JCI The Journal of Clinical Investigation

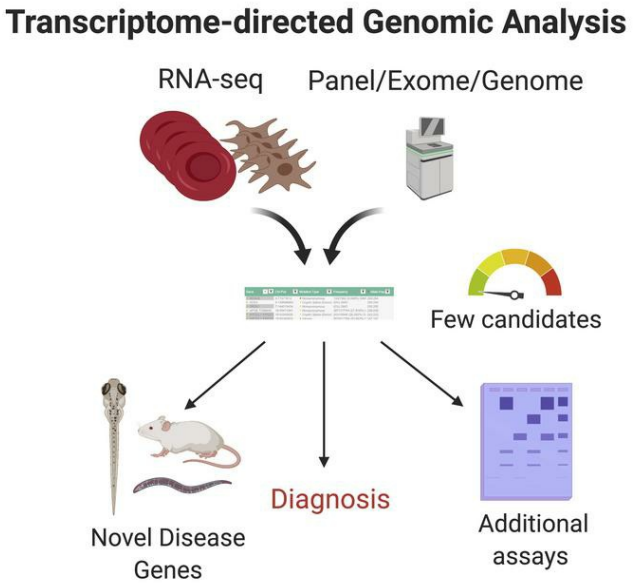
Transcriptome-directed analysis for Mendelian disease diagnosis overcomes limitations of conventional genomic testing

David R. Murdock, … , Undiagnosed Diseases Network, Brendan Lee

J Clin Invest. 2021;131(1):e141500. https://doi.org/10.1172/JCI141500.

Clinical Medicine Genetics

Graphical abstract



Transcriptome-directed analysis for Mendelian disease diagnosis overcomes limitations of conventional genomic testing

David R. Murdock,[1] Hongzheng Dai,[1,2] Lindsay C. Burrage,[1,3] Jill A. Rosenfeld,[1] Shamika Ketkar,[1] Michaela F. Müller,[4] Vicente A. Yépez,[4] Julien Gagneur,[4] Pengfei Liu,[1,2] Shan Chen,[1] Mahim Jain,[1] Gladys Zapata,[1,5] Carlos A. Bacino,[1,3] Hsiao-Tuan Chao,[1,3,6,7,8] Paolo Moretti,[1,9] William J. Craigen,[1,3] Neil A. Hanchard,[1,3,5] Undiagnosed Diseases Network,[10] and Brendan Lee[1,3]

[1]Department of Molecular and Human Genetics, Baylor College of Medicine (BCM), Houston, Texas, USA. [2]Baylor Genetics, Houston, Texas, USA. [3]Texas Children's Hospital, Houston, Texas, USA. [4]Department of Informatics, Technical University of Munich, Garching, Germany. [5]Laboratory for Translational Genomics, Agricultural Research Service (ARS)/United States Department of Agriculture (USDA) Children's Nutrition Research Center, and [6]Departments of Neuroscience and Pediatrics, Division of Neurology and Developmental Neuroscience, BCM, Houston, Texas, USA. [7]Jan and Dan Duncan Neurological Research Institute, Texas Children's Hospital, Houston, Texas, USA. [8]McNair Medical Institute at the Robert and Janice McNair Foundation, Houston, Texas, USA. [9]Department of Neurology, University of Utah and George E. Wahlen Department of Veterans Affairs Medical Center, Salt Lake City, Utah, USA. [10]Undiagnosed Diseases Network is detailed in Supplemental Acknowledgments.

BACKGROUND. Transcriptome sequencing (RNA-seq) improves diagnostic rates in individuals with suspected Mendelian conditions to varying degrees, primarily by directing the prioritization of candidate DNA variants identified on exome or genome sequencing (ES/GS). Here we implemented an RNA-seq–guided method to diagnose individuals across a wide range of ages and clinical phenotypes.

METHODS. One hundred fifteen undiagnosed adult and pediatric patients with diverse phenotypes and 67 family members (182 total individuals) underwent RNA-seq from whole blood and skin fibroblasts at the Baylor College of Medicine (BCM) Undiagnosed Diseases Network clinical site from 2014 to 2020. We implemented a workflow to detect outliers in gene expression and splicing for cases that remained undiagnosed despite standard genomic and transcriptomic analysis.

RESULTS. The transcriptome-directed approach resulted in a diagnostic rate of 12% across the entire cohort, or 17% after excluding cases solved on ES/GS alone. Newly diagnosed conditions included Koolen–de Vries syndrome (*KANSL1*), Renpenning syndrome (*PQBP1*), *TBCK*-associated encephalopathy, *NSD2*- and *CLTC*-related intellectual disability, and others, all with negative conventional genomic testing, including ES and chromosomal microarray (CMA). Skin fibroblasts exhibited higher and more consistent expression of clinically relevant genes than whole blood. In solved cases with RNA-seq from both tissues, the causative defect was missed in blood in half the cases but none from fibroblasts.

CONCLUSIONS. For our cohort of undiagnosed individuals with suspected Mendelian conditions, transcriptome-directed genomic analysis facilitated diagnoses, primarily through the identification of variants missed on ES and CMA.

TRIAL REGISTRATION. Not applicable.

FUNDING. NIH Common Fund, BCM Intellectual and Developmental Disabilities Research Center, Eunice Kennedy Shriver National Institute of Child Health & Human Development.

Introduction

Next-generation sequencing (NGS) has transformed diagnostic capabilities for suspected rare Mendelian disorders, mainly through the widespread adoption of exome sequencing (ES) in clinical practice. Nevertheless, many patients undergoing clinical ES remain undiagnosed, and studies generally report diagnostic rates of 25%–30% (1, 2). The Undiagnosed Diseases Network (UDN), funded by the NIH, is a multicenter effort that takes a multi-omic approach to solve these and other complex medical cases. The UDN has incorporated state-of-the-art testing, including ES and genome sequencing (GS), chromosomal microarray (CMA), metabolomics, model organism screening, and inpatient evaluations, with an overall diagnostic yield of 35% (3). Nevertheless, a significant number of cases remain unsolved, and the diagnostic odyssey for these patients and families continues.

Transcriptome sequencing (RNA-seq) is emerging as another tool in the genetic diagnostic toolbox, leading to a reported 7.5%–36% improvement in the diagnostic rate depending on the sampled tissue and clinical phenotype (4–8), and aiding in the prioritization and resolution of variants of uncertain significance (VUS) (9, 10).

Conflict of interest: The authors have declared that no conflict of interest exists.

Submitted: June 22, 2020; **Accepted:** September 24, 2020; **Published:** January 4, 2021.
Reference information: *J Clin Invest*. 2021;131(1):e141500.
https://doi.org/10.1172/JCI141500.

