genetics Laboratory, Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN 55902, USA. [7]Mitochondrial Medicine Frontier Program, Division of Human Genetics, Department of Pediatrics, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA. [8]Department of Biomedical and Health Informatics, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA. [9]Human Genetics Program, Sanford Children's Health Research Center, La Jolla, CA 92037, USA. [10]Norcliffe Foundation Center for Integrative Brain Research, Seattle Children's Research Institute, Seattle, WA 98109, USA. [11]Division of Genetic Medicine, Department of Pediatrics, University of Washington School of Medicine, Seattle, WA 98195, USA. [12]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA. [13]Division of Neurology, Department of Pediatrics, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA. [14]The Epilepsy NeuroGenetics Initiative (ENGIN), Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA. [15]Department of Neurology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA. [16]Department of Pediatrics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA. [17]Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA. [18]Raymond G. Perelman Center for Cellular and Molecular Therapeutics, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA.
*Corresponding author. Email: falkm@chop.edu (M.J.F.); ganetzkyr@chop.edu (R.D.G.); edmondsona@chop.edu (A.C.E.); linlan@chop.edu (L.L.); xingyi@chop.edu (Y.X.)
†These authors contributed equally to this work.

data. Importantly, RNA-guided genetic diagnosis often relies on clinically accessible tissues (*20*), in which phenotype-relevant genes may have low transcript levels. As a result, the high sequencing cost required for achieving adequate sequencing coverage creates a major barrier to scaling up long-read RNA-seq in clinical settings.

One strategy to overcome this barrier is to perform targeted long-read RNA-seq, which focuses sequencing capacity on genes of interest. Recently, we developed TEQUILA (Target Enrichment and Quantification Utilizing Isothermally Linear-Amplified probes), a highly versatile and low-cost method for generating large quantities of capture oligos for any gene panel (*21*). When coupled with long-read RNA-seq, TEQUILA-seq provides deep coverage of full-length transcripts for target genes in custom-designed gene panels while preserving quantitative gene expression and splicing patterns. Here, we describe the application of this technology for diagnosing patients with suspected rare genetic diseases using RNA obtained from clinically accessible tissues. We report a diagnostic workflow, STRIPE (Sequencing Targeted RNAs Identifies Pathogenic Events), that can comprehensively detect and phase both transcript aberrations and sequence variants for any disease-specific gene panel. Across two different rare disease groups [congenital disorders of glycosylation (CDG) and primary mitochondrial diseases (PMD)], we demonstrate that STRIPE can not only delineate the diverse and frequently unexpected transcript consequences of previously identified variants, including VUS, but also uncover disease-causing variation in genetically undiagnosed patients.

RESULTS

Overview of STRIPE workflow and study design

We developed a versatile workflow, STRIPE, which enables deep sequencing of full-length transcripts for any disease-specific gene panel such that a wide range of clinically informative readouts, including transcript aberrations and sequence variants, can be detected at haplotype-level resolution (Fig. 1, A and B). First, hybridization capture probes for genes of interest are generated by our TEQUILA protocol (*21*) and used to selectively pull down full-length cDNAs derived from a patient sample. The resulting pull-down libraries are subjected to long-read sequencing on Oxford Nanopore Technologies (ONT) platforms, followed by quality control checks on read mapping rates and target enrichment levels. To identify potential disease-causing genes and variants from patient TEQUILA-seq data, we devised an analysis pipeline that can perform haplotype-resolved detection of rare deleterious sequence variants and outlier transcriptomic events across targeted genes. Briefly, STRIPE constructs haplotypes from heterozygous single-nucleotide variants (SNVs) present in TEQUILA-seq reads for each target gene and uses a realignment strategy to partition the reads across the constructed haplotypes. TEQUILA-seq reads are then analyzed for extreme patterns in gene expression and splicing by comparison to a large cohort of healthy individuals with tissue-matched short-read RNA-seq samples from the Genotype-Tissue Expression (GTEx) Consortium (*22*). When paired with DNA sequencing data, these outlier transcriptomic signals can help identify and prioritize nearby rare variants with unusually large regulatory effects, as demonstrated in previous studies (*23*–*27*). STRIPE also deploys multiple tools to perform variant calling on TEQUILA-seq reads and prioritizes rare deleterious variants in each target gene based on population allele frequencies in gnomAD, existing clinical annotations, and predicted severity.

To investigate the clinical utility of STRIPE, we assembled a cohort of 88 individuals, including 68 patients with rare disease who were clinically suspected to have CDG or PMD based on prior phenotyping and biochemical testing, as well as 20 unaffected healthy controls (Fig. 1C). Of the 68 patients with rare disease in our cohort, 22 had known genetic causes from prior exome or genome sequencing analysis, allowing us to assess the capacity of STRIPE to identify disease-causing variants and characterize their effects on gene expression and RNA processing (*28*–*35*). The remaining 46 genetically undiagnosed patients included cases for whom DNA sequencing had (i) identified at least one VUS in a gene related to a patient's clinical phenotype, (ii) found a single disease-causing variant without a second variant in the gene for a recessive condition, or (iii) failed to reveal any candidate genes (table S1). Given the multisystemic nature of the diseases represented in our cohort (*36*, *37*), we reasoned that clinically accessible tissues could adequately capture genes that are commonly disrupted in these disorders. Using short-read RNA-seq samples from GTEx, we found that nearly two to three times more known disease genes for CDG and PMD could be detected with a median TPM (transcripts per million) > 10 in fibroblasts compared to whole blood (fig. S1). Therefore, to maximize power in detecting disease-relevant genes and transcripts, we performed TEQUILA-seq of skin fibroblasts from our cohort using two different custom-designed gene panels tailored to CDG and PMD (fig. S2A, tables S2 to S4, and Materials and Methods).

For both gene panels, we achieved ultradeep coverage of target genes, with median on-target rates of 64.2% for the CDG gene panel and 87.1% for the PMD gene panel (fig. S2B and table S4). Compared to untargeted whole-transcriptome long-read RNA-seq, TEQUILA-seq resulted in a 71.8- and 39.5-fold increase in the fraction of on-target reads for the CDG and PMD gene panels, respectively (Fig. 1D), yielding a substantial boost in read coverage per target gene. Consistent with our prior work (*21*), we observed a strong correlation in the normalized abundances of target genes (Spearman's $\rho = 0.91$ for CDG and 0.90 for PMD) (fig. S3A) and usage frequencies of splice junctions in target genes (Spearman's $\rho = 0.96$ for CDG and 0.95 for PMD) (Fig. 1E) between untargeted whole-transcriptome long-read RNA-seq and TEQUILA-seq data, indicating that our capture probes faithfully recapitulate quantitative gene expression and splicing patterns in skin fibroblasts. Finally, for both gene panels, principal component analysis of target gene expression (fig. S3B) and splicing patterns (Fig. 1F) revealed that fibroblasts from our cohort closely resembled GTEx fibroblasts when compared to other clinically accessible tissues. While batch effects related to differences in sample preparation and sequencing technologies (i.e., short-read RNA-seq versus long-read RNA-seq) exist (fig. S3, C and D), their overall magnitude was modest (Fig. 1F and fig. S3B). Importantly, target gene splicing patterns showed strong agreement (Spearman's $\rho = 0.94$ for CDG and 0.94 for PMD) between cohort and GTEx fibroblasts (fig. S3E). Moreover, measurements of allelic expression are largely unaffected by technical factors (*38*). On the basis of these observations, we reasoned that GTEx fibroblasts could be used as additional controls to facilitate the detection of transcript aberrations from patient TEQUILA-seq data.

STRIPE enables haplotype-resolved identification and transcript characterization of previously known pathogenic variants

We first sought to assess how well STRIPE could capture and phase known pathogenic variants within exonic regions from TEQUILA-seq data on previously diagnosed patients. STRIPE correctly identified

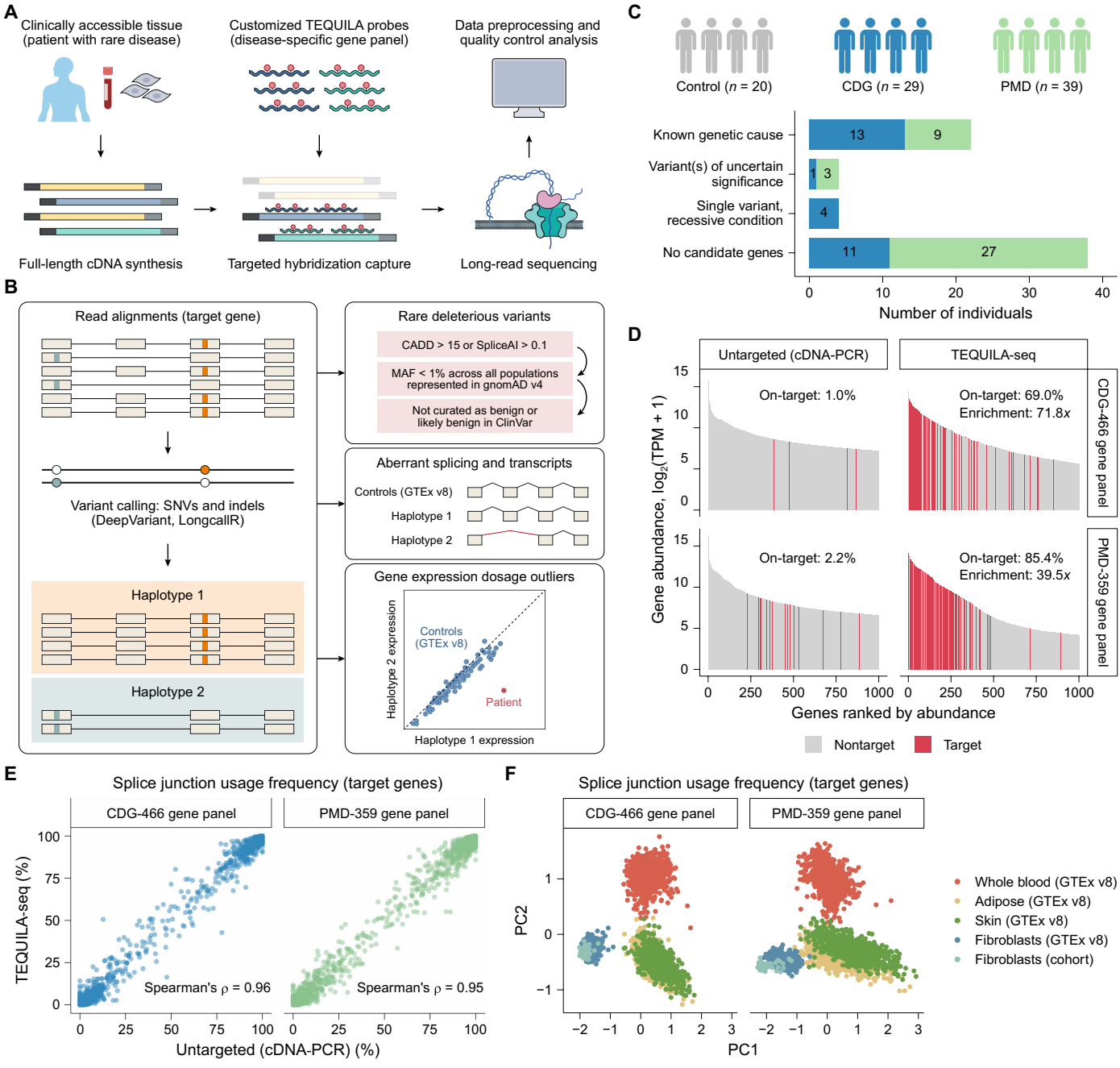

Fig. 1. Overview of STRIPE workflow and quality control. (**A**) Clinically accessible tissue from a patient is collected for RNA extraction and full-length cDNA synthesis. Resulting cDNA is enriched using TEQUILA probes targeting disease-specific genes and prepared for long-read nanopore sequencing. (**B**) For each target gene, heterozygous SNVs in TEQUILA-seq reads are used to construct haplotypes, and a realignment strategy is used to partition reads across the haplotypes (left). The resulting read sets are analyzed for extreme patterns in gene expression dosage and splicing by comparison to tissue-matched short-read RNA-seq samples from the GTEx Consortium, as well as the presence of rare deleterious variants based on predicted severity, population allele frequencies in gnomAD, and existing clinical annotations. (**C**) We sequenced skin fibroblast RNAs from 88 individuals, including 20 unaffected controls and 68 patients with rare disease clinically referred for CDG or PMD. Of the 68 patients, 22 had known genetic causes from prior testing, while the remaining 46 were genetically undiagnosed. (**D**) Gene abundances based on untargeted long-read RNA-seq and TEQUILA-seq data for fibroblast cell lines CDG-P12 (CDG-466 gene panel) and PMD-C01 (PMD-359 gene panel). Each bar represents one gene, and only the 1000 most abundant genes are shown. On-target rates represent the fraction of transcriptional abundance from target genes. Fold enrichment is calculated by dividing the on-target rate in TEQUILA-seq data by the on-target rate in untargeted data. (**E**) Comparison of annotated splice junction usage frequencies in target genes between untargeted and TEQUILA-seq data for CDG-P12 (CDG-466 gene panel) and PMD-C01 (PMD-359 gene panel). (**F**) Principal component analysis (PCA) of annotated splice junction usage frequencies in target genes from the CDG-466 and PMD-359 gene panels in 88 cohort fibroblast samples and GTEx samples for clinically accessible tissues. MAF, minor allele frequency.

all 23 such variants, including 18 missense variants, 3 inframe deletions, 1 nonsense variant, and 1 synonymous variant (fig. S4A and table S5). Furthermore, STRIPE was able to phase 18 of the 23 variants based on sequence information contained within TEQUILA-seq reads alone. For example, in patient CDG-P08, clinical exome sequencing identified two heterozygous missense variants in *PIGN* that were approximately 97.4 kb apart. Using TEQUILA-seq reads spanning both variant positions, STRIPE was able to determine that the two variants were in trans (fig. S4B).

When provided genotyping information from patient DNA sequencing data, STRIPE was able to phase TEQUILA-seq reads for all genes with heterozygous pathogenic variants in previously diagnosed patients, allowing us to study the effects of these variants on gene expression and splicing at haplotype-level resolution. Across 18 patients with heterozygous pathogenic variants (exonic and intronic), we found that gene haplotypes showing unusually lower expression relative to their opposite haplotypes (Materials and Methods) were more likely to harbor putative protein-truncating variants [i.e., structural deletion, nonsense, frameshift, and splice site region (*39*)] (Fisher's exact test; $P = 0.025$). While not all putative protein-truncating variants were associated with lower gene expression (fig. S5 and table S6), those that were likely disrupt transcription initiation or decrease transcript stability (*26*, *39*). For example, in patient PMD-P09 (*35*), we observed little to no expression of *TRMU* from a haplotype carrying a 1.8-kb structural deletion that removes the promoter region and first exon (Fig. 2A). Similarly, in patient CDG-P07, we observed a 77.1% decrease in *PIGN* expression from a haplotype carrying a stop-gain variant (NM_176787.5:c.1759C>T, p.Arg587Ter) that likely triggers nonsense-mediated decay (NMD) (Fig. 2B).

On the other hand, we found that splice site region variants [i.e., variants located in the −3 to +6 region of the donor splice site or the −20 to +3 region of the acceptor splice site (*40*)] tended to have aberrant splice junctions nearby in cis, as expected (fig. S6A and table S7). Notably, we also detected aberrant splice junctions that were proximal to other classes of variants in cis, such as missense and structural variants. For example, gene haplotypes harboring multiexon deletions were found to express aberrant transcripts featuring skipping of the deleted exons (Fig. 2, C to G, and fig. S7). Similarly, in an individual (CDG-P10) with CDG type 1a, we found an unusually high frequency of exon 5 skipping in *PMM2* transcripts expressed from a haplotype carrying a pathogenic missense variant (NM_000303.3:c.415G>A, p.Glu139Lys) that is known to disrupt a splicing enhancer sequence (*41*) in exon 5 (fig. S8). Variants predicted to affect splicing [i.e., SpliceAI (*42*) > 0.1] were significantly more likely to have aberrant splice junctions within a distance of 100 base pairs (bp) on the same haplotype (Wilcoxon rank sum test; $P = 6.4 \times 10^{-5}$), suggesting that these variants may be responsible for the observed splicing defects (fig. S6B). This observation was robust to other commonly used SpliceAI score cutoffs (SpliceAI > 0.2, $P = 1.4 \times 10^{-5}$; SpliceAI > 0.5, $P = 3.7 \times 10^{-5}$). Nonetheless, we also found several aberrant splice junctions that were far away (i.e., > 100 bp) from previously reported variants despite being on the same haplotype. In a patient (PMD-P07) with mitochondrial complex I deficiency, we found an unusually high frequency of exon 10 skipping in *NUBPL* transcripts expressed from a haplotype carrying a rare missense variant on exon 2 (fig. S9, A to E). This missense variant (NM_025152.3:c.166G>A, p.Gly56Arg) is unlikely to explain the observed changes in exon 10 splicing due to its distance and lack of computational evidence supporting a deleterious effect on splicing (SpliceAI = 0). Notably, previous clinical testing had identified a second rare variant (NM_025152.3:c.815-27T>C) upstream of exon 10 that resides on the same haplotype as the p.Gly56Arg missense variant on exon 2, and targeted long-read genomic DNA (gDNA) sequencing of *NUBPL* in patient fibroblasts confirmed this arrangement (fig. S9F). The c.815-27T>C variant is known to increase exon 10 skipping by disrupting a branchpoint motif, and patients with complex I deficiency and the p.Gly56Arg missense variant are almost always found to carry this branchpoint mutation in cis combined with a second deleterious *NUBPL* variant in trans (*43*–*45*). Together, our findings demonstrate how STRIPE enables both local and long-range coupling of disease-relevant alleles with splicing defects in cis.