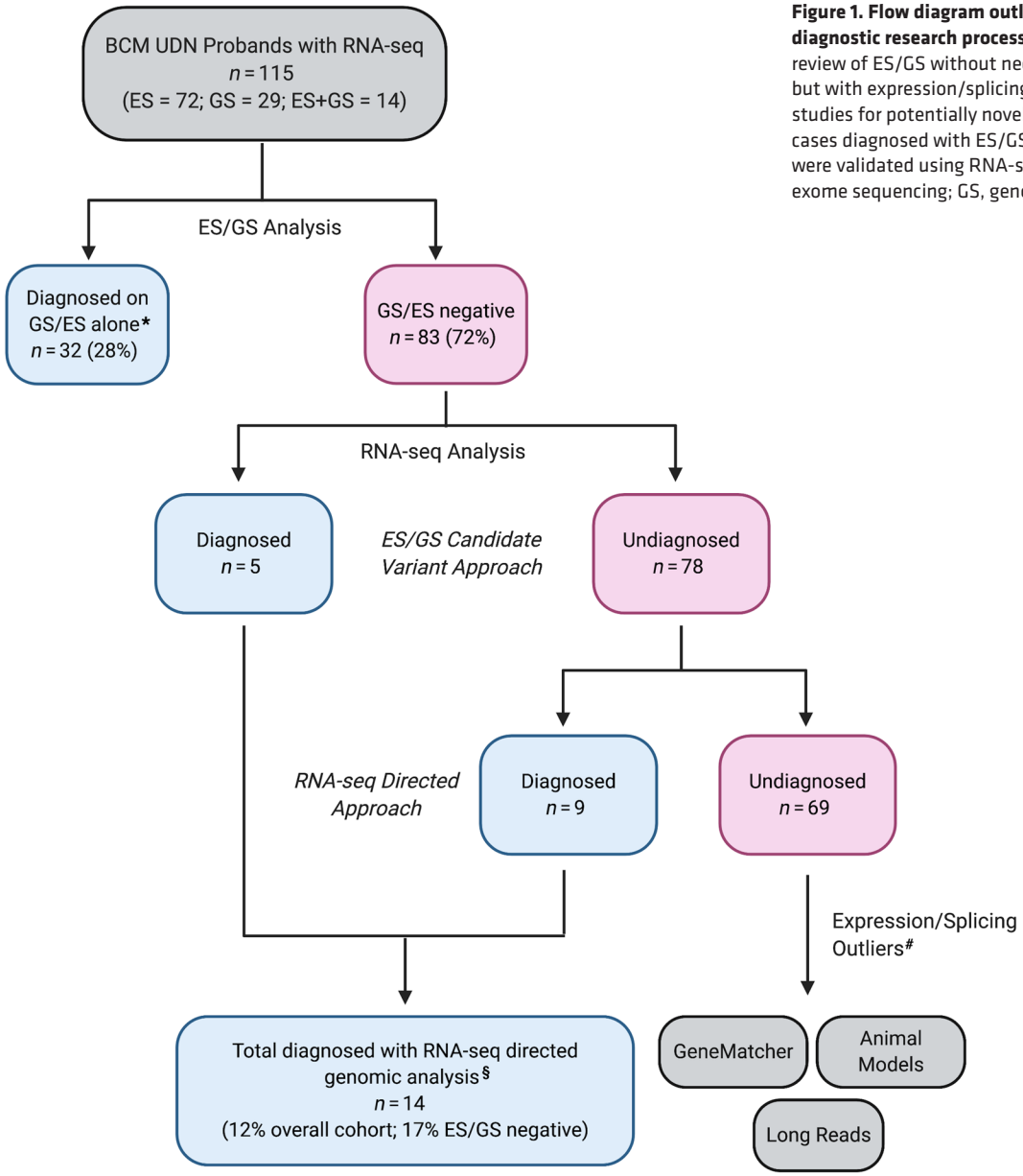


Figure 1. Flow diagram outlining BCM UDN RNA-seq diagnostic research process. *Cases diagnosed on initial review of ES/GS without needing RNA-seq. #Undiagnosed but with expression/splicing outliers prompting follow-up studies for potentially novel disease gene discovery. §Five cases diagnosed with ES/GS candidate variant approach were validated using RNA-seq–directed approach. ES, exome sequencing; GS, genome sequencing.

The approach to RNA-seq analysis varies but generally focuses on differences in splicing and the expression levels of genes. A traditional analytic approach relies on the time-consuming identification of candidate variants first in the ES/GS data that are then manually reviewed in the transcriptome to determine any functional consequences (8–10). While effective, there are several limitations to this strategy. First, as we transition from ES to GS, the number of potential candidate variants grows tremendously, increasing the time required to manually curate every possible effect in the transcriptome. Second, the selection of many candidate splicing variants depends on bioinformatic predictions that still have variable performance and accuracy (10, 11). Third, this approach requires the identification of candidates in the initial genomic analysis. For example, clinically relevant variants like single-exon deletions and repeat expansions are not detected on standard assays like ES or CMA. Last, because this approach prioritizes known disease genes, it is less suited for novel gene discovery.

Here, we describe our approach using RNA-seq first to direct downstream genomic analysis and diagnose patients with rare, Mendelian conditions. We implemented this strategy using transcriptome sequencing data from subjects enrolled in the Baylor College of Medicine (BCM) clinical UDN site. We first validated the method with previously solved cases and then applied it to cases that had eluded diagnosis despite a traditional, candidate-driven analysis. This approach led to multiple new diagnoses while also overcoming limitations in current ES/CMA-first clinical testing strategies. To our knowledge, this is the among the first studies using this technique in a cohort of patients with the variety of phenotypes and age ranges described here using both whole blood and skin fibroblast transcriptome data. We feel this method improves on candidate-driven approaches and supports RNA-seq as a complement to other sequencing modalities in molecular diagnostics, particularly in ES/CMA-negative cases.

Table 1. Demographics, primary phenotypes, RNA-seq tissue source, and ES/GS counts for proband participants

Category	Proband count (%)
Total	115 (100%)
Male	59 (51.3%)
Female	56 (48.7%)
Adult	34 (29.6%)
Pediatric	81 (70.4%)
Primary phenotype	
Neurology	53 (46.1%)
Musculoskeletal and orthopedics	25 (21.7%)
Allergies and disorders of the immune system	9 (7.8%)
Cardiology and vascular conditions	6 (5.2%)
Multiple congenital anomalies	6 (5.2%)
Rheumatology	5 (4.3%)
Endocrinology	3 (2.6%)
Ophthalmology	3 (2.6%)
Pulmonology	3 (2.6%)
Gastroenterology	1 (0.9%)
Oncology	1 (0.9%)
RNA-seq tissue source	
Blood only	18 (15.7%)
Skin fibroblast only	49 (42.6%)
Blood and fibroblast	48 (41.7%)
Sequencing type	
ES only	72 (62.6%)
GS only	29 (25.2%)
ES and GS	14 (12.2%)
Proband ES/GS	23 (20%)
Duo ES/GS	14 (12.2%)
Trio ES/GS	78 (67.8%)

Adult probands were >18 years of age and pediatric probands ≤18 years of age at time of enrollment. ES, exome sequencing; GS, genome sequencing.

Results

Characteristics of patients. From 2014 to 2020, 115 probands enrolled in the BCM UDN clinical site and 67 family members (182 total) underwent RNA-seq from whole blood and skin fibroblasts (Figure 1 and Table 1). Among all probands with RNA-seq, 72 (63%) had ES, 29 (25%) had GS, and another 14 (12%) had both ES and GS. In terms of tissue source, 49 (~42%) of probands had RNA-seq from fibroblasts, 18 (16%) from blood, and 48 (42%) from both tissues. The majority (70%, $n = 81$) of probands were in the pediatric age group (<18 years of age), and nearly half (46%, $n = 53$) had a primary neurologic phenotype, consistent with the overall UDN historical proportions (Table 1). Musculoskeletal and immune phenotypes followed at 22% ($n = 25$) and 8% ($n = 9$), respectively, with many probands having multiple system involvement.

Comparison of skin fibroblasts and whole blood. Principal component analysis (PCA) demonstrated notably better consistency of gene expression in skin fibroblasts than whole blood (Figure 2). Although 2 distinct clusters were present, the fibroblast data showed less variability. This finding suggests that fibroblast RNA-seq is preferable to whole blood for detecting differences in gene expression that have a biological basis and may be clinically relevant. In addition, fibroblast-derived RNA had a higher number of well-expressed genes with transcripts per million (TPM) values greater than 10 across multiple disease gene sets compared with whole blood. In 10 of 16 gene classes, at least half of the genes were well expressed in fibroblasts compared with only 1 of 16 for whole blood (Supplemental Table 1; supplemental material available online with this article; https://doi.org/10.1172/JCI141500DS1). The most significant difference was noted for aortopathy-associated genes where 80% had a TPM greater than 10 in fibroblasts compared with only 24% in whole blood ($P \leq 1 \times 10^{-10}$). This pattern was consistent for genes associated with common UDN patient phenotypes including skeletal dysplasias (60% vs. 15%, $P \leq 1 \times 10^{-10}$), autism/intellectual disability (ID) (58% vs. 25%, $P \leq 1 \times 10^{-10}$), and epilepsy (42% vs. 15%, $P \leq 1 \times 10^{-10}$) (Supplemental Table 1). Consistent with the sample type, only immunodeficiency-related genes had a higher percentage of well-expressed genes in whole blood than fibroblasts (58% vs. 45%, $P \leq 1 \times 10^{-10}$). Notably, fibroblast expression was higher in 92% (12 of 13) of genes identified in the solved cases described here (Supplemental Table 2).

Transcriptome outlier detection. Overall, each proband had an average of 3–4 genes with significantly increased or decreased expression (FDR < 0.05) relative to the entire cohort for both tissues (Supplemental Table 3). We further refined this list by prioritizing known Online Mendelian Inheritance in Man (OMIM) disease genes. For novel disease gene discovery, about 1 in 3 (fibroblasts) and 1 in 6 (whole blood) probands had a gene with low expression predicted to be intolerant of loss-of-function (pLI ≥ 0.9) or predicted to cause a dominant disorder (DOMINO score ≥ 0.8) (Supplemental Table 3). For splicing, we focused on rare events in which a particular splicing junction had not been seen more than twice in the cohort, yielding an average of 60.7 and 22.5 abnormal splicing events per proband in fibroblasts and whole blood, respectively (Supplemental Table 3). We prioritized this list by focusing on known disease genes. Splicing and expression abnormalities were validated by visual inspection of the RNA-seq alignment in the Integrative Genomics Viewer (IGV) (12). Verified results were used for targeted analysis of DNA sequencing data to identify the underlying cause for the transcriptome difference and confirm the diagnosis. For unsolved cases but with expression/splicing outliers, additional workup, including GeneMatcher (13) submissions, animal models, and long-read sequencing, were initiated as part of UDN standard practice for novel gene discovery purposes.

Diagnoses made with transcriptome-guided genomic analysis. Of the 115 probands who underwent RNA-seq, 32 (28%) were diagnosed via other methods such as research ES/GS analysis or clinical evaluation without the need for RNA-seq (Figure 1). We first validated the transcriptome-guided method in the 5 cases previously diagnosed with RNA-seq via a traditional candidate approach (Table 2). These 5 validation cases had all undergone ES, and 1 also had GS. Of the remaining 78 undiagnosed probands (Figure 1), 41 (53%) had ES, 25 (32%) had GS, and 12 (15%) had both ES and GS. A diagnosis was made in an additional 9 of these cases using the new technique (Table 2). All 9 had undergone ES, and an additional 7 also had GS, the latter needed to identify the genomic event responsible for the RNA-seq finding. Across the entire cohort, RNA-seq led to an overall diagnostic rate of 12% (14 of 115; 95% CI, 7%–19%). Excluding cases solved on ES/GS alone

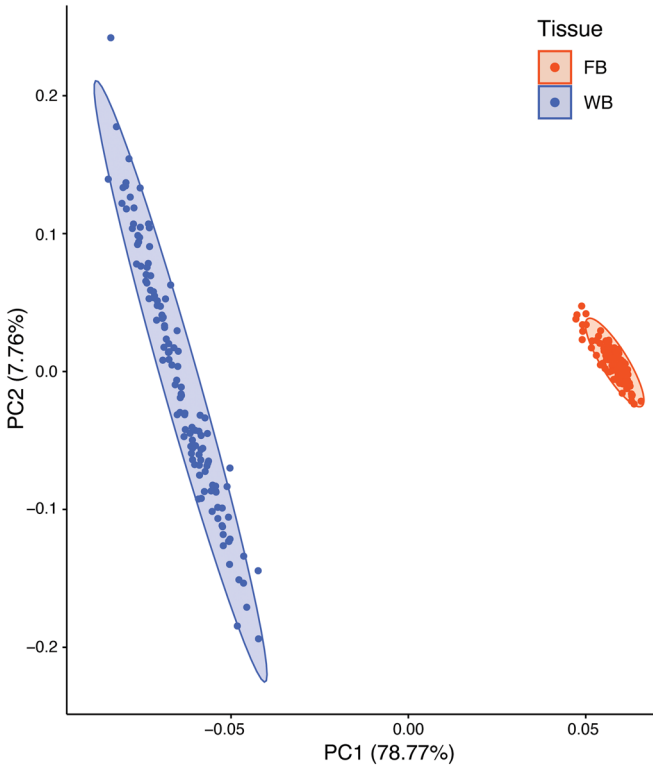


Figure 2. Principal component analysis (PCA) plot of gene expression (TPM) in whole blood (blue) and skin fibroblasts (red). Two distinct tissue clusters are visible; however, less variability is present in skin fibroblasts. This suggests that fibroblasts may be better for detecting clinically relevant differences in gene expression by RNA-seq. TPM, transcripts per million; FB, fibroblast; WB, whole blood.

without the need for RNA-seq, the diagnostic rate was 17% (14 of 82; 95% CI, 10%–26%). The causative genomic variants identified through RNA-seq included synonymous (n = 1), near intronic (3–50 bp from canonical exon boundary, n = 2), deep intronic (>50 bp away from canonical exon boundary, n = 4), promoter (n = 1), and canonical splice site single-nucleotide variants (SNVs) (n = 1) as well as both coding (n = 3) and noncoding (n = 2) deletion copy number variants (CNVs) (Figure 3). Among solved cases, 7 (50%) had RNA-seq from fibroblasts only and 1 (7%) from whole blood only, and 6 (43%) from both (Table 2). Notably, in those 6 cases, the RNA-seq from whole blood failed to identify the causative defect in half (n = 3), while none were missed from fibroblasts. This strategy also streamlined our analysis workflow; after one-time processing, the abnormally expressed genes and splicing events guided targeted analysis of existing sequencing data, and any additional confirmatory testing was done to make the final diagnosis. This contrasts with the 1–4 hours we typically require for an ES/GS research analysis to identify candidate variants and manually inspect the transcriptome for abnormalities. In one recent report, the time required was up to 6–8 hours for genome analysis (14). The following are several case examples of this approach in making diagnoses that also show the limitations of commonly used diagnostic tests.

Case 1 — *PQBP1*: This case involved a 3-year-old male referred to the UDN with multiple congenital anomalies. Medical history was significant for congenital heart defects (ventricular septal defect and patent ductus artery), vertebral anomalies (butterfly vertebrae), ectopic pelvic left kidney, hypospadias, sensorineural hearing loss, and failure to thrive. Developmentally he was delayed, rolling over at 12 months and not yet sitting independently or speaking at age 3 years. On exam, he was small (weight –2.45 SD, height –3.1 SD) and dysmorphic with microbrachycephaly (occipitofrontal circumference –4.88 SD), deep-set eyes, midface hypoplasia, broad nose, low-set ears, high palate, and 4-5 toe syndactyly among other findings (Figure 4A). Family history was significant for a maternal half-brother (not enrolled in the study) with VACTERL association (vertebral defects, anal atresia, cardiac defects, tracheo-esophageal fistula, renal anomalies, and limb abnormalities) and a half-sibling who died in utero with cardiac defect and gastroschisis versus limb-body wall defect. The differential diagnosis included Coffin-Siris syndrome or chromosomal abnormality; however, trio ES and CMA were negative, as well as GS sent through the UDN.

RNA-seq analysis detected a nearly 50% reduction in the expression of *PQBP1* in the proband compared with controls in whole blood. Reanalysis of GS data revealed a hemizygous deep intronic variant in *PQBP1* (c.180-306G>A) inherited from the heterozygous mother that activated a cryptic splice donor near the variant site (Figure 4B). The RNA-seq pipeline also detected an abnormal splicing pattern that resulted in an out-of-frame pseudoexon between exons 3 and 4, as well as more distal intron retention (Figure 4C).

Defects in *PQBP1* cause Renpenning syndrome (MIM 309500), an X-linked ID syndrome characterized by males with microcephaly, short stature, cardiac and renal anomalies, small testes, and dysmorphic features (15). The encoded polyglutamine-binding protein 1 has been shown to have an essential role in neurodevelopment (16). Most of the causative *PQBP1* variants are exonic frameshift deletions leading to markedly reduced gene expression via nonsense-mediated mRNA decay (NMD) and impaired protein function (17, 18). The RNA-seq findings in this proband were consistent with NMD due to the out-of-frame pseudoexon creation and other splicing abnormalities. With the RNA-seq results and substantial phenotypic overlap, we diagnosed Renpenning syndrome in the proband. Notably, no sequencing reads covered this variant on the previous ES, nor did it appear on the GS report. In addition, the SpliceAI prediction tool (19) considered this to be a benign change (score 0.33) unlikely to affect splicing. Therefore, the RNA-seq–directed analysis was indispensable in making the diagnosis. As the proband's mother was heterozygous for this change, there were also important recurrence risk issues discussed with the family.