STRIPE discovers complex transcript consequences of splice site region variants

While standard genetic testing approaches can detect splice site region variants, which are likely to induce mis-splicing, such methods cannot reliably predict whether exon skipping, cryptic splice site activation, or intron retention will occur, greatly limiting interpretations of pathogenicity (*46*). We reasoned that clinical interpretations of splice site region variants could be substantially improved using long-read RNA-seq data, which not only provide phasing information linking disease-associated alleles to mis-splicing events but also reveal such mis-splicing events in the context of full-length mRNA transcripts. To this end, we used STRIPE to further characterize the precise nature of aberrant RNA processing events associated with splice site region variants that were identified from clinical testing of previously diagnosed and undiagnosed patients.

Haplotype-resolved evaluation of TEQUILA-seq reads around 15 splice site region variants revealed a wide range of associated defects in RNA processing (Fig. 3A and table S8). Importantly, two of the 15 splice site region variants were previously reported as VUS, and STRIPE confirmed that both variants resulted in aberrant transcript isoforms with altered coding potential. For example, in an individual (PMD-P11) with mitochondrial DNA depletion syndrome, clinical testing identified a VUS in *SUCLG1* (NM_003849.4:c.98-9A>G) located 9 bp upstream of the intron 1 acceptor splice site. TEQUILA-seq reads revealed that this variant creates a cryptic acceptor splice site, which extends exon 2 by eight out-of-frame nucleotides and results in a premature termination codon (fig. S10). Similarly, an individual (PMD-P12) with combined oxidative phosphorylation deficiency had a synonymous VUS in *TRIT1* (NM_017646.6:c.702A>G, p.Ala234=) that maps to the −2 position of the intron 5 donor splice site. This variant was found to cause several complex splicing changes, predominantly two multiexon skipping events that result in frameshift-induced premature termination codons (fig. S11).

Notably, more than half (8 of 15) of the splice site region variants evaluated resulted in RNA processing changes that were more complex than single exon skipping or cryptic splice site activation. Even for pathogenic splice site region variants with studies characterizing their molecular effects, we uncovered complex splicing defects that were not previously appreciated. For example, an individual (PMD-P01) with Leigh syndrome had a rare homozygous variant in *ATP5MK* (NM_001206427.2:c.87+1G>C) that disrupts the donor splice site of intron 3. Previous reverse transcription

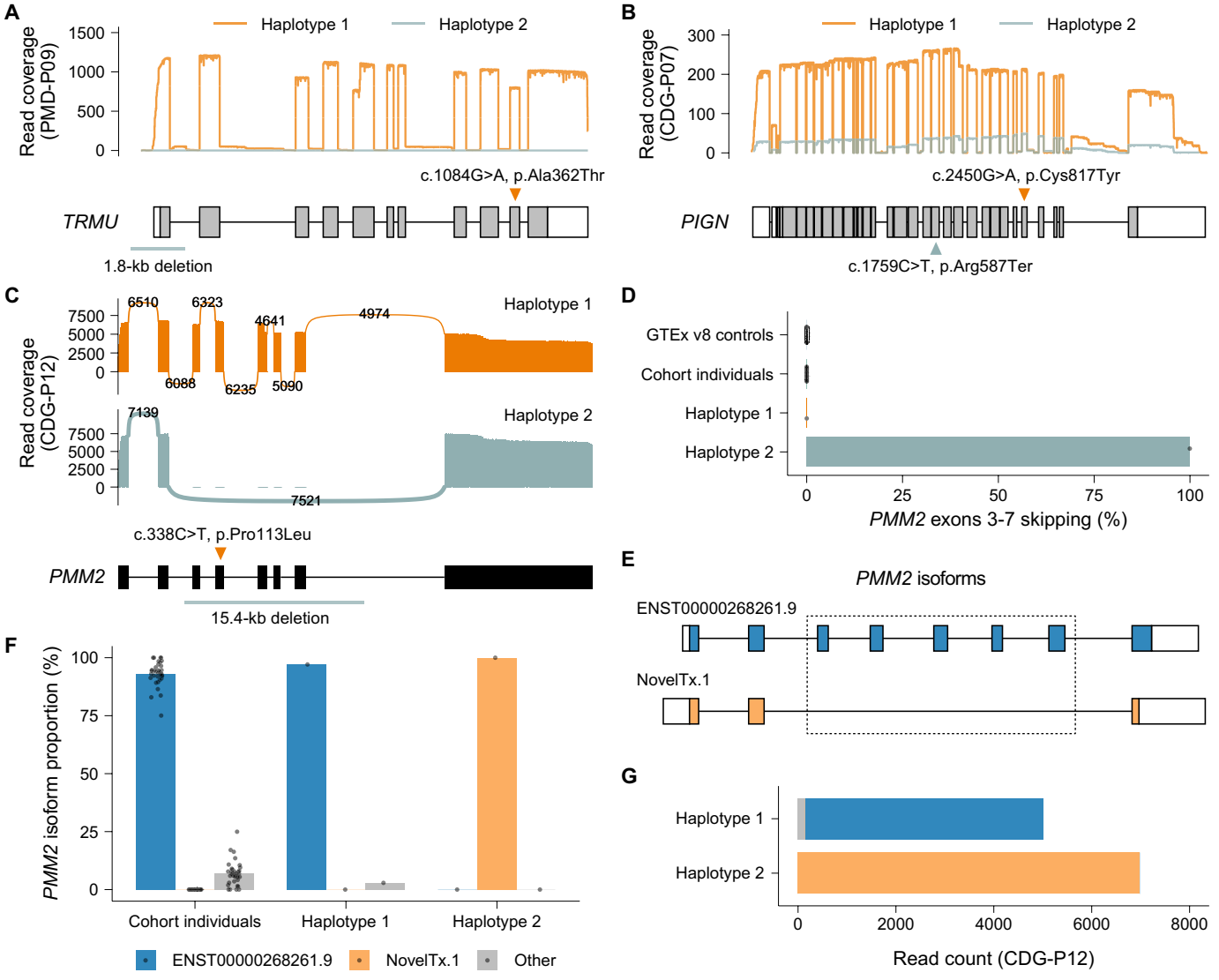

Fig. 2. Haplotype-resolved analysis of transcript consequences induced by previously identified pathogenic variants. (**A**) Haplotype-resolved read coverage over *TRMU* in patient PMD-P09, who is heterozygous for a 1.8-kb structural deletion that removes the promoter region and first exon. (**B**) Haplotype-resolved read coverage over *PIGN* in patient CDG-P07, who is heterozygous for a stop-gain variant (NM_176787.5:c.1759C>T, p.Arg587Ter). (**C**) Sashimi plot visualization of haplotype-resolved *PMM2* splicing patterns in individual CDG-P12. Haplotype 2 carries a 15.4-kb structural deletion that removes a genomic region containing exons 3 to 7. Reads that could not be confidently assigned to a haplotype (Materials and Methods) are not displayed in (A) to (C). (**D**) Usage frequencies for a splice junction involving skipping of *PMM2* exons 3 to 7 in individual CDG-P12 (resolved by haplotype), other cohort individuals, and 504 tissue-matched GTEx controls. (**E**) Full-length structures and (**F**) isoform-level proportions of *PMM2* transcripts detected from TEQUILA-seq data of individual CDG-P12 (resolved by haplotype) and other cohort individuals. Shaded and unshaded regions within transcript structures represent putative coding sequences (CDSs) and untranslated regions, respectively. Transcripts featuring skipping of *PMM2* exons 3 to 7 with the highest isoform-level proportion in haplotype 2 or transcripts with an average isoform-level proportion ≥ 10% across other cohort individuals are displayed in (E) and (F). Skipping of *PMM2* exons 3 to 7 is indicated by the dashed box in (E). (**G**) Read counts for *PMM2* transcripts displayed in (E) and (F) in individual CDG-P12 (resolved by haplotype).

polymerase chain reaction (RT-PCR) and Western blot analysis of fibroblasts from this patient suggested that this variant induces exon 3 skipping and reduces ATP5MK (ATP synthase membrane subunit K) protein level (*31*). While the pathogenicity of this variant was believed to be driven by exon 3 skipping, TEQUILA-seq data generated on the same fibroblast cell line revealed that transcripts skipping exon 3 alone made up a relatively minor proportion (7.2%) of *ATP5MK* transcripts in patient fibroblasts (Fig. 3, B to E). Rather, the predominant splicing changes induced by the donor splice site variant are combined skipping of exons 2 and 3 as well as activation of a cryptic donor splice site within exon 3, 11 bp upstream of intron 3 (49.6 and 26.8% of *ATP5MK* transcripts, respectively). Transcripts skipping either exon 3 or exons 2 and 3 are predicted to be noncoding, as the only known start codon is located in exon 3. On the other hand, transcripts using the cryptic donor splice site within exon 3 are predicted to encode a substantially altered protein that shares 43% sequence identity with the canonical protein.

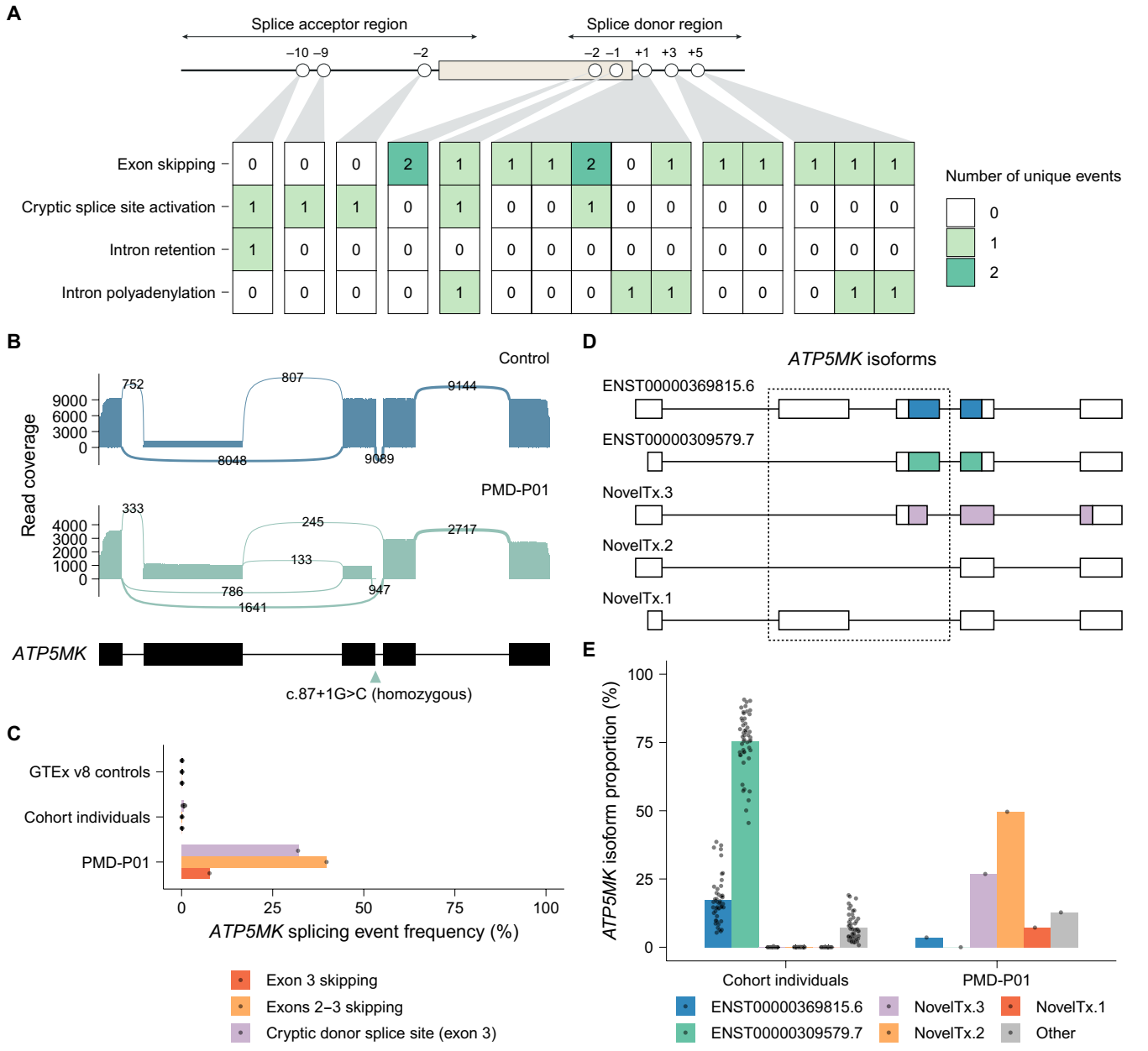

Fig. 3. Diverse RNA processing defects resulting from splice site region variants. (**A**) Number of unique RNA processing defects observed across multiple genes harboring splice site region variants that were identified from clinical testing of previously diagnosed and undiagnosed patients. Variants located in the −3 to +6 region of the donor splice site or the −20 to +3 region of the acceptor splice site were classified as "splice site region" variants. (**B**) Sashimi plot visualization of *ATP5MK* splicing patterns in individual PMD-P01, who is homozygous for a variant (NM_001206427.2:c.87+1G>C) that disrupts the donor splice site of intron 3, and an unaffected healthy control (PMD-C01). (**C**) Usage frequencies for splice junctions involving *ATP5MK* exon 3 skipping, exons 2-3 skipping, and activation of a cryptic donor splice site within exon 3 for individual PMD-P01, other cohort individuals, and 504 tissue-matched GTEx controls. (**D**) Full-length structures and (**E**) isoform-level proportions of *ATP5MK* transcripts detected from TEQUILA-seq data of individual PMD-P01 and other cohort individuals. Shaded and unshaded regions within transcript structures represent putative CDSs and untranslated regions, respectively. Transcripts featuring the events displayed in (C) with the highest isoform-level proportion in individual PMD-P01 or transcripts with an average isoform-level proportion ≥ 10% across other cohort individuals are displayed in (D) and (E). The genomic region spanning the events displayed in (C) is indicated by the dashed box in (D).