Case 2 — *CLTC*: This case involved a 14-year-old male enrolled in the UDN with a history of ID and dysmorphic features. Delays in development were global, with walking occurring at 2.5 years and first words at 4–5 years. At age 14 years, his IQ was measured at 60–70 with academic skills at a second-grade level. He had maladaptive behaviors, including aggressive features, self-harm, violent outbursts, and refusal to eat, necessitating a G-tube placement. Other issues included chronic constipation and seizures. Physical exam was significant for marked hypertelorism, broad forehead, low posterior hairline, and hypotonia (Figure 5A). An

Table 2. Causative variants identified with the transcriptome-directed approach

Diagnosis (Dx)	New Dx	Gene	Sampled Tissue(s)	Genomic Sequencing	Genomic Variant	Variant Category	Zygosity	Inheritance	RNA-seq Finding	Disease Mechanism	Age, Sex	Phenotype
Renpenning syndrome	Yes	*PQBP1*	Blood	ES, GS	NM_001032383.1: c.180-306G>A	Deep intronic SNV	Hemi	Maternal	Decreased expression, abnormal splicing	Deep intronic variant activates cryptic splice site creating out-of-frame pseudoexon	3 yr, M	ID, DD, cardiac defect
CLTC-associated ID syndrome	Yes	*CLTC*	Fibroblast, blood	ES, GS	NC_000017.10: g.57756685_57779426del (22.7 kb)	Coding CNV	Het	Paternal	Decreased expression (~50%) in both tissues	Multi-exon deletion leads to NMD	14 yr, M	ID, DD, seizures
Koolen–de Vries syndrome	Yes	*KANSL1*	Fibroblast	ES, GS	NC_000017.10: g.44174219_44481307 (307 kb)	Coding CNV	Het	De novo	Decreased expression (~50%)	Initiation codon–containing exon deleted	7 yr, F	ID, DD, seizures
DOPA-responsive dystonia	Yes	*SPR*	Fibroblast	ES	NC_000002.11: g.73114483C>T	Promoter SNV	Hom	Unknown, no parental samples	Decreased expression (near zero)	Promoter variant 54 bp before transcription start site	16 yr, F	Slow speech, inactive during the first years of life
NSD2-associated ID syndrome	Yes	*NSD2*	Fibroblast, blood	ES, GS	NC_000004.11: g.1870996_1874851del (3.9 kb)	Noncoding CNV	Het	Not maternal, no paternal sample	Decreased expression (~50%) in both tissues	Exon 1 deletion including transcription start site and promoter/enhancer elements	26 yr, M	DD, failure to thrive, microcephaly, myopathy
CHASERR-associated syndrome	Yes	*CHASERR*	Fibroblast	ES, GS	NC_000015.9: g.93411267_93433682del (22.4 kb)	Noncoding CNV	Het	De novo	Decreased expression (~50%)	Deletion of noncoding RNA gene causes abnormal *CHD2* expression	1 yr, F	DD, hypotonia, abnormal EEG, brain hypomyelination
Brachydactyly syndrome, type B	Yes	*ROR2*	Fibroblast, blood	ES, GS	NM_004560.4:c.98-966T>C	Deep intronic SNV	Het	Maternal	Decreased expression in fibroblasts only	Deep intronic variant alters splicing	8 yr, M	Brachydactyly, nail hypoplasia
TBCK-associated syndromic encephalopathy	Yes	*TBCK*	Fibroblast	ES	NC_000004.11:g.107092252_107092427del (176 bp)	Coding CNV	Hom	Maternal, paternal	Decreased expression (~30%), abnormal splicing	Exon 23 deletion alters reading frame leading to NMD	3 yr, F	DD, hypotonia, seizures
RPL13-associated SEMD	Yes	*RPL13*	Fibroblast	ES, GS	NM_000977.3:c.477+1G>A	Canonical splice site SNV	Het	Paternal	Abnormal splicing (intron retention)	Intron retention and extension of 54 bp	6 yr, F	SEMD
PRUNE1-associated DD syndrome	No[A]	*PRUNE1*	Fibroblast	ES	NM_021222.3:c.933G>A (p.Thr311=)	Synonymous SNV	Hom	Maternal, paternal	Abnormal splicing (exon skipping)	Synonymous splice-site variant causes exon 7 skipping	5 yr, M	Spastic paraplegia, hypotonia, nonverbal, nonambulatory, ID, DD
Noonan-like syndrome	No[A]	*LZTR1*	Fibroblast, blood	ES	NM_006767.4:c.1943-256C>T	Deep intronic SNV	Het	Maternal	Decreased expression, abnormal splicing in fibroblasts only	Intron retention leads to NMD in trans with second variant (ref. 52)	7 yr, M	Noonan-like (cardiac defects, DD, dysmorphic features)
Noonan-like syndrome	No[A]	*LZTR1*	Fibroblast, blood	ES, GS	NM_006767.4:c.1943-256C>T	Deep intronic SNV	Het	Maternal	Decreased expression, abnormal splicing in fibroblasts only	Intron retention leads to NMD in trans with second variant (ref. 52)	2 yr, M (brother of above)	Noonan-like (cardiac defects, DD, dysmorphic features)
Au-Kline syndrome	No[A]	*HNRNPK*	Fibroblast, blood	ES	NM_031263.4:c.214-35A>G	Near intronic SNV	Het	De novo	Decreased expression, abnormal splicing in both tissues	Intronic variant causes intron retention altering reading frame	3 yr, F	ID, DD
AP4M1-associated spastic paraplegia	No[A]	*AP4M1*	Fibroblast	ES	NM_004722.4:c.929+5G>A	Near intronic SNV	Hom	Maternal, paternal	Decreased expression (~25% normal)	Intron retention and extension of 52 bp leading to NMD	17 yr, M	DD, seizures, myopathy

Coordinates according to GRCh37 (hg19). NMD, nonsense-mediated decay; DD, developmental delay; ID, intellectual disability; SEMD, spondyloepimetaphyseal dysplasia; Hemi, hemizygous; Het, heterozygous; Hom, homozygous; M, male; F, female; Zyg, zygosity; ES, exome sequencing; GS, genome sequencing; SNV, single-nucleotide variant; CNV, copy number variant. [A]Validation case previously solved using candidate approach.

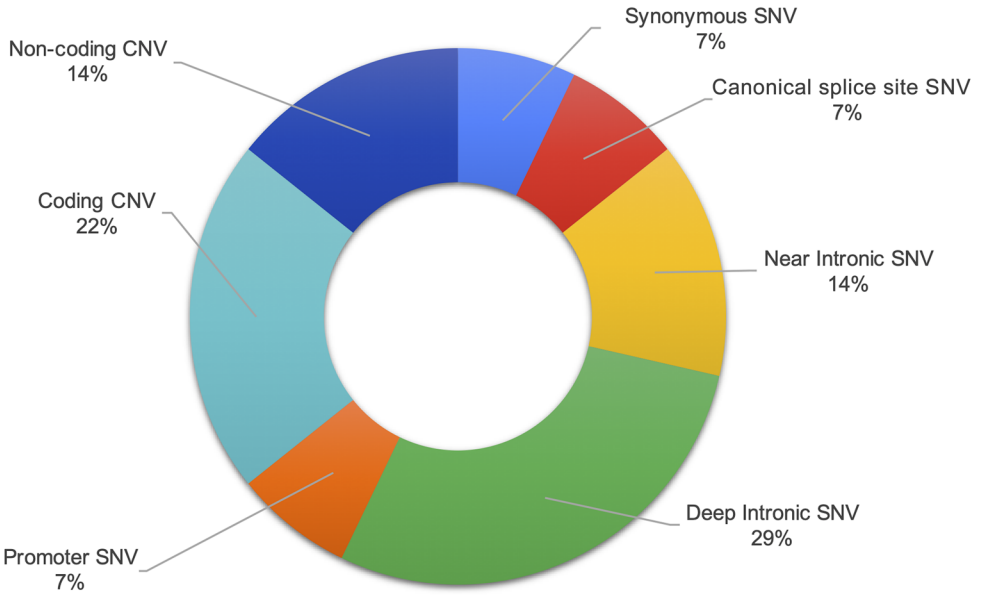


Figure 3. Causative genomic variants identified through RNA-seq–directed genomic analysis. Variant types included synonymous (n = 1), near intronic (3–50 bp from canonical exon boundary, n = 2), deep intronic (>50 bp away from canonical exon boundary, n = 4), promoter (n = 1), and canonical splice site SNVs (n = 1) as well as both coding (n = 3) and noncoding (n = 2) deletion CNVs. SNV, single-nucleotide variant; CNV, copy number variant.

extensive previous genetic workup, including karyotype, CMA, trio ES, and fragile X testing, was negative.

RNA-seq analysis of both whole blood and fibroblast data demonstrated approximately half the normal expression of 2 genes, *CLTC* and *PTRH2*. These 2 genes are adjacent to each other on chromosome 17q23.1, suggesting a possible contiguous deletion. Defects in *CLTC* are associated with an autosomal dominant disorder (MIM 617854) with a variable phenotype that includes ID, developmental delay (DD), and epilepsy (20, 21). In contrast, infantile-onset multisystem neurologic, endocrine, and pancreatic disease (IMNEPD) is caused by biallelic pathogenic variants in *PTRH2* (22).

Genome sequencing revealed a heterozygous 22.7 kb deletion (chr17: 57756685–57779426) that removed the segment from exon 18 to the transcription end of *CLTC* and part of the adjacent *PTRH2* (Figure 5B), consistent with a diagnosis of *CLTC*-associated ID. Polymerase chain reaction (PCR) analysis (Figure 5C) confirmed the deletion in the proband and his father. Notably, the father reported a history of special-education classes due to learning difficulties, a finding consistent with the variable expressivity of the *CLTC*-related syndrome (21). The inherited nature of the deletion also raised important genetic counseling issues for the family. Of note, this deletion was not called on trio ES, and there was no coverage of *CLTC* on the previous CMA.