For 5 of 12 donor splice site region variants, TEQUILA-seq revealed robust activation of cryptic polyadenylation sites (PASs) downstream of the affected splice sites. In an individual (CDG-P09) with early infantile epileptic encephalopathy, clinical exome sequencing revealed two compound heterozygous variants in *PIGQ*, one of which was a paternally inherited donor splice site region variant affecting intron 4 (NM_004204.5:c.942+1G>A) (*29*). Analysis of phased TEQUILA-seq reads revealed that 82.8% of *PIGQ* transcripts expressed from the paternal haplotype featured usage of a cryptic PAS located ~1 kb downstream of the mutated splice site together with retention of the upstream intron (Fig. 4, A to D). The resulting transcript isoform is predicted to encode a severely

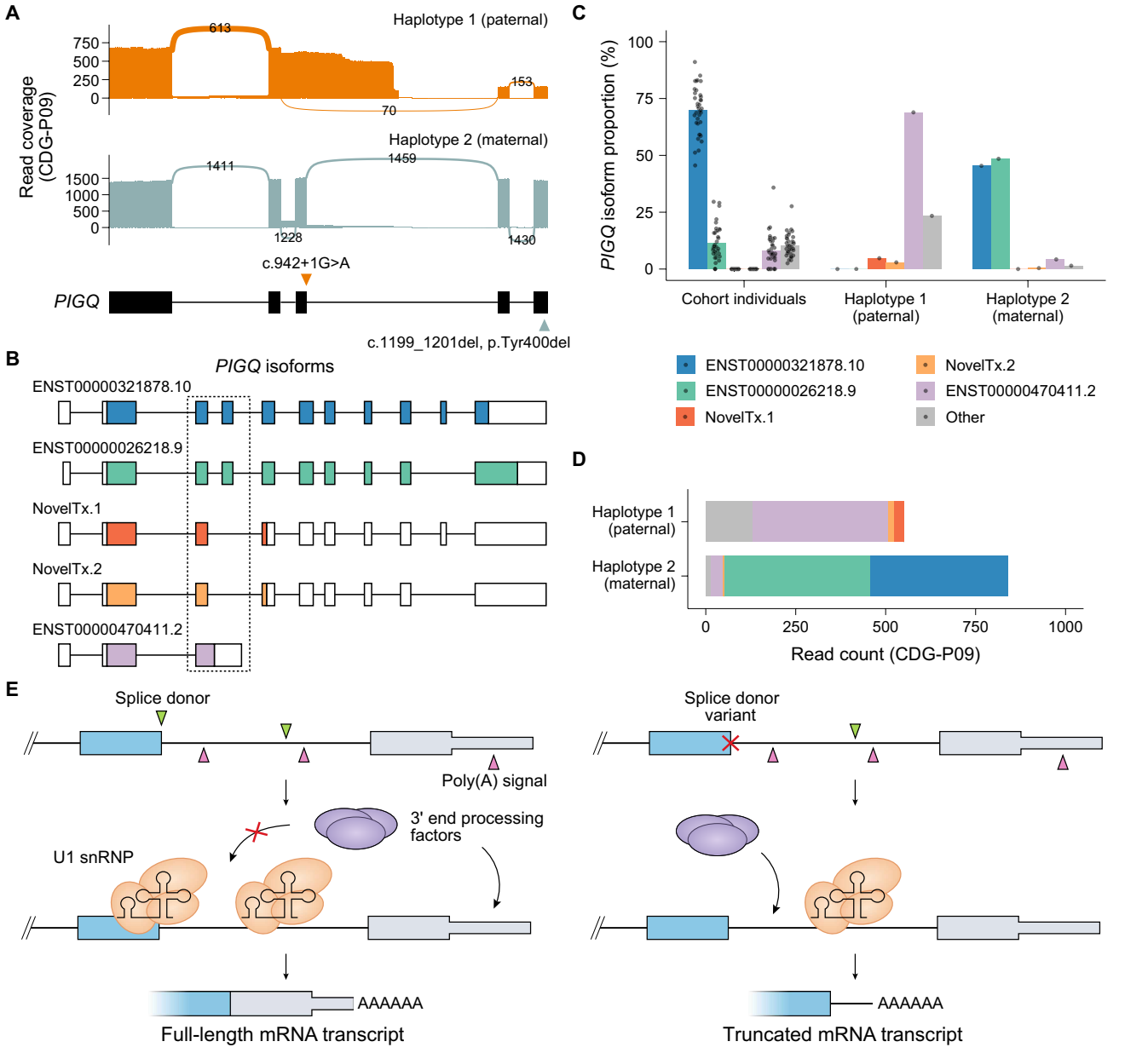

Fig. 4. Donor splice site mutations activate downstream cryptic PASs. (**A**) Sashimi plot visualization of haplotype-resolved splicing patterns around *PIGQ* exons 2 to 6 in individual CDG-P09. Haplotype 1 (paternal) carries a variant (NM_004204.5:c.942+1G>A) that disrupts the donor splice site of intron 4. Reads that could not be confidently assigned to a haplotype (Materials and Methods) are not displayed. (**B**) Full-length structures and (**C**) isoform-level proportions of *PIGQ* transcripts detected from TEQUILA-seq data of individual CDG-P09 and other cohort individuals. Shaded and unshaded regions within transcript structures represent putative CDSs and untranslated regions, respectively. Transcripts featuring *PIGQ* exon 4 skipping or usage of a PAS within intron 4, with the highest isoform-level proportion in haplotype 1 or transcripts with an average isoform-level proportion ≥ 10% across other cohort individuals are displayed in (B) and (C). The genomic region spanning these events is indicated by the dashed box in (B). (**D**) Read counts for *PIGQ* transcripts displayed in (B) and (C) in individual CDG-P09 (resolved by haplotype). (**E**) Most human introns harbor cryptic PASs (pink) and cryptic donor splice sites (green). Base pairing between U1 snRNP and its cognate sequences (canonical and cryptic donor splice sites) along nascent pre-mRNAs is known to prevent 3′ end processing factors from acting upon cryptic PASs within introns. However, mutations in donor splice sites disrupt base pairing with U1 snRNP, allowing 3′ end processing factors to act on the first immediate PAS that becomes accessible. poly(A), polyadenylate.

truncated protein that lacks the catalytic domain of the PIGQ enzyme. By contrast, transcripts skipping exon 4 made up a small proportion (8.8%) of *PIGQ* transcripts from the paternal haplotype and are predicted to harbor frameshift-induced premature termination codons. In four other individuals with donor splice site region variants, we also found that activation of cryptic PASs within the affected introns frequently emerged as the predominant, if not only, transcript consequence (fig. S12). Variants in donor splice sites are known to affect splicing by disrupting base-pairing interactions with U1 small nuclear ribonucleoprotein (snRNP), a core component of the spliceosome (*47*). Notably, previous studies have shown that U1 snRNP also plays a nonsplicing role in suppressing cryptic

PASs within introns by binding to neighboring canonical and cryptic donor splice sites (Fig. 4E) (*48*, *49*). When interactions between U1 snRNP and pre-mRNA transcripts are disrupted by variants in canonical donor splice sites, cleavage and polyadenylation can take place at the first immediate PAS that becomes accessible, leading to severely truncated transcripts, consistent with our TEQUILA-seq data on patients with rare disease.

STRIPE uncovers disease-causing variants in previously undiagnosed patients

Lastly, we used STRIPE to perform a gene panel-wide analysis of transcript aberrations and sequence variants in 42 patients with rare disease for whom prior genetic testing revealed only a single pathogenic variant in a gene associated with a recessive condition ($n = 4$) or no candidate genes ($n = 38$). Of note, we restricted this analysis to target genes with sufficient read coverage (Materials and Methods) to ensure reliable variant detection as well as robust quantification of splicing and gene expression dosage in individual patient samples. Across the 42 patients, STRIPE reported a total of 8, 78, and 192 target genes with gene expression dosage imbalance, aberrant splice junctions, and rare deleterious sequence variants, respectively (tables S9 to S11). From these results, we were able to uncover disease-causing variants in three individuals, all of which highlight several key advantages of long-read RNA-seq in rare disease diagnosis.

In one individual (CDG-P16), trio exome sequencing revealed a maternally inherited frameshift deletion in *NGLY1* exon 10 (NM_018297.4:c.1533_1536del, p.Asn511LysfsTer51). However, neither exome nor short-read RNA-seq analysis of patient blood (*50*) revealed a second variant in trans. Patient TEQUILA-seq reads derived from the maternal haplotype revealed a substantial decrease in *NGLY1* transcripts relative to the paternal haplotype, likely due to NMD (Fig. 5A). We also observed an unusually high proportion of transcripts (53.0%) expressed from the maternal haplotype featuring skipping of exon 11, which encodes the oligosaccharide-binding domain of the NGLY1 enzyme (*51*) and is naturally skipped at low frequencies in GTEx fibroblasts (interquartile range: 6.5 to 11.1%) and other fibroblast samples from our cohort (7.5 to 10.1%) (Fig. 5, B to E). This observed increase in skipping is likely due to NMD preferentially degrading transcripts that include exon 11, as the frameshift deletion in exon 10 introduces a premature termination codon only when exon 11 is present. On the other hand, we found that *NGLY1* transcripts expressed from the paternal haplotype showed nearly complete skipping of exon 11. Targeted long-read gDNA sequencing of *NGLY1* subsequently revealed a paternally inherited intronic variant (NM_018297.4:c.1612-34A>G) located upstream of exon 11 (Fig. 5F). While this variant is absent in the general population (gnomAD allele count: 0), it was previously reported as a VUS in ClinVar (4 April 2023). To ascertain whether the intronic c.1612-34A>G variant is responsible for exon 11 skipping, we generated minigene reporter plasmids containing *NGLY1* exon 11 and its flanking intronic sequences. We then transfected these minigenes with or without the c.1612-34A>G variant into K562 cells and analyzed splicing patterns by RT-PCR. While 18.6 to 19.4% of minigene transcripts show exon 11 inclusion in the presence of the reference A allele, the mutant G allele causes complete skipping of exon 11 (Fig. 5G). Furthermore, RT-PCR analysis of *NGLY1* intron 10 lariats in control fibroblasts revealed that the c.1612-34A>G variant disrupts the branchpoint adenosine (fig. S13), providing a mechanism by which this variant induces exon 11 skipping. On the basis of this information, we prioritized the paternally inherited branchpoint variant as a previously unknown disease-causing variant, which helped complete a molecular diagnosis of NGLY1 deficiency for this patient. This diagnosis was further confirmed with detection of a urine biomarker for NGLY1 deficiency (fig. S14).

We were also able to achieve molecular diagnoses in two patients with no candidate genes. In one individual (PMD-P23) with Leigh syndrome, we identified two aberrant splicing events in *MTFMT* from patient TEQUILA-seq data, including partial exon 4 skipping and inclusion of a 53-bp pseudoexon within intron 6. Both splicing events result in unproductive *MTFMT* transcript isoforms with premature termination codons, and phasing information from TEQUILA-seq reads confirmed that the two events occur in trans (Fig. 6, A to E). Concordantly, targeted long-read gDNA sequencing revealed two compound heterozygous *MTFMT* variants underlying the observed splicing defects: a missense variant on exon 4 (NM_139242.4:c.626C>T, p.Ser209Leu) and a deep intronic variant within intron 6 (NM_139242.4:c.814-2763A>G) (Fig. 6F). The c.626C>T missense variant has been previously reported in patients with *MTFMT*-related Leigh syndrome and is known to cause exon 4 skipping by eliminating a splicing enhancer sequence (*52*, *53*). On the other hand, the deep intronic c.814-2763A>G variant, which is absent in the general population (gnomAD allele count: 0), creates a cryptic AG splice acceptor site flanking the observed pseudoexon. The c.626C>T missense variant had been previously identified through clinical testing. However, we were blinded to this information, and STRIPE resolved this case through a gene panel-wide analysis, without knowledge of prior clinical testing data. In addition, clinical short-read RNA-seq and whole-transcriptome long-read RNA-seq data failed to capture the two aberrant *MTFMT* splicing events due to low read coverage (fig. S15), likely driven by degradation of the resulting transcripts via NMD.

Finally, in another individual (CDG-P22) with a clinical diagnosis of type I CDG but nondiagnostic findings from exome sequencing, STRIPE detected a rare homozygous missense variant in *ALG1* exon 13 (NM_019109.5:c.1312C>T, p.Arg438Trp). This variant, which disrupts a glycosyltransferase domain of the ALG1 enzyme, was missed by exome sequencing due to a lack of read coverage over exon 13 (Fig. 7A). The absence of reads covering this exon can be explained by the genomic architecture of *ALG1*, which is known to have at least eight pseudogenes that share strong sequence identity with *ALG1* exons 10 to 13 (Fig. 7B). The coding region of exon 13, in particular, possesses a remarkably high sequence identity (92.4 to 97.7%) with known *ALG1* pseudogenes, further complicating mapping of short sequencing reads. Notably, the c.1312C>T variant observed in *ALG1* is also present in the reference genome sequences of three *ALG1* pseudogenes (*ALG1L9P*, *ALG1L7P*, and *ALG1L8P*), likely due to the variant being located in a CpG dinucleotide context (Fig. 7C). While it is possible for variants in *ALG1* pseudogenes to masquerade as variants in the primary *ALG1* sequence based on short-read sequencing, nearly all TEQUILA-seq reads harboring the c.1312C>T variant (16,332 of 16,679) were uniquely mapped to the *ALG1* locus (Fig. 7D). Furthermore, targeted long-read gDNA sequencing of *ALG1* confirmed that this variant is present in a homozygous state in the patient (fig. S16). Subsequent *N*-glycan analysis of patient serum revealed unusually high levels of *N*-tetrasaccharide (Fig. 7E), which biochemically confirmed the diagnosis of ALG1-CDG in this patient.

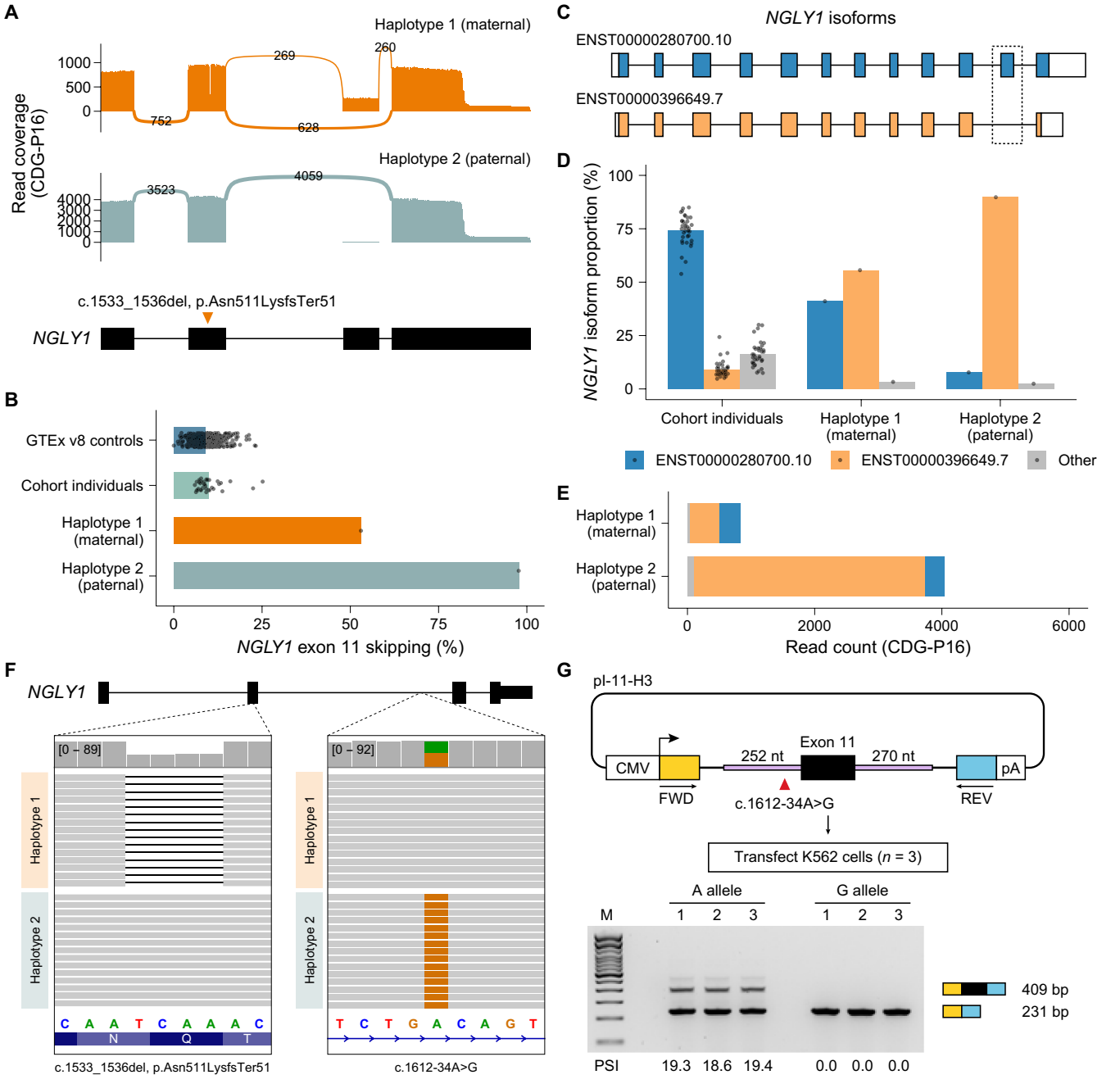

Fig. 5. STRIPE uncovers a disease-causing variant in an unsolved case of NGLY1 deficiency. (**A**) Sashimi plot visualization of haplotype-resolved splicing patterns around *NGLY1* exons 9 to 12 in individual CDG-P16. Haplotype 1 (maternal) carries a frameshift deletion variant on exon 10 (NM_018297.4:c.1533_1536del, p.Asn511LysfsTer51). Reads that could not be confidently assigned to a haplotype (Materials and Methods) are not displayed. (**B**) Usage frequencies for a splice junction involving *NGLY1* exon 11 skipping in individual CDG-P16 (resolved by haplotype), other cohort individuals, and 504 tissue-matched GTEx controls. (**C**) Full-length structures and (**D**) isoform-level proportions of *NGLY1* transcripts detected from TEQUILA-seq data of individual CDG-P16 (resolved by haplotype) and other cohort individuals. Shaded and unshaded regions within transcript structures represent putative CDSs and untranslated regions, respectively. Transcripts featuring *NGLY1* exon 11 skipping with the highest isoform-level proportion in haplotype 2 or transcripts with an average isoform-level proportion ≥ 10% across other cohort individuals are displayed in (C) and (D). *NGLY1* exon 11 skipping is indicated by the dashed box in (C). (**E**) Read counts for *NGLY1* transcripts displayed in (C) and (D) in individual CDG-P16 (resolved by haplotype). (**F**) ONT gDNA sequencing data (via adaptive sampling) for individual CDG-P16 identified the following variants in *NGLY1*: a maternally inherited frameshift deletion variant on exon 10 (NM_018297.4:c.1533_1536del, p.Asn511LysfsTer51) and a paternally inherited intronic variant upstream of exon 11 (NM_018297.4:c.1612-34A>G). (**G**) Schematic illustration of a minigene reporter plasmid (pI-11-H3) containing *NGLY1* exon 11 and its flanking intronic sequences (top). Minigenes with or without the intronic c.1612-34A>G variant were transfected into K562 cells ($n = 3$), and resulting splicing products were analyzed using RT-PCR (bottom). nt, nucleotide; CMV, cytomegalovirus; FWD, forward; REV, reverse; PSI, percent spliced in; M, marker; pA, poly(A) tail.