Case 3 — *KANSL1*: The third case involved a 7-year-old female referred to the UDN with ID, DD, dysmorphic features, and epilepsy. Developmentally, she sat at 7 months and walked at 23 months. At age 7 years, her IQ was in the 50s, and she was only able to combine 2–3 words. Dysmorphic features included blepharophimosis, epicanthal folds, protruding ears, and a tubular nose with a broad tip (Figure 6A). Other significant findings in her history included scoliosis, hyperopia, strabismus, and mild joint hypermobility. Her parents and sister were in good health, and other family history was noncontributory. Trio ES, including subsequent reanalysis, was negative. CMA was negative in 2012 and 2018.

RNA-seq analysis of fibroblast data identified a nearly 50% decrease in expression of *KANSL1*. Defects in *KANSL1* cause Koolen–de Vries (KdV) syndrome (MIM 610443), an autosomal dominant ID syndrome with distinctive facial features, epilepsy, congenital heart defects, and renal and urologic anomalies (23, 24). The KdV syndrome is caused by either a heterozygous microdeletion at chromosome 17q21.31 that includes *KANSL1* or heterozygous loss-of-function variants in *KANSL1* (23, 24). On manual review of this patient's sequencing, a heterozygous *KANSL1* single-nucleotide polymorphism (SNP) in exon 14 inherited from the father was present in the ES data but absent from the RNA-seq, suggesting that allele was not expressed (Figure 6B). Manual inspection of the ES data revealed approximately half the expected coverage of the initiation codon–containing exon 2 in the proband compared with the parents (Supplemental Figure 1A). A subsequent review of CMA data showed evidence of a deletion at that locus (Supplemental Figure 1B). However, the deletion was considered benign, as it lay in a known complex region with common background CNV variation where other similar benign losses have been reported in ClinVar and Decipher (25–27). GS failed to call any variants here due to problems in read-mapping in the region. Nevertheless, given the RNA-seq results and phenotypic fit, PCR analysis was done and identified a 307 kb heterozygous deletion (chr17: 44174219–44481307) at 17q21.31 removing the first 2 exons of *KANSL1* (Figure 6C), consistent with a diagnosis of KdV syndrome. The deletion was not present in either parent, indicating a low risk of recurrence. Additional KdV-specific management, including screening for cardiac and urogenital defects, was initiated.

Case 4 — *NSD2*: The fourth case involved a 26-year-old male with DD, failure to thrive, unilateral hearing loss, microcephaly, and myopathy. He walked at 2 years and had delays in fine motor control and language. He completed 12th grade with special education. Physical exam was significant for microcephaly (occipitofrontal circumference –2.93 SD), brachycephaly, microstomia, and decreased muscle bulk and tone. An extensive genetic workup, including karyotype, CMA, myotonic dystrophy type 1, mitochondrial testing, and duo ES, were nondiagnostic.

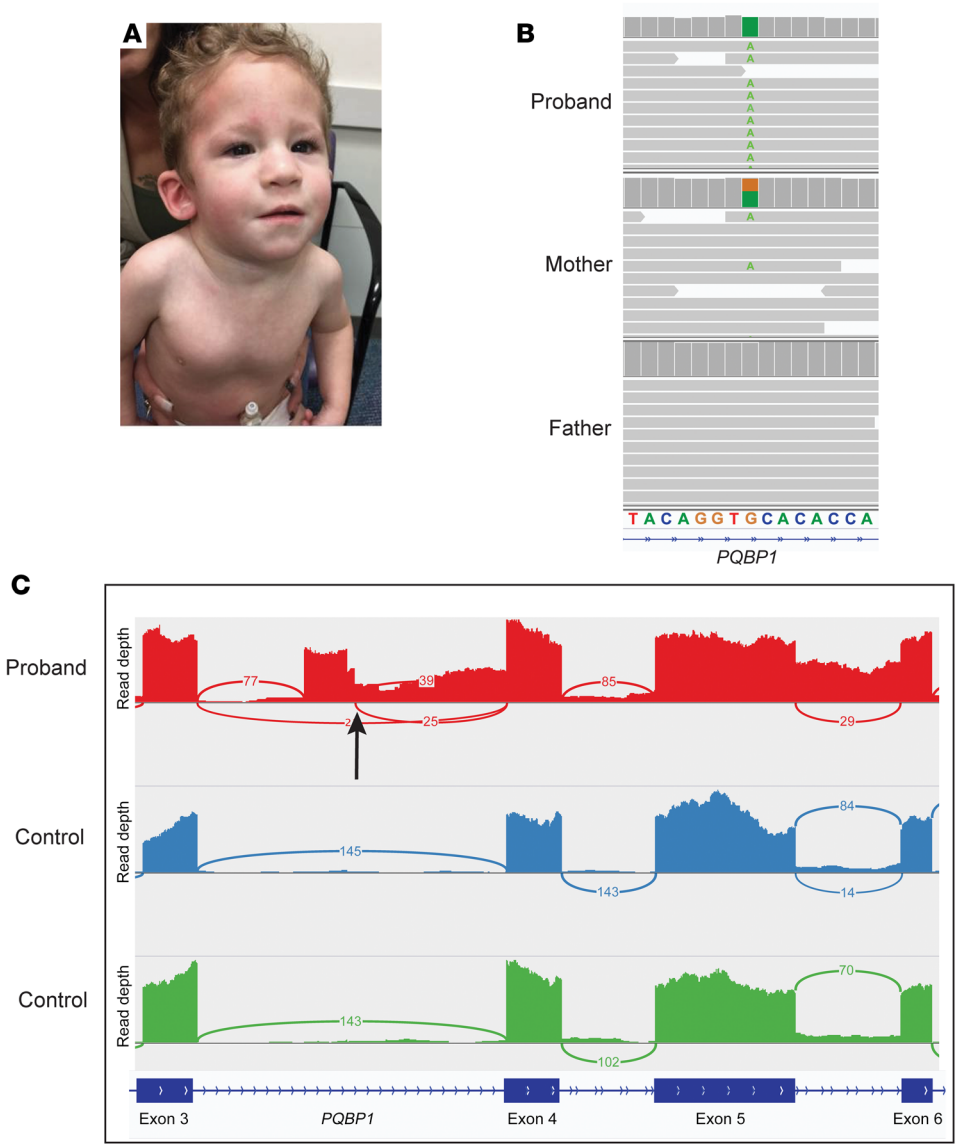


Figure 4. Case 1 – Renpenning syndrome. (**A**) Dysmorphic features, including microbrachycephaly, deep-set eyes, midface hypoplasia, broad nose, and low-set ears. (**B**) GS with hemizygous deep intronic *PQBPQ1* variant (green) inherited from heterozygous mother. (**C**) RNA-seq sashimi plot from whole blood showing out-of-frame pseudoexon and distal intron retention in the proband (red) but absent from controls (blue/green). Black arrow indicates the location of *PQBP1* intronic variant. GS, genome sequencing.

RNA-seq analysis of both whole blood and fibroblast data demonstrated approximately half-normal expression of *NSD2*. Also known as *WHSC1*, *NSD2* is one of 2 genes within the Wolf-Hirschhorn syndrome (WHS) critical region (WHSCR) on chromosome 4p16.3. *NSD2* is predicted to be intolerant of loss of function (pLI = 1), and reports have described truncating *NSD2* variants in association with a phenotype resembling a mild form of WHS (28–30).

Suspecting a noncoding causative variant given the negative prior ES and CMA, GS was requested and revealed a heterozygous 3.9 kb deletion (chr4: 1870996–1874851) containing part of *NSD2* (Figure 7A). The deletion encompassed all of *NSD2* exon 1 (representing >80% of the 5′UTR), including the transcription start site and the upstream region containing the promoter and enhancer elements (31). PCR analysis confirmed that the deletion was not inherited from the mother (Figure 7B); however, a paternal sample was not available for segregation. Notably, the small noncoding deletion was not called on ES or CMA due to the lack of coverage in this region.

With the deletion finding as well as both whole-blood and fibroblast transcriptome data demonstrating half-normal *NSD2* expression, the diagnosis of *NSD2*-associated neurodevelopmental disorder was made. The proband's phenotype was consistent with a mild form of WHS in that he had learning disabilities, decreased muscle bulk/hypotonia, hearing loss, microcephaly, and postnatal growth retardation.