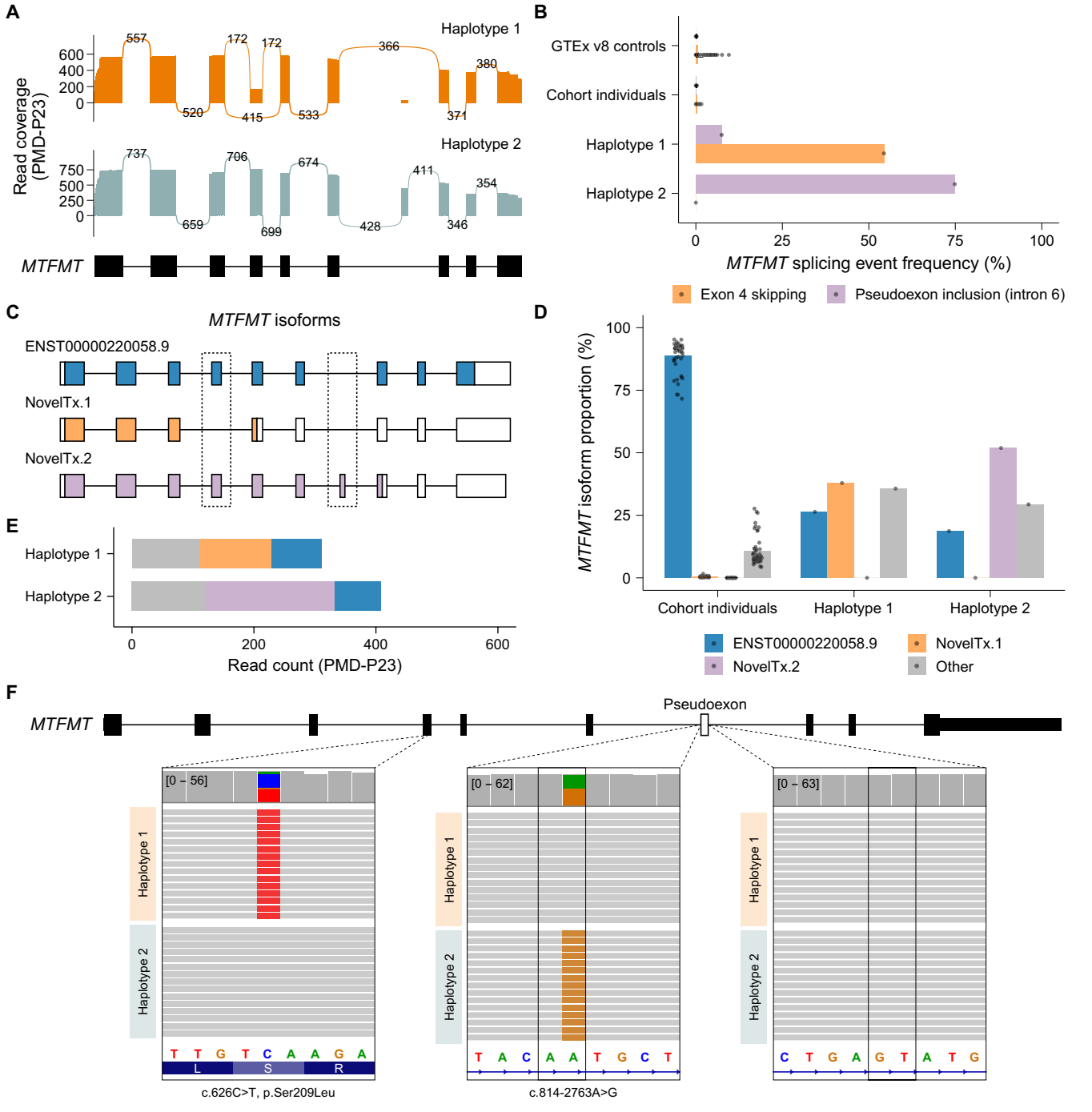

Fig. 6. Detection and phasing of two *MTFMT* splicing defects in an undiagnosed Leigh syndrome patient. (**A**) Sashimi plot visualization of haplotype-resolved *MTFMT* splicing patterns in individual PMD-P23. Reads that could not be confidently assigned to a haplotype (Materials and Methods) are not displayed. (**B**) Usage frequencies for splice junctions involving *MTFMT* exon 4 skipping and inclusion of a 53-bp pseudoexon within *MTFMT* intron 6 in individual PMD-P23 (resolved by haplotype), other cohort individuals, and 504 tissue-matched GTEx controls. (**C**) Full-length structures and (**D**) isoform-level proportions of *MTFMT* transcripts detected from TEQUILA-seq data of individual PMD-P23 (resolved by haplotype) and other cohort individuals. Shaded and unshaded regions within transcript structures represent putative CDSs and untranslated regions, respectively. Transcripts featuring *MTFMT* exon 4 skipping with the highest isoform-level proportion in haplotype 1, transcripts featuring inclusion of the pseudoexon within intron 6 with the highest isoform-level proportion in haplotype 2, or transcripts with an average isoform-level proportion ≥ 10% across other cohort individuals are displayed in (C) and (D). These events are indicated by the dashed boxes in (C). (**E**) Read counts for *MTFMT* transcripts displayed in (C) and (D) in individual PMD-P23 (resolved by haplotype). (**F**) ONT gDNA sequencing data (via adaptive sampling) for individual PMD-P23 identified the following compound heterozygous variants in *MTFMT*: a rare missense variant on exon 4 (NM_139242.4:c.626C>T, p.Ser209Leu) and a deep intronic variant within intron 6 (NM_139242.4:c.814-2763A>G).

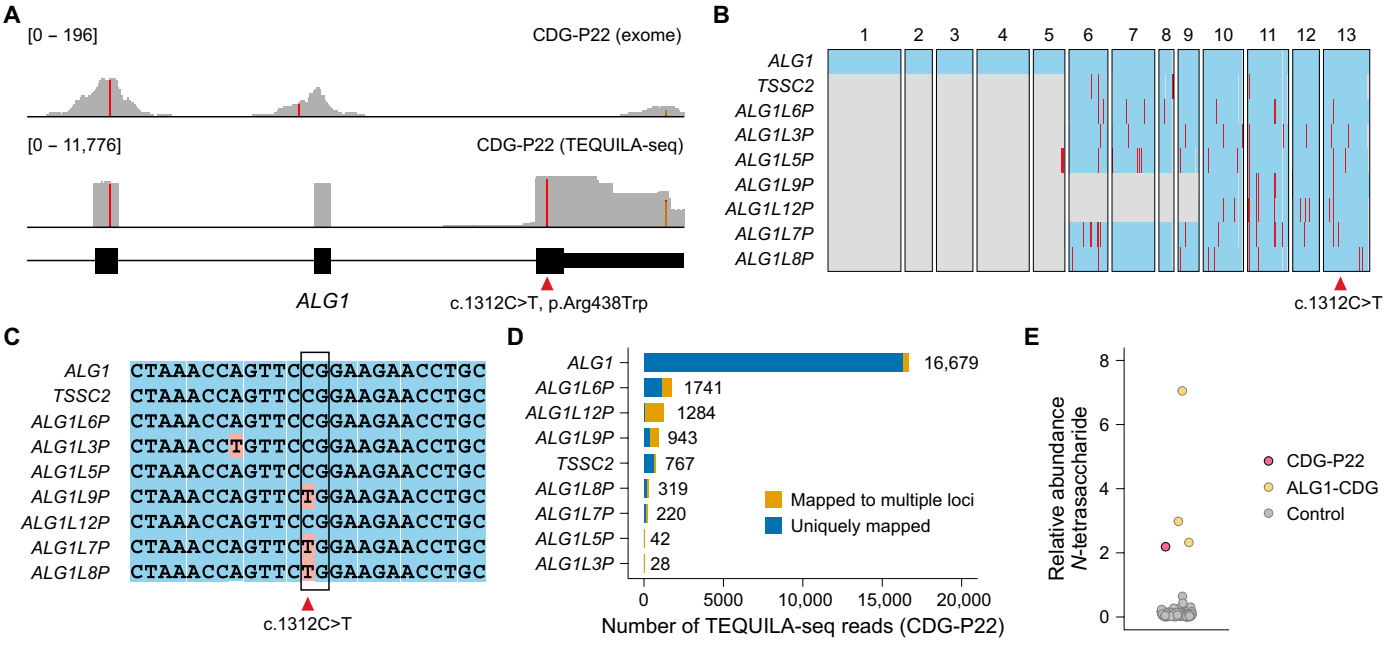

Fig. 7. STRIPE identifies a missense variant within a highly duplicated region of *ALG1*. (**A**) Exome sequencing and TEQUILA-seq read coverage of *ALG1* exons 11 to 13 for patient CDG-P22, who was found to carry a missense variant (NM_019109.5:c.1312C>T, p.Arg438Trp) on exon 13. (**B**) Multiple sequence alignment between the CDS of *ALG1* (partitioned by exon) and its eight known pseudogenes (*TSSC2*, *ALG1L6P*, *ALG1L3P*, *ALG1L5P*, *ALG1L9P*, *ALG1L12P*, *ALG1L7P*, and *ALG1L8P*). Matches, mismatches, and gaps are represented in blue, red, and gray, respectively. (**C**) The CDS in *ALG1* surrounding the c.1312C>T missense variant shares strong sequence identity with sequences from known *ALG1* pseudogenes. The variant is found in a CpG dinucleotide context, which likely explains its appearance in the reference sequences of *ALG1L9P*, *ALG1L7P*, and *ALG1L8P*. (**D**) Number of uniquely mapped and multimapped TEQUILA-seq reads from patient CDG-P22 that are aligned to *ALG1* and its known pseudogenes. (**E**) Relative abundances of *N*-tetrasaccharide, a biomarker for ALG1-CDG, in serum samples from patient CDG-P22, three confirmed cases of ALG1-CDG, and 98 age-matched controls, as measured by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (at 1124 mass/charge ratio).

DISCUSSION

Long-read RNA-seq has substantially enhanced our ability to analyze cis effects of rare genetic variants on the transcriptome (*15*, *16*). Recent applications of this technology to individual rare disease cases have yielded insights into the molecular consequences of known disease-causing variants (*18*, *19*). However, the high sequencing depths required, especially when analyzing clinically accessible tissues, make long-read RNA-seq cost prohibitive for regular genetic diagnoses. In this work, we present STRIPE, a long-read RNA-seq–based strategy for rare disease diagnosis and variant interpretation (Fig. 1, A and B). By harnessing TEQUILA-seq (*21*), a cost-effective and versatile method for targeted long-read RNA-seq, STRIPE can provide ultradeep coverage of full-length transcripts for any disease-specific gene panel while preserving quantitative gene expression and splicing patterns. In terms of sequencing costs, the per-sample RNA-to-data cost of running STRIPE is currently estimated to be as low as $100 on Oxford Nanopore platforms, making it feasible to deploy at scale in various settings. Building on this foundation, STRIPE incorporates several analytical components designed to facilitate interpretation of targeted long-read RNA-seq data in a rare disease context. While STRIPE leverages existing variant calling tools to identify candidate sequence variants, it also implements approaches for detecting outlier splicing events at haplotype-level resolution and genes with extreme dosage imbalance. Although such approaches are established in short-read RNA-seq–based workflows (*12*, *27*), their systematic adaptation to long-read transcriptomes across a large rare disease cohort has not been previously achieved, representing a key advance of this study.

Applying STRIPE to skin fibroblast RNAs from 88 individuals spanning two different rare disease groups (CDG and PMD), we demonstrate that our approach can provide multiple readouts that are clinically informative for both previously diagnosed and undiagnosed patients. First, the long read lengths and deep sequencing coverage provided by STRIPE enable robust identification and phasing of sequence variants in exonic regions. STRIPE reidentified all 23 known pathogenic exonic variants that were reported from clinical testing of previously diagnosed patients (fig. S4A) and phased 18 of the 23 variants using TEQUILA-seq data alone, including two compound heterozygous missense variants that were separated by ~97.4 kb (fig. S4B). Notably, the median genomic distance between biallelic pathogenic variants in our cohort was ~16.7 kb (table S1), which far exceeds the typical read span of short-read sequencing libraries. Computational phasing approaches are also unlikely to succeed in this context, given that these variants are often extremely rare and absent from population reference panels (*54*, *55*). In addition, we show that STRIPE has the capacity to ascertain sequence variants within highly duplicated and repetitive regions of the transcriptome, which are known to elude detection by traditional short-read sequencing methods (*14*, *56*, *57*). For example, in a patient (CDG-P22) with type I CDG, STRIPE revealed a homozygous missense variant within a highly duplicated region of *ALG1* that prior clinical exome sequencing failed to capture (Fig. 7).

STRIPE can also comprehensively profile gene expression and splicing patterns at haplotype-level resolution, providing an opportunity to map transcript aberrations to rare variants, including disease-causing ones. While traditional short-read RNA-seq can also detect aberrant changes in gene expression or splicing, it has limited capacity to link such changes, especially splicing defects, to individual haplotypes given that short reads rarely cover both heterozygous loci and isoform-informative splice junctions (*58*, *59*). Using phased TEQUILA-seq data for previously diagnosed patients with heterozygous pathogenic variants, we show that haplotypes with reduced gene expression or aberrant splicing often carry putative protein-truncating variants (figs. S5 and S6), which are known to have profound effects on gene regulation through various mechanisms such as NMD or disruption of splicing (*39*, *60*). Additional studies, such as coupling NMD inhibition with transcriptome profiling (*61*), could help further clarify the molecular mechanisms by which these putative protein-truncating variants exert their effects.

Furthermore, our ability to detect haplotype-specific aberrations in gene expression and splicing allowed us to not only clarify VUS but also uncover disease-causing variants with large regulatory effects (Figs. 5 and 6 and figs. S10 and S11) in four individuals with nondiagnostic findings from previous genetic testing. Notably, in two of these individuals, short-read RNA-seq did not resolve the genetic diagnosis. In a patient with NGLY1 deficiency, both alleles showed skipping of the same exon (exon 11): One allele carried a known frameshift deletion in exon 10 that would trigger transcript degradation when exon 11 is included, and the other carried an exon 11 branchpoint mutation missed by exome sequencing (Fig. 5). While short-read RNA-seq detected the exon skipping event, it could not determine whether this signal was driven solely by the frameshift variant due to a lack of haplotype resolution (*50*). In a patient with MTFMT deficiency, both alleles carried splicing defects that were tightly coupled to NMD, which led to negligible read coverage over *MTFMT* in standard short-read and long-read RNA-seq data (fig. S15). By contrast, targeted long-read RNA-seq of the same patient sample yielded sufficient depth to both detect and phase the two splicing defects, enabling us to uncover their causal variants and complete the patient's diagnosis (Fig. 6).

An important observation made in our study is that the precise nature of variant-induced mis-splicing is often not obvious. Even in the case of splice site region variants, it often remains unclear whether these variants will lead to exon skipping, cryptic splice site activation, or intron retention (*46*). Understanding the exact transcript consequences of splice-altering variants is especially important in clinical settings because it not only informs assessments of variant pathogenicity but also creates opportunities to design personalized splice-modulating oligonucleotide therapies (*62*). Across our patient cohort, we observed a wide range of RNA processing defects, including many unexpected events, caused by splice site region variants found from prior testing (Fig. 3A). More than half of these variants (8 of 15) resulted in more complex RNA processing defects beyond single exon skipping or cryptic splice site activation alone. In particular, we demonstrate that donor splice site variants frequently activate cryptic PASs downstream of the affected splice sites (Fig. 4, A to C, and fig. S12). This particular outcome has not been described in previous studies using RNA-seq for rare disease investigation (*5*–*12*). Nonetheless, the coupling of donor splice site variants to cryptic PASs observed in our patients with rare disease is consistent with and provides compelling genetic evidence for the proposed dual roles of U1 snRNP in recognizing donor splice sites and suppressing cryptic PASs (Fig. 4E) (*48*, *49*). In the absence of U1 snRNP bound to the canonical donor splice site, 3′ end processing factors can cleave pre-mRNA transcripts from the first intronic PAS that becomes accessible. Given the pervasive contribution of donor splice site variants to rare diseases (*20*, *63*), we envision that the activation of cryptic intronic PASs may be a prevalent yet underappreciated mechanism by which these variants exert their pathogenicity.

Because STRIPE is based on targeted long-read RNA-seq, it is best suited for patient cohorts that have been prescreened for disease phenotypes that fall into well-defined rare disease categories (e.g., CDG and PMD). Ideally, the relevant disease genes should also be adequately detectable in clinically accessible tissues, such as whole blood and skin fibroblasts. Nonetheless, in certain disease contexts, phenotype-relevant genes may show restricted or low expression in accessible tissues, limiting the diagnostic utility of standard transcriptome sequencing (*20*). The targeted design of STRIPE helps to mitigate this limitation by concentrating sequencing capacity on predefined genes of interest, thus achieving substantially improved read depth even for genes with low baseline expression. This advantage is exemplified by the patient with MTFMT deficiency, where the gene is well covered by TEQUILA-seq (Fig. 6) but barely detectable in whole-transcriptome RNA-seq data (fig. S15).

In a separate vein, an inherent tradeoff of targeted long-read RNA-seq is constraining diagnostic breadth to predefined disease genes. This limitation becomes particularly relevant for disease phenotypes with high genetic heterogeneity, where causal variants may occur in genes not represented on a given panel. While in principle deep whole-transcriptome long-read RNA-seq could broaden coverage to all genes, achieving the sequencing depth required to confidently investigate lowly expressed disease genes across large patient cohorts remains prohibitively expensive (*64*). To mitigate this limitation and expand breadth, curated gene panels can be designed to include not only known disease genes but also biologically relevant genes that currently lack definitive disease associations, thereby enabling opportunities for gene discovery. Moreover, owing to the relative ease and substantially lower cost of generating large pools of capture oligos with TEQUILA compared to commercial providers (e.g., IDT and Twist Bioscience) (*21*), targeted long-read RNA-seq panels used by STRIPE can be readily and regularly updated to include new genes, as more insights are gained into the genetic basis of different conditions. It is also worth noting, however, that while the current implementation of STRIPE is based on targeted sequencing, the underlying computational framework is fully adaptable to deep whole-transcriptome long-read RNA-seq data, as such datasets become more available in the future. In this regard, downstream analyses and reporting can focus on selected gene subsets, guided by patient phenotypes and gene-phenotype relationships.

Finally, while the present study demonstrates the development and application of STRIPE in a research setting, its impact could be substantially amplified through clinical deployment. As with any transcriptome-based diagnostic workflow, implementing STRIPE in clinical settings will require addressing several practical challenges. Many of these are shared with ongoing efforts to implement clinical short-read RNA-seq, including procurement and processing of RNA samples, as well as standardization of experimental protocols and bioinformatics workflows to ensure reproducibility (*12*, *50*, *65*). The added complexity of targeted long-read RNA-seq also introduces additional considerations, such as training clinical laboratory personnel

and validating disease-specific gene panels. Although further work is needed to establish STRIPE as a clinical test, we envision that STRIPE can serve as a uniform supplement to exome and genome sequencing within a multiomic diagnostic framework. In addition, our ability to resolve genetic diagnoses without relying on information from prior testing (e.g., the patient with MTFMT deficiency; Fig. 6) suggests that STRIPE could potentially be deployed as a frontline test, particularly in resource-constrained settings, although further studies will be needed to systematically evaluate this possibility.

Overall, we demonstrate that targeted long-read RNA-seq is a powerful and scalable strategy that can expand our understanding of rare disease mechanisms, facilitate new molecular diagnoses, and inform the development of future personalized therapies.

MATERIALS AND METHODS

Experimental design

This study aimed to develop a targeted long-read RNA-seq strategy to enhance variant interpretation and facilitate rare disease diagnosis. We reasoned that sequencing full-length transcripts from patient-derived tissues could more precisely characterize the cis functional effects of genetic variants identified through prior clinical DNA testing, including VUS, and uncover disease-causing variants missed by standard approaches. To address this, we implemented a targeted long-read RNA-seq workflow, STRIPE, and applied it to skin fibroblast RNAs from 88 individuals spanning two groups of rare diseases: CDG and PMD. Capture panels were designed to selectively enrich disease-relevant genes and enable deep sequencing of full-length transcripts with haplotype-level resolution. We first validated STRIPE's capacity to detect and phase known pathogenic variants and their transcript-level consequences in 22 previously diagnosed patients. We then performed panel-wide analyses across 46 genetically undiagnosed patients to identify gene expression dosage imbalance, aberrant splicing events, and rare deleterious sequence variants, with the goal of resolving VUS and uncovering disease-causing variants.

Patient cohorts and clinical samples

Local institutional review boards of participating institutions provided study approval and oversight for research protocols enabling skin punch biopsies for fibroblast cell line establishment and genetic/transcriptomic analysis of individuals suspected of having rare diseases (IRB #14-0112233 for CDG; IRB #08-6177 and #24-22190 for PMD). All patients and/or guardians of patients provided informed consent before inclusion in the studies. This study was performed in accordance with ethical principles for medical research outlined in the Declaration of Helsinki of 1975, as revised in 2013.

Cell culture

Skin fibroblasts from individuals subjected to TEQUILA-seq with the CDG gene panel were cultured in Dulbecco's modified Eagle's medium (DMEM; Gibco, #11885084) supplemented with 10% fetal bovine serum (FBS; Corning, #45000-734), 1% GlutaMAX (Gibco, #35050061), and penicillin-streptomycin (100 U/ml; Gibco, #15140-122). Skin fibroblasts from individuals subjected to TEQUILA-seq with the PMD gene panel were cultured in DMEM supplemented with 10% FBS, uridine (50 μg/ml; Thermo Fisher Scientific Chemicals, #A15227.06), and penicillin-streptomycin (100 U/ml). K562 cells (American Type Culture Collection, #CCL-243) were cultured in RPMI 1640 medium (Gibco, #11875093) supplemented with 10% FBS and penicillin-streptomycin (100 U/ml). Cells were maintained at 37°C in a humidified incubator with 5% CO_2.

RNA isolation and full-length cDNA synthesis

Total RNA was isolated from fibroblast cell lines using TRIzol reagent (Invitrogen, #15596018) according to the manufacturer's instructions. RNA concentration and integrity were measured with a NanoDrop 2000 Spectrophotometer and an Agilent 4200 TapeStation, respectively. A total of 200 ng of total RNA was used in a reverse transcription and template switching reaction with Maxima H Minus Reverse Transcriptase (Thermo Fisher Scientific, #EP0753) for 90 min at 42°C, followed by 5 min at 85°C. First-strand cDNA was then PCR amplified using the KAPA HiFi HotStart ReadyMix (KAPA Biosystems, #KK2602) under the following program: initial denaturation at 95°C for 3 min, 12 cycles of (98°C for 20 s, 67°C for 20 s, and 72°C for 5 min), and final extension at 72°C for 8 min. PCR products were purified with 0.8X SPRIselect beads (Beckman Coulter, #B23318) and quantified using the Qubit dsDNA High Sensitivity Assay and Agilent High Sensitivity D5000 ScreenTape Assay on an Agilent 4200 TapeStation. The sequences of oligos and primers used are listed in table S12.

TEQUILA probe synthesis and hybridization capture

Two custom gene panels were designed for probe synthesis: a CDG panel targeting 466 genes and a PMD panel targeting 359 genes (fig. S2A and tables S2 and S3). The CDG panel was compiled across five different sources, including 183 genes known to underlie CDG (*37*), 386 genes from the GlyCosmos Portal (*66*), 87 genes encoding glycoside hydrolases (*67*), 349 murine glycosylation-related genes with known human orthologs (*68*), and 213 genes encoding glycosyltransferases and sulfotransferases (*69*). On the other hand, the PMD panel was constructed on the basis of the Combined Mito Genome Plus Mito Focused Nuclear Gene Panel from GeneDx, the Nuclear Mitochondrial Disorders Panel from Invitae, and expert consultation (R.D.G.). Notably, the PMD panel focuses exclusively on nuclear-encoded mitochondrial genes, as genes within the mitochondrial genome are routinely assessed through existing mitochondrial DNA diagnostic workflows.

TEQUILA probes for these gene panels were synthesized as previously described (*21*). All probe sequences are available on GitHub (https://github.com/Xinglab/STRIPE) and archived at Zenodo (https://doi.org/10.5281/zenodo.17538307). Approximately 500 ng of amplified cDNA was denatured at 95°C for 10 min and incubated with 100 ng of TEQUILA probes for 12 hours at 65°C. The mixture was then incubated with 50 μl of M-270 streptavidin beads (Invitrogen, #65306) for 45 min at 68°C. The resulting mixture was subjected to a series of washes according to the IDT xGen Lockdown protocol, followed by resuspension in 40 μl of TE buffer (Invitrogen, #12090015). Captured cDNA was amplified using on-bead PCR with KAPA HiFi HotStart ReadyMix under the following program: initial denaturation at 95°C for 3 min, 14 cycles of (98°C for 20 s, 67°C for 20 s, and 72°C for 5 min), and final extension at 72°C for 8 min. PCR products were purified with 0.7X SPRIselect beads and quantified using the Qubit dsDNA High Sensitivity Assay and Agilent High Sensitivity D5000 ScreenTape Assay on an Agilent 4200 TapeStation.

1D cDNA library construction and nanopore sequencing

Pre- and postcapture amplified cDNA products were end-repaired and dA-tailed using the NEBNext Ultra II End Repair/dA-Tailing Module (NEB, #E7546) for 20 min at 20°C, followed by 20 min at 65°C. The resulting cDNA was purified with 0.7X SPRIselect beads and used to make nanopore 1D sequencing libraries according to the ONT ligation sequencing kit (SQK-LSK109) or native barcoding kit (SQK-NBD114-24) protocols. Libraries were then loaded onto R9.4.1/R10.4.1 flow cells and sequenced on MinION/PromethION devices. Summary statistics for all nanopore sequencing runs are provided in table S4.

Genomic DNA isolation and selective nanopore sequencing

Genomic DNA was isolated from fibroblast cell lines using the Quick-DNA Miniprep Plus Kit (Zymo Research, #D4068) according to the manufacturer's instructions. DNA concentration and integrity were measured with a NanoDrop 2000 Spectrophotometer and a Genomic DNA ScreenTape Assay on an Agilent 4200 TapeStation, respectively. Extracted DNA was sheared to a target fragment size of 20 kbp using a g-TUBE (Covaris, #520079), and ~1 μg of sheared DNA was used to make nanopore 1D sequencing libraries according to the ONT ligation sequencing kit (SQK-LSK114) protocol. Libraries were then loaded onto R10.4.1 flow cells and sequenced on a PromethION device. Target regions were enriched using ONT's adaptive sampling option in MinKNOW (v24.02.19).

Minigene assay and RT-PCR analysis

Gene fragments containing *NGLY1* exon 11 and its flanking intronic regions (252 bp upstream and 270 bp downstream of exon 11) with and without the variant NM_018297.4:c.1612-34A>G were ordered from Twist Bioscience and cloned into the XbaI and XhoI sites of the pI-11-H3 minigene vector (*70*). All sequences and mutations were verified by Sanger sequencing. Reference and variant-containing plasmids were transfected into K562 cells ($n = 3$) using 4D-Nucleofector (Lonza, #AAF-1003X) according to the manufacturer's instructions. Total RNA was extracted from transfected cells 3 days after transfection using TRIzol reagent (Invitrogen, #15596018) according to the manufacturer's instructions and treated with deoxyribonuclease (DNase) I (Invitrogen, #18068015). Approximately 1 μg of total RNA was reverse transcribed using oligo(dT) primers (Invitrogen, #18418012), Moloney Murine Leukemia Virus (M-MLV) Reverse Transcriptase (Promega, #M1701), and 1× M-MLV Reverse Transcriptase Buffer (Promega, #M5313) according to the manufacturer's instructions for 1 hour at 37°C. RT products were then PCR amplified using primers pI-11-H3-RT-For and pI-11-H3-RT-Rev, and resulting PCR products were separated on a 1.5% agarose gel and visualized in ChemiDoc MP Imaging System (BioRad, #12003154). The intensity of individual bands was analyzed with ChemiDoc Image Lab software (v6.0.0). The sequences of oligos and primers used are listed in table S12.

Branchpoint mapping by nested PCR and Sanger sequencing

Total RNA from healthy control fibroblast cell line CDG-C03 was treated with RNase-free DNase I (TURBO DNA-free kit, Invitrogen, #AM1907). Approximately 1 μg of total RNA was reverse transcribed using Maxima H Minus Reverse Transcriptase and the Rev1 primer at 42°C for 60 min. PCR was performed in a Veriti 96-Well Thermal Cycler (Applied Biosystems, #43-757-86) using the KAPA HiFi HotStart ReadyMix with primers For and Rev1, followed by nested primers For and Rev2, under the following program: initial denaturation at 95°C for 3 min, 32 cycles of (98°C for 15 s, 65°C for 15 s, and 72°C for 15 s), and final extension at 72°C for 1 min. PCR products were analyzed using a D1000 ScreenTape assay on an Agilent 4200 TapeStation. Sanger sequencing of the nested PCR amplicon was used to determine the location of the branch point. The sequences of oligos and primers used are listed in table S12.

Basecalling and alignment of nanopore sequencing data

Raw nanopore data were basecalled using Guppy (v6.5.7) in super-accurate mode (dna_r9.4.1_450bps_sup.cfg) for MinION sequencing runsandhigh-accuracymode(dna_r10.4.1_e8.2_400bps_hac_prom.cfg) for PromethION sequencing runs. Basecalled reads from multiplexed sequencing runs were also demultiplexed using Guppy (v6.5.7). Reads were then aligned to the GRCh38 reference genome using minimap2 (v2.26) with presets "splice" and "map-ont" for nanopore long-read RNA and DNA sequencing data, respectively.

Reprocessing of patient exome and genome sequencing data

Exome and genome sequencing reads were aligned to the GRCh38 reference genome using BWA mem (v0.7.18) with default parameters. SNVs and indels were called from resulting alignment files using DeepVariant (v1.6.1) with parameter "--model_type=WES" for exome sequencing data and parameter "--model_type=WGS" for genome sequencing data.

Transcript detection and characterization

Genes and transcripts were detected and quantified from primary alignments with MAPQ ≥ 1 in long-read RNA-seq data using StringTie (v2.2.3) with GENCODE v45 annotations (using the -G parameter) and the following additional parameters: "-L -s 5 -c 5 -u -M 0". We used the following strategy to determine coding sequences (CDSs) of transcripts without existing CDS annotations. If the transcript is derived from a known protein-coding gene, we selected the longest open reading frame (ORF) starting from any annotated start codon that overlaps the transcript sequence. If there are no overlapping annotated start codons, then the longest ORF starting from any ATG contained in the transcript sequence was selected. All predicted ORFs are required to contain ≥ 20 amino acids.

Quantification of splice junction usage frequencies

We estimated the usage frequency of a splice junction using a slightly modified version of the Intron Jaccard Index (*71*). Briefly, for some splice junction J with donor splice site D and acceptor splice site A, let n_D be the number of RNA-seq reads supporting donor splice site D, n_A be the number of RNA-seq reads supporting acceptor splice site A, and n_{DA} be the number of RNA-seq reads supporting splice junction J. The usage frequency for splice junction J is calculated as $n_{DA}/(n_D + n_A - n_{DA})$, where $n_D + n_A - n_{DA} \geq 20$. Here, $n_D + n_A - n_{DA}$ represents the total read coverage around splice sites D and A.

SNV and indel calling from TEQUILA-seq data

SNVs and indels were called from TEQUILA-seq alignment files using LongcallR (v0.1.0) with preset "ont-cdna" and DeepVariant (v1.6.1) with parameter "--model_type=ONT_R104." For each target gene, we curated a set of high-confidence biallelic variants across both tools using the following criteria: (i) not located within a dense

cluster of variants (i.e., ≤3 variants within any 200-bp window); (ii) variant position is covered by ≥100 reads, and ≥30% of these reads support the variant; (iii) flagged as "PASS", "LowQual", or "RnaEdit". The resulting call set was then filtered for rare, predicted deleterious variants using the following sequential filters: (i) CADD (v1.6) score > 15 or SpliceAI (v1.3) prediction > 0.1 [tool thresholds were taken directly from (*72*)], (ii) maximum allele frequency across gnomAD v4 populations < 1%, and (iii) not annotated as "Benign" or "Likely benign" in ClinVar (17 September 2024). To remove potential false positive variants corresponding to basecalling or alignment artifacts, we further filtered for variants that were identified and prioritized in no more than three cohort individuals.