Discussion

Here we describe a transcriptome-directed approach to genomic analysis that facilitated new diagnoses in previously unsolved cases and overcame limitations in widely applied genetic testing tools like ES and CMA. In contrast to RNA-seq efforts that rely on the initial identification of candidate genomic variants, we searched for global outliers in the transcriptome to guide the subsequent analysis, resulting in a diagnostic rate of 12% across the entire cohort, or 17% after excluding cases previously solved on ES/GS alone. In reality, this strategy is not only a different choice of analytical model, it also emphasizes a different philosophy in seeking

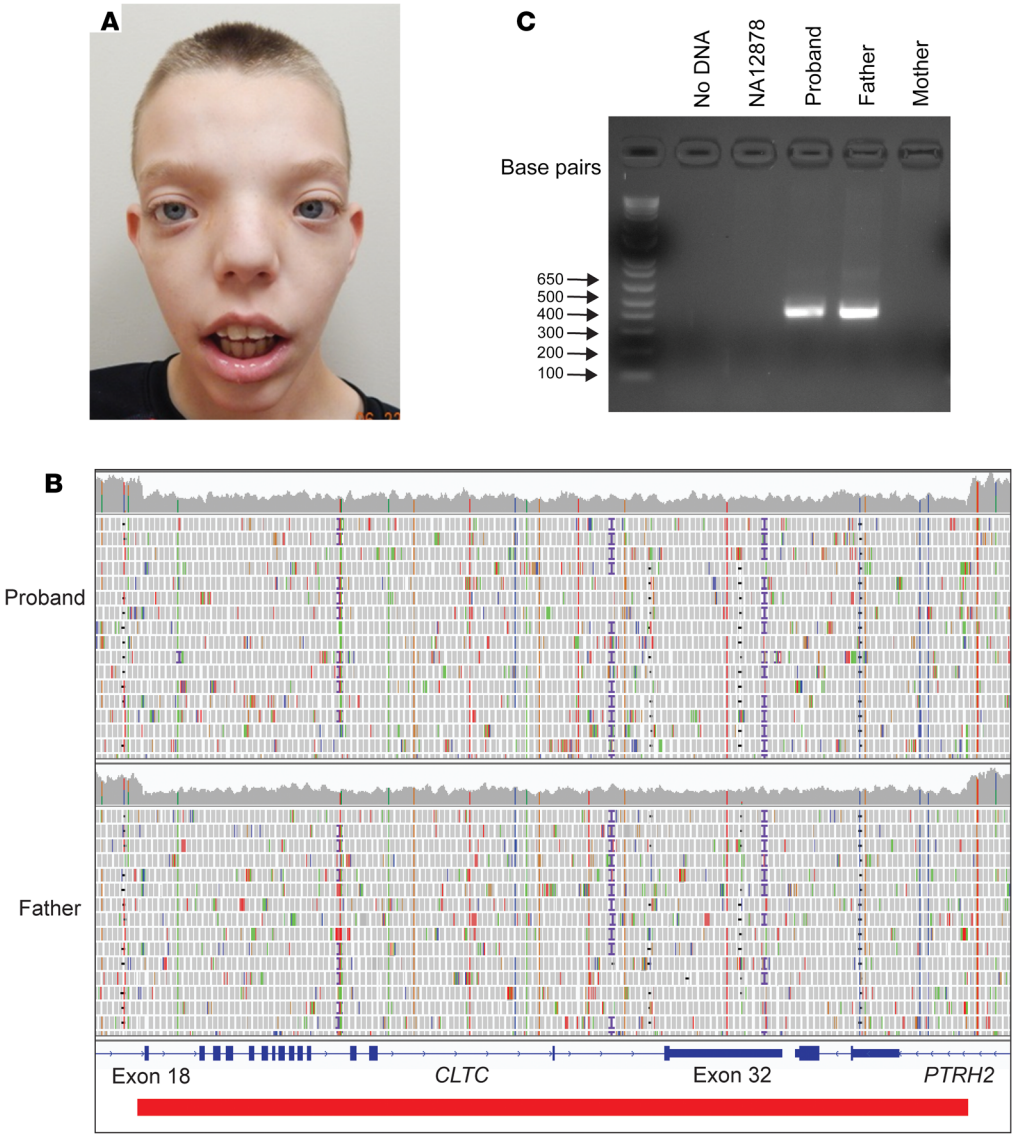
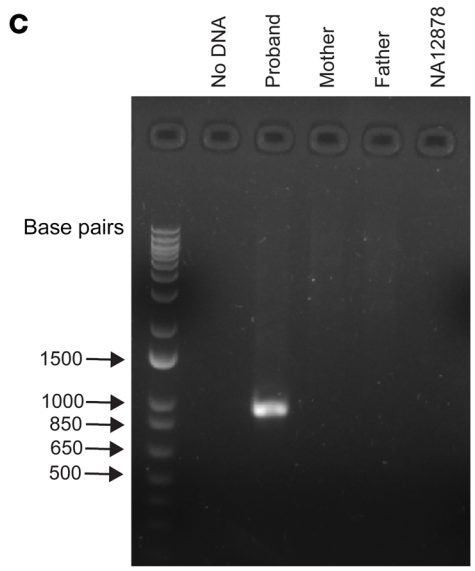


Figure 5. Case 2 – *CLTC*-associated ID syndrome. (**A**) Dysmorphic features, including hypertelorism, broad forehead, and low posterior hairline. (**B**) *CLTC* deletion (red bar) on GS encompassing exons 18–32. (**C**) PCR confirmation of deletion in proband and father but absent from mother and control (NA12878). Expected size with deletion = 391 bp. GS, genome sequencing; ID, intellectual disability.

genotype-phenotype associations (top-down vs. bottom-up). In effect, the transcriptome profile can be treated as the phenotype compared with the genotype deciphered by ES/GS/CMA. Any change of the transcriptome phenotype is the most direct reflection of changes occurring at the genetic or epigenetic levels.

Other studies that have used control data sets to detect outliers include Gonorazky et al. and Cummings et al. who achieved a diagnostic rate of around 35% in patients with neuromuscular disorders using mainly muscle tissue (4, 6), Fresard et al. who diagnosed 7.5% of undiagnosed cases from whole blood (7), and Kremer et al. who diagnosed 10% of patients with mitochondrial dysfunction using skin fibroblasts (5). However, our approach is the first report to our knowledge taking this approach for subjects with the diverse phenotypes present in the UDN in both whole blood and fibroblasts. Lee et al. recently used a traditional, candidate genomic variant approach to diagnose subjects enrolled in the UDN with a variety of phenotypes using multiple tissue types (whole blood, skin [fibroblasts], muscle) (8). While this resulted in an 18% diagnostic rate, most of the candidates analyzed ended up being benign on further review of RNA-seq data. Also, that strategy requires that the genomic variant be known a priori, having been identified on primary analysis. In the case of splice-altering variants, this is primarily based on bioinformatic predictions that frequently do not agree and lead to false negatives and false positives (10, 11). It will also miss variants that fall below the primary assay's resolution but may still be clinically relevant, for example, single-exon deletions or mosaic variants. In contrast, our approach identifies expression and splicing outliers in the transcriptome first to directly guide the subsequent genomic analysis, and as such, it is agnostic to the underlying mechanism (e.g., coding vs. noncoding, synonymous vs. nonsynonymous). It is also agnostic to presumed inheritance patterns that may be incorrectly assigned due to variable expressivity, as was the case in the *CLTC* example. The main requirement is a large control transcriptome data set derived from the same tissue, ideally sequenced on the same platform.

RNA-seq is beneficial to both ES and GS, but for different reasons. In our experience here, RNA-seq–directed analysis showed