Gene-level phasing of nanopore sequencing reads

TEQUILA-seq reads mapping to a given gene were phased using an alignment-based strategy as follows. First, phased heterozygous SNVs called from TEQUILA-seq reads via LongcallR are used to construct a pair of haplotypes for the given gene. Gene haplotypes are further updated to include additional phased variants called from previous exome or genome sequencing data, if available. Haplotype-specific reference sequences are then constructed for the given gene, and TEQUILA-seq reads are realigned to both haplotype-specific sequences using minimap2 (v2.26) with preset "splice". We subsequently assigned reads with a MAPQ score ≥ 1 to the haplotype with a higher alignment score. Reads with a MAPQ score of 0, as well as reads with identical alignment scores to either of the two haplotypes, were not assigned to a haplotype. To assess the quality of our read phasing results for a given gene G, we computed a quality score Q_G as

$$Q_G = \frac{\sum_{k \in H_1 \cup H_2} w_k}{\sum_k w_k}, \quad w_k = \frac{\max[\min(T_k, T_G) - \max(S_k, S_G), 0]}{T_G - S_G}$$

where (S_G, T_G) denotes the start and end coordinates of gene G and (S_k, T_k) denotes the alignment start and alignment end coordinates of some read k mapping to gene G. For each haplotype, SNVs and indels were called from haplotype-assigned reads using bcftools (version 1.21) with parameters "-B -Q 5 --ploidy 1" and filtered on the basis of the following criteria: (i) not located within a dense cluster of variants (i.e., ≤3 variants within any 200-bp window) and (ii) variant position is covered by ≥30 reads. Haplotype-resolved variant calls were further filtered for rare, predicted deleterious variants using the same criteria as previously described. Separately, nanopore DNA sequencing reads mapping to a given gene were phased using the wf-human-variation workflow (v2.4.1) from EPI2ME Labs with parameters '--basecaller_cfg "dna_r10.4.1_e8.2_400bps_hac_prom" --snp --phased'.

Identification of target genes with gene expression dosage imbalance

Let m_1 and m_2 represent read counts for both haplotypes of a target gene (minimum phasing quality score of 0.7). We define a "haplotype expression ratio" as $m_1/(m_1 + m_2)$, where $m_1 + m_2 \geq 20$. For each target gene, we fitted a beta distribution to haplotype expression ratios observed among tissue-matched GTEx short-read RNA-seq samples (minimum of 100 samples), which were used to represent the general population. We assume that this beta distribution has a fixed mean $\mu = 1/2$ and some precision parameter θ. For some individual P, let m_{1P} and m_{2P} be the number of reads supporting both haplotypes of a target gene. Suppose m_{1P} is drawn from a beta-binomial distribution with $m_{1P} + m_{2P}$ trials and parameters μ_P (individual-specific mean) and θ (precision). Under the null hypothesis, we assume that $\mu_P = 1/2$ (i.e., individual P shows balanced expression of both gene haplotypes). Thus, the probability of observing $m_1 = k$ reads in individual P for some $k \in \{0, \ldots, m_{1P} + m_{2P}\}$ is given by

$$p(m_1 = k) = \binom{m_{1P} + m_{2P}}{k} \cdot \frac{B(k + \theta/2, m_{1P} + m_{2P} - k + \theta/2)}{B(\theta/2, \theta/2)}$$

where B is the beta function. The P value is calculated as

$$P = 2 \cdot \min\left[\sum_{k=0}^{m_{1P}} p(m_1 = k), \sum_{k=m_{1P}}^{m_{1P}+m_{2P}} p(m_1 = k)\right] - p(m_1 = m_{1P})$$

Detection of aberrant splice junctions in target genes

Splice junctions in target genes were identified from TEQUILA-seq reads, filtered for those with at least one annotated splice site (relative to GENCODE v45 annotations), and annotated with the following: (i) genomic coordinates, (ii) target gene in which the junction was observed, (iii) read support for the junction, and (iv) total read coverage around the corresponding splice sites. Junctions were then further filtered for those supported by ≥20 reads. For each junction, we fitted a beta distribution to usage frequencies observed among tissue-matched GTEx short-read RNA-seq samples (minimum of 100 samples). Suppose our fitted beta distribution is parameterized by μ (mean) and θ (precision) such that for some random variable $z \sim \text{beta}(\mu, \theta)$, the probability density function is given by

$$f(z; \mu, \theta) = \frac{z^{\mu\theta - 1}(1 - z)^{(1-\mu)\theta - 1}}{B[\mu\theta, (1 - \mu)\theta]}$$

For some individual P, let y_P be the number of reads supporting a given junction, and let N_P be the number of reads covering the corresponding splice sites. Suppose y_P is drawn from a beta-binomial distribution with N_P trials and parameters μ_P (individual-specific mean) and θ (precision). Under the null hypothesis, we assume that $\mu_P = \mu$ (i.e., the mean usage frequency of a given splice junction in individual P is identical to that of the general population). Thus, the probability of observing $y = k$ reads in individual P for some $k \in \{0, \ldots, N_P\}$ is given by

$$p(y = k) = \binom{N_P}{k} \cdot \frac{B[k + \mu\theta, N_P - k + (1 - \mu)\theta]}{B[\mu\theta, (1 - \mu)\theta]}$$

The P value is calculated as

$$P = 2 \cdot \min\left[\sum_{k=0}^{y_P} p(y = k), \sum_{k=y_P}^{N_P} p(y = k)\right] - p(y = y_P)$$

Statistical analysis and code availability

Gene expression dosage imbalance calls were pooled across all cohort individuals and filtered for significance [false discovery rate (FDR) < 1%]. We subsequently filtered out target genes that were called as gene expression dosage outliers in more than three cohort individuals. On the other hand, outlier splice junction calls were pooled across all cohort individuals and filtered for significance (FDR < 1%) and the following sequential criteria: (i) usage frequency for a junction is >10% higher than the mean usage frequency observed among tissue-matched GTEx controls, and (ii) junction is called an outlier in no more than three cohort individuals. Each

aberrant splice junction was visualized in a sashimi plot using ggsashimi (v1.1.5). Scripts for analyzing data and generating figures in the manuscript are available on GitHub (https://github.com/Xinglab/STRIPE) and archived at Zenodo (https://doi.org/10.5281/zenodo.17538307).

Supplementary Materials

The PDF file includes:
Figs. S1 to S16
Legends for tables S1 to S12

Other Supplementary Material for this manuscript includes the following:
Tables S1 to S12

REFERENCES

1. E. Turro, W. J. Astle, K. Megy, S. Graf, D. Greene, O. Shamardina, H. L. Allen, A. Sanchis-Juan, M. Frontini, C. Thys, J. Stephens, R. Mapeta, O. S. Burren, K. Downes, M. Haimel, S. Tuna, S. V. V. Deevi, T. J. Aitman, D. L. Bennett, P. Calleja, K. Carss, M. J. Caulfield, P. F. Chinnery, P. H. Dixon, D. P. Gale, R. James, A. Koziell, M. A. Laffan, A. P. Levine, E. R. Maher, H. S. Markus, J. Morales, N. W. Morrell, A. D. Mumford, E. Ormondroyd, S. Rankin, A. Rendon, S. Richardson, I. Roberts, N. B. A. Roy, M. A. Saleem, K. G. C. Smith, H. Stark, R. Y. Y. Tan, A. C. Themistocleous, A. J. Thrasher, H. Watkins, A. R. Webster, M. R. Wilkins, C. Williamson, J. Whitworth, S. Humphray, D. R. Bentley, NIHR BioResource for the 100,000 Genomes Project, N. Kingston, N. Walker, J. R. Bradley, S. Ashford, C. J. Penkett, K. Freson, K. E. Stirrups, F. L. Raymond, W. H. Ouwehand, Whole-genome sequencing of patients with rare diseases in a national health system. *Nature* **583**, 96–102 (2020).
2. S. M. Baxter, J. E. Posey, N. J. Lake, N. Sobreira, J. X. Chong, S. Buyske, E. E. Blue, L. H. Chadwick, Z. H. Coban-Akdemir, K. F. Doheny, C. P. Davis, M. Lek, C. Wellington, S. N. Jhangiani, M. Gerstein, R. A. Gibbs, R. P. Lifton, D. G. MacArthur, T. C. Matise, J. R. Lupski, D. Valle, M. J. Bamshad, A. Hamosh, S. Mane, D. A. Nickerson, Centers for Mendelian Genomics Consortium, H. L. Rehm, A. O'Donnell-Luria, Centers for Mendelian Genomics: A decade of facilitating gene discovery. *Genet. Med.* **24**, 784–797 (2022).
3. C. C. Y. Chung, S. P. Y. Hue, N. Y. T. Ng, P. H. L. Doong, Hong Kong Genome Project, A. T. W. Chu, B. H. Y. Chung, Meta-analysis of the diagnostic and clinical utility of exome and genome sequencing in pediatric and adult patients with rare diseases across diverse populations. *Genet. Med.* **25**, 100896 (2023).
4. M. H. Wojcik, G. Lemire, E. Berger, M. S. Zaki, M. Wissmann, W. Win, S. M. White, B. Weisburd, D. Wieczorek, L. B. Waddell, J. M. Verboon, G. E. VanNoy, A. Topf, T. Y. Tan, S. Syrbe, V. Strehlow, V. Straub, S. L. Stenton, H. Snow, M. Singer-Berk, J. Silver, S. Shril, E. G. Seaby, R. Schneider, V. G. Sankaran, A. Sanchis-Juan, K. A. Russell, K. Reinson, G. Ravenscroft, M. Radtke, D. Popp, T. Polster, K. Platzer, E. A. Pierce, E. M. Place, S. Pajusalu, L. Pais, K. Ounap, I. Osei-Owusu, H. Opperman, V. Okur, K. T. Oja, M. O'Leary, E. O'Heir, C. F. Morel, A. Merkenschlager, R. G. Marchant, B. E. Mangilog, J. A. Madden, D. MacArthur, A. Lovgren, J. P. Lerner-Ellis, J. Lin, N. Laing, F. Hildebrandt, J. Hentschel, E. Groopman, J. Goodrich, J. G. Gleeson, R. Ghaoui, C. A. Genetti, J. Gburek-Augustat, H. T. Gazda, V. S. Ganesh, M. Ganapathi, L. Gallacher, J. M. Fu, E. Evangelista, E. England, S. Donkervoort, S. DiTroia, S. T. Cooper, W. K. Chung, J. Christodoulou, K. R. Chao, L. D. Cato, K. M. Bujakowska, S. J. Bryen, H. Brand, C. G. Bonnemann, A. H. Beggs, S. M. Baxter, T. Bartolomaeus, P. B. Agrawal, M. Talkowski, C. Austin-Tse, R. A. Jamra, H. L. Rehm, A. O'Donnell-Luria, Genome sequencing for diagnosing rare diseases. *N. Engl. J. Med.* **390**, 1985–1997 (2024).
5. B. B. Cummings, J. L. Marshall, T. Tukiainen, M. Lek, S. Donkervoort, A. R. Foley, V. Bolduc, L. B. Waddell, S. A. Sandaradura, G. L. O'Grady, E. Estrella, H. M. Reddy, F. Zhao, B. Weisburd, K. J. Karczewski, A. H. O'Donnell-Luria, D. Birnbaum, A. Sarkozy, Y. Hu, H. Gonorazky, K. Claeys, H. Joshi, A. Bournazos, E. C. Oates, R. Ghaoui, M. R. Davis, N. G. Laing, A. Topf, C. Genotype-Tissue Expression, P. B. Kang, A. H. Beggs, K. N. North, V. Straub, J. J. Dowling, F. Muntoni, N. F. Clarke, S. T. Cooper, C. G. Bonnemann, D. G. MacArthur, Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. *Sci. Transl. Med.* **9**, eaal5209 (2017).
6. L. S. Kremer, D. M. Bader, C. Mertes, R. Kopajtich, G. Pichler, A. Iuso, T. B. Haack, E. Graf, T. Schwarzmayr, C. Terrile, E. Konarikova, B. Repp, G. Kastenmuller, J. Adamski, P. Lichtner, C. Leonhardt, B. Funalot, A. Donati, V. Tiranti, A. Lombes, C. Jardel, D. Glaser, R. W. Taylor, D. Ghezzi, J. A. Mayr, A. Rotig, P. Freisinger, F. Distelmaier, T. M. Strom, T. Meitinger, J. Gagneur, H. Prokisch, Genetic diagnosis of Mendelian disorders via RNA sequencing. *Nat. Commun.* **8**, 15824 (2017).
7. H. D. Gonorazky, S. Naumenko, A. K. Ramani, V. Nelakuditi, P. Mashouri, P. Wang, D. Kao, K. Ohri, S. Viththiyapaskaran, M. A. Tarnopolsky, K. D. Mathews, S. A. Moore, A. N. Osorio, D. Villanova, D. U. Kemaladewi, R. D. Cohn, M. Brudno, J. J. Dowling, Expanding the boundaries of RNA sequencing as a diagnostic tool for rare Mendelian disease. *Am. J. Hum. Genet.* **104**, 466–483 (2019).
8. L. Frésard, C. Smail, N. M. Ferraro, N. A. Teran, X. Li, K. S. Smith, D. Bonner, K. D. Kernohan, S. Marwaha, Z. Zappala, B. Balliu, J. R. Davis, B. Liu, C. J. Prybol, J. N. Kohler, D. B. Zastrow, C. M. Reuter, D. G. Fisk, M. E. Grove, J. M. Davidson, T. Hartley, R. Joshi, B. J. Strober, S. Utiramerur, N. Undiagnosed Diseases, Care4Rare Canada Consortium, L. Lind, E. Ingelsson, A. Battle, G. Bejerano, J. A. Bernstein, E. A. Ashley, K. M. Boycott, J. D. Merker, M. T. Wheeler, S. B. Montgomery, Identification of rare-disease genes using blood transcriptome sequencing and large control cohorts. *Nat. Med.* **25**, 911–919 (2019).
9. H. Lee, A. Y. Huang, L.-K. Wang, A. J. Yoon, G. Renteria, A. Eskin, R. H. Signer, N. Dorrani, S. Nieves-Rodriguez, J. Wan, E. D. Douine, J. D. Woods, E. C. Dell'Angelica, B. L. Fogel, M. G. Martin, M. J. Butte, N. H. Parker, R. T. Wang, P. B. Shieh, D. A. Wong, N. Gallant, K. E. Singh, Y. J. Tavyev Asher, J. S. Sinsheimer, D. Krakow, S. K. Loo, P. Allard, J. C. Papp, N. Undiagnosed Diseases, C. G. S. Palmer, J. A. Martinez-Agosto, S. F. Nelson, Diagnostic utility of transcriptome sequencing for rare Mendelian diseases. *Genet. Med.* **22**, 490–499 (2020).
10. S. Rentas, K. S. Rathi, M. Kaur, P. Raman, I. D. Krantz, M. Sarmady, A. A. Tayoun, Diagnosing Cornelia de Lange syndrome and related neurodevelopmental disorders using RNA sequencing. *Genet. Med.* **22**, 927–936 (2020).
11. D. R. Murdock, H. Dai, L. C. Burrage, J. A. Rosenfeld, S. Ketkar, M. F. Muller, V. A. Yepez, J. Gagneur, P. Liu, S. Chen, M. Jain, G. Zapata, C. A. Bacino, H. T. Chao, P. Moretti, W. J. Craigen, N. A. Hanchard, Undiagnosed Diseases Network, B. Lee, Transcriptome-directed analysis for Mendelian disease diagnosis overcomes limitations of conventional genomic testing. *J. Clin. Invest.* **131**, e141500 (2021).
12. V. A. Yepez, M. Gusic, R. Kopajtich, C. Mertes, N. H. Smith, C. L. Alston, R. Ban, S. Beblo, R. Berutti, H. Blessing, E. Ciara, F. Distelmaier, P. Freisinger, J. Haberle, S. J. Hayflick, M. Hempel, Y. S. Itkis, Y. Kishita, T. Klopstock, T. D. Krylova, C. Lamperti, D. Lenz, C. Makowski, S. Mosegaard, M. F. Muller, G. Munoz-Pujol, A. Nadel, A. Ohtake, Y. Okazaki, E. Procopio, T. Schwarzmayr, J. Smet, C. Staufner, S. L. Stenton, T. M. Strom, C. Terrile, F. Tort, R. Van Coster, A. Vanlander, M. Wagner, M. Xu, F. Fang, D. Ghezzi, J. A. Mayr, D. Piekutowska-Abramczuk, A. Ribes, A. Rotig, R. W. Taylor, S. B. Wortmann, K. Murayama, T. Meitinger, J. Gagneur, H. Prokisch, Clinical implementation of RNA sequencing for Mendelian disease diagnostics. *Genome Med.* **14**, 38 (2022).
13. G. A. Logsdon, M. R. Vollger, E. E. Eichler, Long-read human genome sequencing and its applications. *Nat. Rev. Genet.* **21**, 597–614 (2020).
14. D. Beyter, H. Ingimundardottir, A. Oddsson, H. P. Eggertsson, E. Bjornsson, H. Jonsson, B. A. Atlason, S. Kristmundsdottir, S. Mehringer, M. T. Hardarson, S. A. Gudjonsson, D. N. Magnusdottir, A. Jonasdottir, A. Jonasdottir, R. P. Kristjansson, S. T. Sverrisson, G. Holley, G. Palsson, O. A. Stefansson, G. Eyjolfsson, I. Olafsson, O. Sigurdardottir, B. Torfason, G. Masson, A. Helgason, U. Thorsteinsdottir, H. Holm, D. F. Gudbjartsson, P. Sulem, O. T. Magnusson, B. V. Halldorsson, K. Stefansson, 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).
15. S. Marwaha, J. W. Knowles, E. A. Ashley, A guide for the diagnosis of rare and undiagnosed disease: Beyond the exome. *Genome Med.* **14**, 23 (2022).
16. H. Tilgner, F. Grubert, D. Sharon, M. P. Snyder, Defining a personal, allele-specific, and single-molecule long-read transcriptome. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 9869–9874 (2014).
17. D. A. Glinos, G. Garborcauskas, P. Hoffman, N. Ehsan, L. Jiang, A. Gokden, X. Dai, F. Aguet, K. L. Brown, K. Garimella, T. Bowers, M. Costello, K. Ardlie, R. Jian, N. R. Tucker, P. T. Ellinor, E. D. Harrington, H. Tang, M. Snyder, S. Juul, P. Mohammadi, D. G. MacArthur, T. Lappalainen, B. B. Cummings, Transcriptome variation in human tissues revealed by long-read sequencing. *Nature* **608**, 353–359 (2022).
18. A. Dainis, E. Tseng, T. A. Clark, T. Hon, M. Wheeler, E. Ashley, Targeted long-read RNA sequencing demonstrates transcriptional diversity driven by splice-site variation in MYBPC3. *Circ. Genom. Precis. Med.* **12**, e002464 (2019).
19. F. Sedaghat-Hamedani, S. Rebs, E. Kayvanpour, C. Zhu, A. Amr, M. Muller, J. Haas, J. Wu, L. M. Steinmetz, P. Ehlermann, K. Streckfuss-Bomeke, N. Frey, B. Meder, Genotype complements the phenotype: Identification of the pathogenicity of an LMNA splice variant by nanopore long-read sequencing in a large DCM family. *Int. J. Mol. Sci.* **23**, 12230 (2022).
20. R. Wang, I. Helbig, A. C. Edmondson, L. Lin, Y. Xing, Splicing defects in rare diseases: Transcriptomics and machine learning strategies towards genetic diagnosis. *Brief. Bioinform.* **24**, bbad284 (2023).
21. F. Wang, Y. Xu, R. Wang, B. Zhang, N. Smith, A. Notaro, S. Gaerlan, E. Kutschera, K. E. Kadash-Edmondson, Y. Xing, L. Lin, TEQUILA-seq: A versatile and low-cost method for targeted long-read RNA sequencing. *Nat. Commun.* **14**, 4760 (2023).
22. GTEx Consortium, The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
23. S. B. Montgomery, T. Lappalainen, M. Gutierrez-Arcelus, E. T. Dermitzakis, Rare and common regulatory variation in population-scale sequenced human genomes. *PLOS Genet.* **7**, e1002144 (2011).

24. Y. Zeng, G. Wang, E. Yang, G. Ji, C. L. Brinkmeyer-Langford, J. J. Cai, Aberrant gene expression in humans. *PLOS Genet.* **11**, e1004942 (2015).
25. J. Zhao, I. Akinsanmi, D. Arafat, T. J. Cradick, C. M. Lee, S. Banskota, U. M. Marigorta, G. Bao, G. Gibson, A burden of rare variants associated with extremes of gene expression in human peripheral blood. *Am. J. Hum. Genet.* **98**, 299–309 (2016).
26. X. Li, Y. Kim, E. K. Tsang, J. R. Davis, F. N. Damani, C. Chiang, G. T. Hess, Z. Zappala, B. J. Strober, A. J. Scott, A. Li, A. Ganna, M. C. Bassik, J. D. Merker, GTEx Consortium, I. M. Hall, A. Battle, S. B. Montgomery, The impact of rare variation on gene expression across tissues. *Nature* **550**, 239–243 (2017).
27. N. M. Ferraro, B. J. Strober, J. Einson, N. S. Abell, F. Aguet, A. N. Barbeira, M. Brandt, M. Bucan, S. E. Castel, J. R. Davis, E. Greenwald, G. T. Hess, A. T. Hilliard, R. L. Kember, B. Kotis, Y. Park, G. Peloso, S. Ramdas, A. J. Scott, C. Smail, E. K. Tsang, S. M. Zekavat, M. Ziosi, Aradhana, TOPMed Lipids Working Group, K. G. Ardlie, T. L. Assimes, M. C. Bassik, C. D. Brown, A. Correa, I. Hall, H. K. Im, X. Li, P. Natarajan, GTEx Consortium, T. Lappalainen, P. Mohammadi, S. B. Montgomery, A. Battle, Transcriptomic signatures across human tissues identify functional rare genetic variation. *Science* **369**, eaaz5900 (2020).
28. H. Alsharhan, B. G. Ng, E. J. P. Daniel, J. Friedman, E. K. Pivnick, A. Al-Hashem, E. A. Faqeih, P. Liu, N. M. Engelhardt, K. N. Keller, J. Chen, P. A. Mazzeo, University of Washington Center for Mendelian Genomics (UW-CMG), J. A. Rosenfeld, M. J. Bamshad, D. A. Nickerson, K. M. Raymond, H. H. Freeze, M. He, A. C. Edmondson, C. Lam, Expanding the phenotype, genotype and biochemical knowledge of ALG3-CDG. *J. Inherit. Metab. Dis.* **44**, 987–1000 (2021).
29. J. Sidpra, S. Sudhakar, A. Biswas, F. Massey, V. Turchetti, T. Lau, E. Cook, J. R. Alvi, H. M. Elbendary, J. L. Jewell, A. Riva, A. Orsini, A. Vignoli, Z. Federico, J. Rosenblum, A. S. Schoonjans, M. de Wachter, I. Delgado Alvarez, A. Felipe-Rucian, N. A. Haridy, S. Haider, M. Zaman, S. Banu, N. Anwaar, F. Rahman, S. Maqbool, R. Yadav, V. Salpietro, R. Maroofian, R. Patel, R. Radhakrishnan, S. P. Prabhu, K. Lichtenbelt, H. Stewart, Y. Murakami, U. Lobel, F. D'Arco, E. Wakeling, W. Jones, E. Hay, S. Bhate, T. S. Jacques, D. M. Mirsky, M. T. Whitehead, M. S. Zaki, T. Sultan, P. Striano, A. C. Jansen, M. Lequin, L. S. de Vries, M. Severino, A. C. Edmondson, L. Menzies, P. M. Campeau, H. Houlden, A. McTague, S. Efthymiou, K. Mankad, The clinical and genetic spectrum of inherited glycosylphosphatidylinositol deficiency disorders. *Brain* **147**, 2775–2790 (2024).
30. B. G. Ng, P. Sosicka, S. Agadi, M. Almannai, C. A. Bacino, R. Barone, L. D. Botto, J. E. Burton, C. Carlston, B. H. Chung, J. S. Cohen, D. Coman, K. M. Dipple, N. Dorrani, W. B. Dobyns, A. F. Elias, L. Epstein, W. A. Gahl, D. Garozzo, T. B. Hammer, J. Haven, D. Heron, M. Herzog, G. E. Hoganson, J. M. Hunter, M. Jain, J. Juusola, S. Lakhani, H. Lee, J. Lee, K. Lewis, N. Longo, C. M. Lourenco, C. C. Y. Mak, D. McKnight, B. A. Mendelsohn, C. Mignot, G. Mirzaa, W. Mitchell, H. Muhle, S. F. Nelson, M. Olczak, C. G. S. Palmer, A. Partikian, M. C. Patterson, T. M. Pierson, S. C. Quinonez, B. M. Regan, M. E. Ross, M. J. G. Sacoto, F. Scaglia, I. E. Scheffer, D. Segal, N. S. Singhal, P. Striano, L. Sturiale, J. D. Symonds, S. Tang, E. Vilain, M. Willis, L. A. Wolfe, H. Yang, S. Yano, Z. Powis, S. F. Suchy, J. A. Rosenfeld, A. C. Edmondson, S. Grunewald, H. H. Freeze, SLC35A2-CDG: Functional characterization, expanded molecular, clinical, and biochemical phenotypes of 30 unreported individuals. *Hum. Mutat.* **40**, 908–925 (2019).
31. E. Barca, R. D. Ganetzky, P. Potluri, M. Juanola-Falgarona, X. Gai, D. Li, C. Jalas, Y. Hirsch, V. Emmanuele, S. Tadesse, M. Ziosi, H. O. Akman, W. K. Chung, K. Tanji, E. M. McCormick, E. Place, M. Consugar, E. A. Pierce, H. Hakonarson, D. C. Wallace, M. Hirano, M. J. Falk, USMG5 Ashkenazi Jewish founder mutation impairs mitochondrial complex V dimerization and ATP synthesis. *Hum. Mol. Genet.* **27**, 3305–3312 (2018).
32. M. Ganapathi, G. Friocourt, N. Gueguen, M. W. Friederich, G. Le Gac, V. Okur, N. Loaec, T. Ludwig, C. Ka, K. Tanji, P. Marcorelles, E. Theodorou, A. Lignelli-Dipple, C. Voisset, M. A. Walker, L. C. Briere, A. Bourhis, M. Blondel, C. LeDuc, J. Hagen, C. Cooper, C. Muraresku, C. Ferec, A. Garenne, S. Lelez-Soquet, C. A. Rogers, Y. Shen, D. K. Strode, P. Bizargity, A. Iglesias, A. Goldstein, F. A. High, U. D. Network, D. A. Sweetser, R. Ganetzky, J. L. K. Van Hove, V. Procaccio, C. Le Marechal, W. K. Chung, A homozygous splice variant in ATP5PO, disrupts mitochondrial complex V function and causes Leigh syndrome in two unrelated families. *J. Inherit. Metab. Dis.* **45**, 996–1012 (2022).
33. H. Dai, V. W. Zhang, A. W. El-Hattab, C. Ficicioglu, M. Shinawi, M. Lines, A. Schulze, M. McNutt, G. Gotway, X. Tian, S. Chen, J. Wang, W. J. Craigen, L. J. Wong, FBXL4 defects are common in patients with congenital lactic acidemia and encephalomyopathic mitochondrial DNA depletion syndrome. *Clin. Genet.* **91**, 634–639 (2017).
34. N. J. Lake, B. D. Webb, D. A. Stroud, T. R. Richman, B. Ruzzenente, A. G. Compton, H. S. Mountford, J. Pulman, C. Zangarelli, M. Rio, N. Boddaert, Z. Assouline, M. D. Sherpa, E. E. Schadt, S. M. Houten, J. Byrnes, E. M. McCormick, Z. Zolkipli-Cunningham, K. Haude, Z. Zhang, K. Retterer, R. Bai, S. E. Calvo, V. K. Mootha, J. Christodoulou, A. Rotig, A. Filipovska, I. Cristian, M. J. Falk, M. D. Metodiev, D. R. Thorburn, Biallelic mutations in MRPS34 lead to instability of the small mitoribosomal subunit and Leigh syndrome. *Am. J. Hum. Genet.* **101**, 239–254 (2017).
35. C. N. Murali, C. Soler-Alfonso, K. M. Loomes, A. A. Shah, D. Monteil, C. D. Padilla, F. Scaglia, R. Ganetzky, TRMU deficiency: A broad clinical spectrum responsive to cysteine supplementation. *Mol. Genet. Metab.* **132**, 146–153 (2021).
36. G. S. Gorman, P. F. Chinnery, S. DiMauro, M. Hirano, Y. Koga, R. McFarland, A. Suomalainen, D. R. Thorburn, M. Zeviani, D. M. Turnbull, Mitochondrial diseases. *Nat. Rev. Dis. Primers* **2**, 16080 (2016).
37. D. J. Lefeber, H. H. Freeze, R. Steet, T. Kinoshita, "Congenital disorders of glycosylation," in *Essentials of Glycobiology*, A. Varki, R. D. Cummings, J. D. Esko, P. Stanley, G. W. Hart, M. Aebi, D. Mohnen, T. Kinoshita, N. H. Packer, J. H. Prestegard, R. L. Schnaar, P. H. Seeberger, Eds. (Cold Spring Harbor Laboratory Press, 2022), pp. 599–614.
38. P. Mohammadi, S. E. Castel, B. B. Cummings, J. Einson, C. Sousa, P. Hoffman, S. Donkervoort, Z. Jiang, P. Mohassel, A. R. Foley, H. E. Wheeler, H. K. Im, C. G. Bonnemann, D. G. MacArthur, T. Lappalainen, Genetic regulatory variation in populations informs transcriptome analysis in rare disease. *Science* **366**, 351–356 (2019).
39. M. A. Rivas, M. Pirinen, D. F. Conrad, M. Lek, E. K. Tsang, K. J. Karczewski, J. B. Maller, K. R. Kukurba, D. S. DeLuca, M. Fromer, P. G. Ferreira, K. S. Smith, R. Zhang, F. Zhao, E. Banks, R. Poplin, D. M. Ruderfer, S. M. Purcell, T. Tukiainen, E. V. Minikel, P. D. Stenson, D. N. Cooper, K. H. Huang, T. J. Sullivan, J. Nedzel, GTEx Consortium, Geuvadis Consortium, C. D. Bustamante, J. B. Li, M. J. Daly, R. Guigo, P. Donnelly, K. Ardlie, M. Sammeth, E. T. Dermitzakis, M. I. McCarthy, S. B. Montgomery, T. Lappalainen, D. G. MacArthur, Human genomics. Effect of predicted protein-truncating genetic variants on the human transcriptome. *Science* **348**, 666–669 (2015).
40. G. Yeo, C. B. Burge, Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J. Comput. Biol.* **11**, 377–394 (2004).
41. S. Vuillaumier-Barrot, A. Barnier, M. Cuer, G. Durand, B. Grandchamp, N. Seta, Characterization of the 415G>A (E139K) PMM2 mutation in carbohydrate-deficient glycoprotein syndrome type Ia disrupting a splicing enhancer resulting in exon 5 skipping. *Hum. Mutat.* **14**, 543–544 (1999).
42. K. Jaganathan, S. Kyriazopoulou Panagiotopoulou, J. F. McRae, S. F. Darbandi, D. Knowles, Y. I. Li, J. A. Kosmicki, J. Arbelaez, W. Cui, G. B. Schwartz, E. D. Chow, E. Kanterakis, H. Gao, A. Kia, S. Batzoglou, S. J. Sanders, K. K. Farh, Predicting splicing from primary sequence with deep learning. *Cell* **176**, 535–548.e24 (2019).
43. S. E. Calvo, E. J. Tucker, A. G. Compton, D. M. Kirby, G. Crawford, N. P. Burtt, M. Rivas, C. Guiducci, D. L. Bruno, O. A. Goldberger, M. C. Redman, E. Wiltshire, C. J. Wilson, D. Altshuler, S. B. Gabriel, M. J. Daly, D. R. Thorburn, V. K. Mootha, High-throughput, pooled sequencing identifies mutations in NUBPL and FOXRED1 in human complex I deficiency. *Nat. Genet.* **42**, 851–858 (2010).
44. E. J. Tucker, M. Mimaki, A. G. Compton, M. McKenzie, M. T. Ryan, D. R. Thorburn, Next-generation sequencing in molecular diagnosis: NUBPL mutations highlight the challenges of variant detection and interpretation. *Hum. Mutat.* **33**, 411–418 (2012).
45. S. H. Kevelam, R. J. Rodenburg, N. I. Wolf, P. Ferreira, R. J. Lunsing, L. G. Nijtmans, A. Mitchell, H. A. Arroyo, D. Rating, A. Vanderver, C. G. van Berkel, T. E. Abbink, P. Heutink, M. S. van der Knaap, NUBPL mutations in patients with complex I deficiency and a distinct MRI pattern. *Neurology* **80**, 1577–1583 (2013).
46. R. Dawes, A. M. Bournazos, S. J. Bryen, S. Bommireddipalli, R. G. Marchant, H. Joshi, S. T. Cooper, SpliceVault predicts the precise nature of variant-associated mis-splicing. *Nat. Genet.* **55**, 324–332 (2023).
47. M. E. Rogalska, C. Vivori, J. Valcarcel, Regulation of pre-mRNA splicing: Roles in physiology and disease, and therapeutic prospects. *Nat. Rev. Genet.* **24**, 251–269 (2023).
48. D. Kaida, M. G. Berg, I. Younis, M. Kasim, L. N. Singh, L. Wan, G. Dreyfuss, U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. *Nature* **468**, 664–668 (2010).
49. B. R. So, C. Di, Z. Cai, C. C. Venters, J. Guo, J. M. Oh, C. Arai, G. Dreyfuss, A complex of U1 snRNP with cleavage and polyadenylation factors controls telescripting, regulating mRNA transcription in human cells. *Mol. Cell* **76**, 590–599.e4 (2019).
50. J. Marquez, J. N. Cech, C. R. Paschal, B. Dingmann, A. I. Scott, J. M. Thies, M. R. Mills, C. M. Albert, A. E. Beck, E. Beckman, E. S. Bonkowski, D. L. Earl, C. T. Lam, H. C. Mefford, J. L. Merritt II, Z. Nelson, T. J. Ohlsen, M. R. Taylor, S. J. Perlman, E. R. Rudzinski, M. C. Sikes, N. Waligorski, T. L. Wenger, M. P. Adam, G. M. Mirzaa, J. T. Bennett, I. A. Glass, D. L. Sternen, D. E. Miller, Clinical RNA sequencing clarifies variants of uncertain significance identified by prior testing. *Genet. Med. Open* **2**, 101886 (2024).
51. X. Zhou, G. Zhao, J. J. Truglio, L. Wang, G. Li, W. J. Lennarz, H. Schindelin, Structural and biochemical studies of the C-terminal domain of mouse peptide-*N*-glycanase identify it as a mannose-binding module. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 17214–17219 (2006).
52. E. J. Tucker, S. G. Hershman, C. Kohrer, C. A. Belcher-Timme, J. Patel, O. A. Goldberger, J. Christodoulou, J. M. Silberstein, M. McKenzie, M. T. Ryan, A. G. Compton, J. D. Jaffe, S. A. Carr, S. E. Calvo, U. L. RajBhandary, D. R. Thorburn, V. K. Mootha, Mutations in MTFMT underlie a human disorder of formylation causing impaired mitochondrial translation. *Cell Metab.* **14**, 428–434 (2011).
53. H. Hayhurst, I. F. M. de Coo, D. Piekutowska-Abramczuk, C. L. Alston, S. Sharma, K. Thompson, R. Rius, L. He, S. Hopton, R. Ploski, E. Ciara, N. J. Lake, A. G. Compton, M. B. Delatycki, A. Verrips, P. E. Bonnen, S. A. Jones, A. A. Morris, D. Shakespeare, J. Christodoulou, D. Wesol-Kucharska, D. Rokicki, H. J. M. Smeets, E. Pronicka, D. R. Thorburn, G. S. Gorman, R. McFarland, R. W. Taylor, Y. S. Ng, Leigh syndrome caused