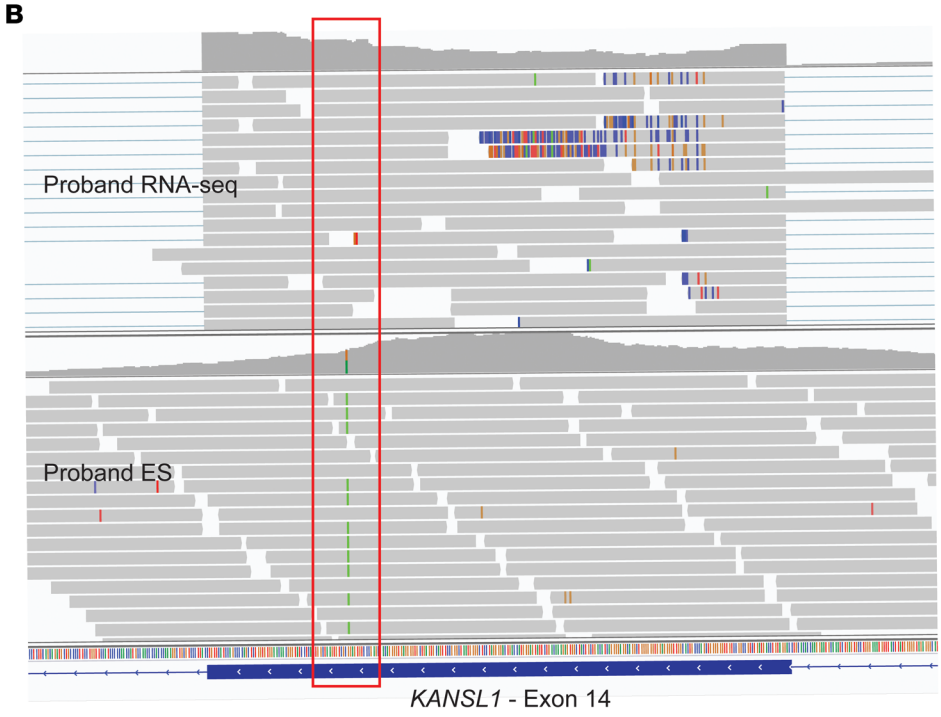


Figure 6. Case 3 – Koolen–de Vries syndrome. (**A**) Dysmorphic features, including blepharophimosis, epicanthal folds, protruding ears, and a tubular nose with a broad tip. (**B**) Exon 14 SNP (red box) in ES but absent in RNA-seq consistent with loss of that allele. (**C**) PCR confirmation of de novo deletion in proband absent from parents and control (NA12878). Expected size with deletion = 926 bp. ES, exome sequencing; SNP, single-nucleotide polymorphism.

that ES does miss clinically relevant variants, mainly because it only sequences 1%–2% of the entire genome, leaving behind potentially pathogenic noncoding variants and small CNVs. In contrast, the challenge with GS is interpreting the tremendous number of VUS identified, primarily in noncoding regions. This issue is likely a significant deterrent to the broader application of GS and contributes to the relatively modest improvement in diagnostic rate with GS alone, which one recent study showed being only 7% higher than ES (32). In our experience, GS cases require additional time to analyze compared with ES due to their complexity and the volume of data generated. Furthermore, interpretation guidelines set by the American College of Medical Genetics and Genomics (ACMG) apply mainly to coding changes (33), making interpreting noncoding variants found on GS challenging and serving as an opportunity for our RNA-seq–directed approach to streamline analysis. For example, in case 1, the causative *PQBP1* deep intronic variant was not listed on the clinical GS report despite full review following the ACMG guidelines and was considered benign by even the recent, machine learning–based SpliceAI prediction tool (19). Only through the RNA-seq–guided GS reanalysis was the abnormal *PQBP1* expression and splicing noted, directing us to the causative deep intronic variant. This suggests that the combination of RNA-seq and GS analysis may be significantly more informative than GS analysis alone.

This study also shows that CMA has notable limitations, as it missed all 5 of the causative CNVs identified with the RNA-seq–directed approach described here. The ACMG recommends that CMA designs allow detection of gains or losses of 400 kb or larger (34), and some modern CMA platforms offer single-exon resolution of clinically relevant genes (35). In case 2, the 22.7 kb multiex-

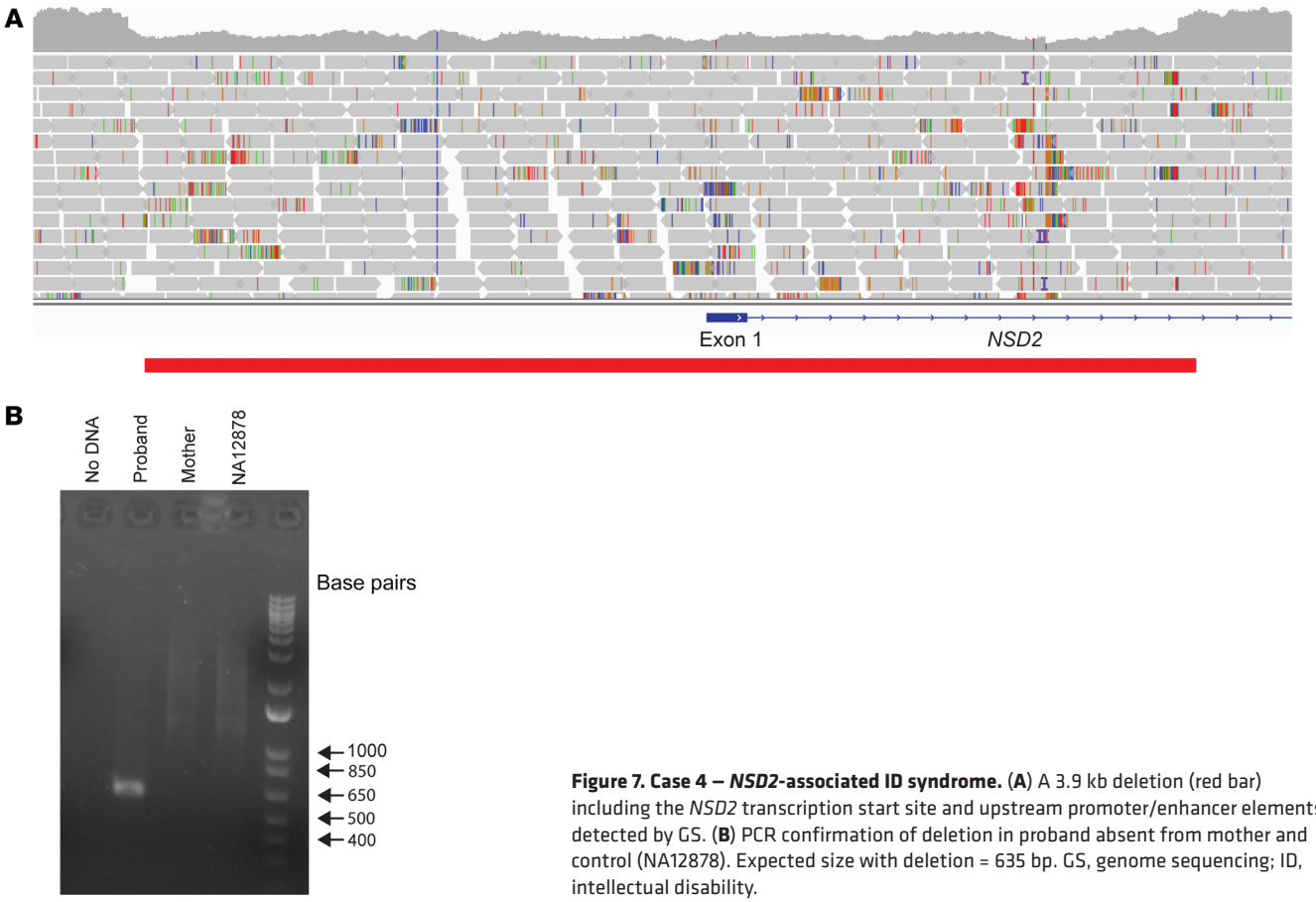


Figure 7. Case 4 — *NSD2*-associated ID syndrome. (**A**) A 3.9 kb deletion (red bar) including the *NSD2* transcription start site and upstream promoter/enhancer elements detected by GS. (**B**) PCR confirmation of deletion in proband absent from mother and control (NA12878). Expected size with deletion = 635 bp. GS, genome sequencing; ID, intellectual disability.

on *CLTC* deletion was missed due to a lack of CMA probe coverage, likely because it was only disease associated in 2016 (20). Case 3 demonstrates that CMA cannot be accurately used to detect copy number changes in the complex 17q21.31 region near the 5′ end of *KANSL1*. In case 4, there was no CMA coverage of the noncoding *NSD2* deletion containing the 5′UTR and promoter/enhancer elements that resulted in a 50% reduction in this gene's expression. Last, a pathogenic homozygous exon 23 deletion in *TBCK* and a de novo 22.4-kb-long noncoding RNA deletion of *CHASERR* were missed on CMA in 2 additional cases, respectively (Table 2). Although ES may also detect some larger CNVs, especially those 3 exons or larger (36), none of the causative deletions in the above cases were reported by ES, as current pipelines to identify CNVs still suffer from batch-to-batch variation (37). Ultimately, GS may overcome CMA coverage limitations, although costs continue to prohibit broad adoption, and mapping short reads in complex regions remains a major challenge (e.g., *KANSL1* case).