by mutations in MTFMT is associated with a better prognosis. *Ann. Clin. Transl. Neurol.* **6**, 515–524 (2019).
54. R. Wertenbroek, R. J. Hofmeister, I. Xenarios, Y. Thoma, O. Delaneau, Improving population scale statistical phasing with whole-genome sequencing data. *PLOS Genet.* **20**, e1011092 (2024).
55. Q. Sun, Y. Li, Advances in haplotype phasing and genotype imputation. *Nat. Rev. Genet.* **27**, 155–169 (2026).
56. M. J. P. Chaisson, A. D. Sanders, X. Zhao, A. Malhotra, D. Porubsky, T. Rausch, E. J. Gardner, O. L. Rodriguez, L. Guo, R. L. Collins, X. Fan, J. Wen, R. E. Handsaker, S. Fairley, Z. N. Kronenberg, X. Kong, F. Hormozdiari, D. Lee, A. M. Wenger, A. R. Hastie, D. Antaki, T. Anantharaman, P. A. Audano, H. Brand, S. Cantsilieris, H. Cao, E. Cerveira, C. Chen, X. Chen, C. S. Chin, Z. Chong, N. T. Chuang, C. C. Lambert, D. M. Church, L. Clarke, A. Farrell, J. Flores, T. Galeev, D. U. Gorkin, M. Gujral, V. Guryev, W. H. Heaton, J. Korlach, S. Kumar, J. Y. Kwon, E. T. Lam, J. E. Lee, J. Lee, W. P. Lee, S. P. Lee, S. Li, P. Marks, K. Viaud-Martinez, S. Meiers, K. M. Munson, F. C. P. Navarro, B. J. Nelson, C. Nodzak, A. Noor, S. Kyriazopoulou-Panagiotopoulou, A. W. C. Pang, Y. Qiu, G. Rosanio, M. Ryan, A. Stutz, D. C. J. Spierings, A. Ward, A. E. Welch, M. Xiao, W. Xu, C. Zhang, Q. Zhu, X. Zheng-Bradley, E. Lowy, S. Yakneen, S. McCarroll, G. Jun, L. Ding, C. L. Koh, B. Ren, P. Flicek, K. Chen, M. B. Gerstein, P. Y. Kwok, P. M. Lansdorp, G. T. Marth, J. Sebat, X. Shi, A. Bashir, K. Ye, S. E. Devine, M. E. Talkowski, R. E. Mills, T. Marschall, J. O. Korbel, E. E. Eichler, C. Lee, Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat. Commun.* **10**, 1784 (2019).
57. E. Kucuk, B. van der Sanden, L. O'Gorman, M. Kwint, R. Derks, A. M. Wenger, C. Lambert, S. Chakraborty, P. Baybayan, W. J. Rowell, H. G. Brunner, L. Vissers, A. Hoischen, C. Gilissen, Comprehensive de novo mutation discovery with HiFi long-read sequencing. *Genome Med.* **15**, 34 (2023).
58. L. Demirdjian, Y. Xu, E. Bahrami-Samani, Y. Pan, S. Stein, Z. Xie, E. Park, Y. N. Wu, Y. Xing, Detecting allele-specific alternative splicing from population-scale RNA-seq data. *Am. J. Hum. Genet.* **107**, 461–472 (2020).
59. K. Amoah, Y. E. Hsiao, J. H. Bahn, Y. Sun, C. Burghard, B. X. Tan, E. W. Yang, X. Xiao, Allele-specific alternative splicing and its functional genetic variants in human tissues. *Genome Res.* **31**, 359–371 (2021).
60. C. DeBoever, Y. Tanigawa, M. E. Lindholm, G. McInnes, A. Lavertu, E. Ingelsson, C. Chang, E. A. Ashley, C. D. Bustamante, M. J. Daly, M. A. Rivas, Medical relevance of protein-truncating variants across 337,205 individuals in the UK Biobank study. *Nat. Commun.* **9**, 1612 (2018).
61. M. R. Vollger, J. Korlach, K. C. Eldred, E. Swanson, J. G. Underwood, S. C. Bohaczuk, Y. Mao, Y. H. Cheng, J. Ranchalis, E. E. Blue, U. Schwarze, K. M. Munson, C. T. Saunders, A. M. Wenger, A. Allworth, S. Chanprasert, B. L. Duerden, I. Glass, M. Horike-Pyne, M. Kim, K. A. Leppig, I. J. McLaughlin, J. Ogawa, E. A. Rosenthal, S. Sheppeard, S. M. Sherman, S. Strohbehn, A. L. Yuen, A. W. Stacey, University of Washington Center for Rare Disease Research, Undiagnosed Diseases Network, T. A. Reh, P. H. Byers, M. J. Bamshad, F. M. Hisama, G. P. Jarvik, Y. Sancak, K. M. Dipple, A. B. Stergachis, Synchronized long-read genome, methylome, epigenome and transcriptome profiling resolve a Mendelian condition. *Nat. Genet.* **57**, 469–479 (2025).
62. J. Kim, S. Woo, C. M. de Gusmao, B. Zhao, D. H. Chin, R. L. DiDonato, M. A. Nguyen, T. Nakayama, C. A. Hu, A. Soucy, A. Kuniholm, J. K. Thornton, O. Riccardi, D. A. Friedman, C. M. El Achkar, Z. Dash, L. Cornelissen, C. Donado, K. N. W. Faour, L. W. Bush, V. Suslovitch, C. Lentucci, P. J. Park, E. A. Lee, A. Patterson, A. A. Philippakis, B. Margus, C. B. Berde, T. W. Yu, A framework for individualized splice-switching oligonucleotide therapy. *Nature* **619**, 828–836 (2023).
63. M. M. Scotti, M. S. Swanson, RNA mis-splicing in disease. *Nat. Rev. Genet.* **17**, 19–32 (2016).
64. I. H. Ament, N. DeBruyne, F. Wang, L. Lin, Long-read RNA sequencing: A transformative technology for exploring transcriptome complexity in human diseases. *Mol. Ther.* **33**, 883–894 (2025).
65. S. Li, S. Zhao, J. C. Sinson, A. Bajic, J. A. Rosenfeld, M. B. Neeley, M. Pena, K. C. Worley, L. C. Burrage, M. Weisz-Hubshman, S. Ketkar, W. J. Craigen, G. D. Clark, S. Lalani, C. A. Bacino, K. Machol, H. T. Chao, L. Potocki, L. Emrick, J. Sheppard, M. T. T. Nguyen, A. Khoramnia, P. P. Hernandez, S. C. Nagamani, Z. Liu, N. Undiagnosed Diseases, C. M. Eng, B. Lee, P. Liu, The clinical utility and diagnostic implementation of human subject cell transdifferentiation followed by RNA sequencing. *Am. J. Hum. Genet.* **111**, 841–862 (2024).
66. I. Yamada, M. Shiota, D. Shinmachi, T. Ono, S. Tsuchiya, M. Hosoda, A. Fujita, N. P. Aoki, Y. Watanabe, N. Fujita, K. Angata, H. Kaji, H. Narimatsu, S. Okuda, K. F. Aoki-Kinoshita, The GlyCosmos Portal: A unified and comprehensive web resource for the glycosciences. *Nat. Methods* **17**, 649–650 (2020).
67. L. Hansen, D. M. Husein, B. Gericke, T. Hansen, O. Pedersen, M. A. Tambe, H. H. Freeze, H. Y. Naim, B. Henrissat, H. H. Wandall, H. Clausen, E. P. Bennett, A mutation map for human glycoside hydrolase genes. *Glycobiology* **30**, 500–515 (2020).
68. A. V. Nairn, W. S. York, K. Harris, E. M. Hall, J. M. Pierce, K. W. Moremen, Regulation of glycan structures in animal tissues: Transcript profiling of glycan-related genes. *J. Biol. Chem.* **283**, 17298–17313 (2008).
69. K. T. Schjoldager, Y. Narimatsu, H. J. Joshi, H. Clausen, Global view of human protein glycosylation pathways and functions. *Nat. Rev. Mol. Cell Biol.* **21**, 729–749 (2020).
70. R. H. Hovhannisyan, R. P. Carstens, A novel intronic cis element, ISE/ISS-3, regulates rat fibroblast growth factor receptor 2 splicing through activation of an upstream exon and repression of a downstream exon containing a noncanonical branch point sequence. *Mol. Cell. Biol.* **25**, 250–263 (2005).
71. I. F. Scheller, K. Lutz, C. Mertes, V. A. Yepez, J. Gagneur, Improved detection of aberrant splicing with FRASER 2.0 and the intron Jaccard index. *Am. J. Hum. Genet.* **110**, 2056–2067 (2023).
72. D. E. Miller, A. Sulovari, T. Wang, H. Loucks, K. Hoekzema, K. M. Munson, A. P. Lewis, E. P. A. Fuerte, C. R. Paschal, T. Walsh, J. Thies, J. T. Bennett, I. Glass, K. M. Dipple, K. Patterson, E. S. Bonkowski, Z. Nelson, A. Squire, M. Sikes, E. Beckman, R. L. Bennett, D. Earl, W. Lee, R. Allikmets, S. J. Perlman, P. Chow, A. V. Hing, T. L. Wenger, M. P. Adam, A. Sun, C. Lam, I. Chang, X. Zou, S. L. Austin, E. Huggins, A. Safi, A. K. Iyengar, T. E. Reddy, W. H. Majoros, A. S. Allen, G. E. Crawford, P. S. Kishnani, University of Washington Center for Mendelian Genomics, M.-C. King, T. Cherry, J. X. Chong, M. J. Bamshad, D. A. Nickerson, H. C. Mefford, D. Doherty, E. E. Eichler, Targeted long-read sequencing identifies missing disease-causing variation. *Am. J. Hum. Genet.* **108**, 1436–1449 (2021).

Acknowledgments: We thank all participants and clinical staff for their support and contributions to this study. We would also like to acknowledge R. T. Starosta (University of Colorado School of Medicine), A. Larson (University of Colorado School of Medicine), and S. Madan-Khetarpal (Children's Hospital of Pittsburgh) for providing fibroblast samples from undiagnosed patients who remained undiagnosed at the end of the study. **Funding:** This work was funded by the National Institutes of Health grant R01DK099551 (B.G.N. and H.H.F.), the National Institutes of Health grant R01GM088342 (Y.X.), the National Institutes of Health grant R01GM121827 (L.L.), the National Institutes of Health grant R35GM134863 (M.J.F.), the National Institutes of Health grant R35GM151098 (R.D.G.), the National Institutes of Health grant R35GM158057 (L.L.), the National Institutes of Health grant R56HG012310 (Y.X.), the National Institutes of Health grant T32HG000046 (R.W.), the National Institutes of Health grant U54NS115198 (A.C.E., H.H.F., E.M., and C.T.L.), and the Rocket Fund (B.G.N. and H.H.F.). **Author contributions:** Conceptualization: Y.X., L.L., A.C.E., R.D.G., R.W., F.W., I.H., and E.M. Data curation: R.W., F.W., N.M.E., S.G., K.K., E.M., M.J.F., and A.C.E. Formal analysis: R.W., N.D., N.M.E., M.J.S., M.J.F., and R.D.G. Funding acquisition: Y.X., L.L., A.C.E., R.D.G., M.J.F., E.M., C.T.L., H.H.F., and I.H. Investigation: R.W., F.W., X.J., J.J.-H.P., A.N., S.G., R.P., E.M.M., C.T.L., R.D.G., and A.C.E. Methodology: R.W., F.W., R.D.G., A.C.E., L.L., and Y.X. Project administration: Y.X., L.L., A.C.E., R.D.G., M.J.F., H.H.F., R.W., F.W., and K.E.K.-E. Resources: Y.X., L.L., A.C.E., R.D.G., M.J.F., E.M., C.T.L., H.H.F., S.G., S.C., E.M.M., K.K., and B.G.N. Software: R.W. and N.D. Supervision: Y.X., L.L., A.C.E., R.D.G., M.J.F., E.M., and H.H.F. Validation: F.W., X.J., N.M.E., J.J.-H.P., A.N., S.G., B.G.N., E.M., M.J.F., R.D.G., and A.C.E. Visualization: R.W., X.J., and M.J.S. Writing—original draft: R.W., F.W., and Y.X. Writing—review and editing: R.W., Y.X., L.L., A.C.E., R.D.G., M.J.F., I.H., E.M., C.T.L., H.H.F., K.E.K.-E., B.G.N., E.M.M., M.J.S., N.M.E., and X.J. **Competing interests:** L.L., Y.X., and F.W. are inventors on a patent application filed by The Children's Hospital of Philadelphia covering the TEQUILA method (US 63/277,894). Y.X., L.L., F.W., and R.W. are inventors on a patent application filed by The Children's Hospital of Philadelphia covering the STRIPE method (US 64/009,481). M.J.F. has served as a sponsored research collaborator for Aadi Bioscience, Adjuvia Therapeutics, Astellas, Cyclerion Therapeutics, Epirium Bio, Imel Therapeutics, Khondrion, Merck, Minovia Therapeutics, Mission Therapeutics, NeuroVive Pharmaceutical AB, Precision Biosciences, PTC Therapeutics, Raptor Pharmaceuticals, REATA Inc., Reneo Therapeutics, RiboNova, Saol Therapeutics, Standigm, Stealth BioTherapeutics, and Thiogenesis. All other authors declare that they have no competing interests. **Data, code, and materials availability:** GTEx v8 datasets (gene TPMs, exon-exon junction read counts, and haplotype expression matrix) were downloaded from the GTEx portal (https://gtexportal.org/home/downloads/adult-gtex/). To protect subject confidentiality and patient privacy, individual-level sequencing data generated in this study are not publicly available but can be accessed through the Office of Collaborative and Corporate Research Contracts at the Children's Hospital of Philadelphia (researchcontracts@chop.edu), under appropriate data-use agreements and in compliance with institutional and ethical guidelines. The corresponding author (Y.X.; xingyi@chop.edu) should also be copied and will assist in coordinating requests. All data and code needed to evaluate and reproduce the results in the paper are present in the paper and/or the Supplementary Materials. This study did not generate new materials.

Submitted 14 May 2025
Accepted 3 March 2026
Published 15 April 2026
10.1126/sciadv.ady9895