When performing RNA-seq, the correct choice of source material is essential since gene expression is cell and tissue specific. Commonly used tissues include skin (fibroblasts), muscle, and whole blood, with only the latter not requiring a biopsy. Our results demonstrate that gene expression in fibroblasts is significantly higher for a greater number of genes than whole blood, particularly for clinically relevant genes. This is especially pertinent to genes expressed in the nervous system given the common neurologic phenotypes seen in the UDN and undiagnosed patients in general. From genes associated with autism/ID and epilepsy to CNS malformations, our data show markedly higher expression in fibroblasts (Supplemental Table 1). Among the solved cases described here with a neurologic component, more than 90% of the causative genes had higher expression in fibroblasts, and fewer than half were well expressed (TPM > 10) in whole blood (Supplemental Table 2). Also, PCA indicates that gene expression is less variable in fibroblasts and may be better for detecting subtle differences that are biologic in origin and clinically relevant. These findings prompted our decision to prioritize skin biopsies on probands for RNA-seq. Although more invasive to obtain and subsequent processing was needed, we felt the improved consistency and number of well-expressed (TPM > 10) genes in fibroblasts associated with common proband phenotypes justified the additional effort. Other recent studies have borne this out including Gonorazky et al. who concluded that whole blood is an inadequate source for most neuromuscular diseases (4), and Aicher et al. who showed that fibroblasts had the lowest number of inadequately expressed genes among clinically accessible tissues, including whole blood (38). Our diagnostic rate of 12% across the entire cohort or 17% after excluding cases diagnosed with ES/GS contrasts with the 7.5% reported by Fresard et al. using whole blood alone. In fact, of the 6 cases described here where we sampled both tissues, RNA-seq from whole blood failed to identify the causative defect in 3 cases (50%), while none were missed using fibroblasts. Nevertheless, many genes are still not well expressed in clinically accessible tissues or may undergo tissue-specific alternative splicing (38). Techniques like fibroblast transdifferen-

tiation to other disease-relevant cell types show promise for overcoming this limitation (4).

In conclusion, we made multiple new diagnoses in our undiagnosed patient cohort using a transcriptome-directed approach to genomic analysis. In line with other studies, the diagnostic rate with the addition of RNA-seq was 12% across the entire cohort or 17% after excluding cases solved on ES/GS alone. We showed that RNA-seq derived from fibroblasts exhibited higher and less variable gene expression in clinically relevant genes. Central to the UDN's overall mission, this approach also identified potentially novel disease genes that are under further investigation. Last, we demonstrated that disease-causing variants are missed on commonly used testing platforms such as ES and CMA. Our findings suggest a transcriptome-directed approach to rare disease diagnosis may improve diagnostic rates, in particular as a complement to GS in ES/CMA-negative cases. However, controlled studies comparing standard RNA-seq implementations to the approach used here are needed to determine the statistical significance and applicability to clinical diagnostic practice.

Methods

Study design. Patients were enrolled in the UDN according to standard inclusion criteria that included objective findings pertinent to the phenotype and no diagnosis despite thorough evaluation by health care providers. Molecular testing (e.g., ES or GS) is a mainstay of the UDN and was done when a patient's medical history and physical exam strongly suggested an underlying genetic cause. RNA-seq was done only if such testing done before or as part of the UDN evaluation was nondiagnostic (no pathogenic or likely pathogenic variants) according to the ACMG interpretation guidelines (33). In some cases (28%, n = 32), a diagnosis was reached through a research reanalysis of the ES/GS data alone, clinical evaluation, or other diagnostic tests without the need for the RNA-seq analysis (Figure 1). As our data set grew over time, we performed RNA-seq on fewer family members and prioritized additional probands. For validation, we selected the 5 previously solved RNA-seq cases where a traditional candidate approach had been used to identify a causative variant from research ES/GS analysis (Table 2). For fibroblasts, a skin biopsy was taken, and we followed our internal protocol to generate cell cultures followed by RNA extraction. Whole-blood samples were collected in PAXgene (QIAGEN) whole-blood RNA tubes, and intracellular RNA was extracted and processed according to the manufacturer's recommendation.

RNA-seq. RNA from whole blood and skin fibroblasts was quantified and processed using a stranded, polyA-tailed kit (Illumina) before being multiplexed and subjected to 150 bp paired-end sequencing at the BCM Laboratory for Translational Genomics, with approximately 30–50 million reads generated per sample. Sequencing was performed in separate batches over the course of the study. The sequencing data were processed with a pipeline adapted from one developed by the GTEx Consortium (39). Briefly, fastq files were aligned to the GRCh37/hg19 reference sequence using STAR-2.6.1b (40) in 2-pass mode, and duplicates were marked with Picard (41). We quantified gene expression using RSEM (42) to generate TPM values for expressed genes in each sample. The processed alignment files were then used as input for the outlier detection step.

Outlier detection and prioritization. For aberrant expression and splicing detection, we used the Detection of RNA Outlier Pipeline (DROP) using the default, recommended settings (43). This workflow analyzes RNA-seq data sets to identify genes with aberrant expression levels using OUTRIDER (44), and aberrant splicing using FRASER (45), among a large group of samples, while automatically controlling for latent confounders. To minimize tissue-specific differences, we processed the data from skin fibroblasts and whole blood separately, yielding more than 100 cases from each tissue, which was well above the recommended minimum of 50–60 samples to maximize outlier sensitivity (43). Our large cohort effectively served as its own control data set and obviated the need for parental samples to detect deviations in expression or splicing. Each tissue data set required approximately 24 hours to complete the entire pipeline on a 12-core, 24-thread Linux server with 64 GB RAM. This was a one-time processing step, and subsequent analysis did not require re-running the pipeline.

The output of the DROP expression module is a list of outlier genes in each sample along with statistical information such as multiple-testing-adjusted P values, z scores, and fold changes for each deviation compared with the cohort. We identified outlier genes with large under- or overexpression in each sample at a false discovery rate (FDR) of 0.05 per OUTRIDER recommendations (44). The splicing module similarly provides relevant statistics and frequency of each abnormal splicing event within the data set. Theorizing that rare splicing events are more likely to be pathogenic, we focused on potentially novel splicing events that were seen no more than 2 times in their respective tissue using established FRASER statistical cutoffs (45).

To facilitate the identification of clinically relevant aberrations, we developed custom Perl scripts to annotate the DROP output with information such as OMIM disease gene status (46), gnomAD loss-of-function intolerant (pLI) scores (47), DOMINO probabilities of causing dominant disease (48), ClinGen haploinsufficiency and triplosensitivity scores (49), and GTEx TPM values (39). The annotated output was then used to guide the analysis of genomic sequencing data or request additional confirmatory testing (e.g., Sanger sequencing, PCR studies).

Data and materials availability. UDN sequencing data are available through dbGaP (accession: phs001232.v2.p1) and the UDN Gateway. Phenotype data with flagged genes of interest have been submitted to Phenome Central. Variants thought to diagnose the patients have been submitted to ClinVar. The DROP pipeline to compute expression and splicing outliers is available for download at https://github.com/gagneurlab/drop. Disease gene lists described here are available for download at https://github.com/drmurdock/rnaseq. RNA-seq count tables generated in this study from fibroblasts and whole blood are available for download from Zenodo (https://doi.org/10.5281/zenodo.3963474 and https://doi.org/10.5281/zenodo.3963470).

Statistics. We performed a PCA comparing gene expression in the 2 tissues and analyzed expression profiles of genes in 16 disease classes reflecting the varied phenotypes of probands enrolled in the BCM UDN site (Supplemental Table 1), and considered well-expressed genes as those having a minimum of 10 TPM (38). We determined the number and percentage of genes well expressed for each respective disease class in aggregate across all samples and compared the statistical difference between tissues using the 2-sample t test in R (version 3.6.0) (50). The 95% CIs were calculated for the reported diagnostic rates with the binomCI function (51) in R using the Wald method.

Study approval. The Institutional Review Boards approved the study at the National Human Genome Research Institute (protocol

15HG0130) and BCM (approval H-34433). Written informed consent was obtained from all study participants, including consent for publication of patient photographs.

Author contributions

DRM and BL conceived and designed the experiments. DRM, HD, SC, and MJ analyzed RNA-seq data. JAR provided clinical support. HD, LCB, SC, and MJ analyzed exome and genome data. PL performed validation experiments. CAB, HTC, PM, WJC, NAH, and BL provided patient samples and clinical information. GZ and NAH performed RNA-seq. SK performed statistical analyses. MFM, VAY, and JG developed RNA-seq analysis tools. LCB, HD, JAR, MFM, VAY, JG, CAB, HTC, PM, WJC, NAH, and BL critically reviewed the manuscript. DRM wrote the manuscript.

Acknowledgments

The authors thank the families for their participation in this study. Research reported in this manuscript was supported by the NIH Common Fund, through the Office of Strategic Coordination/Office of the NIH Director under the following awards: U01HG007709, U01HG007942, and U01HG007943. Support also came from the BCM Intellectual and Developmental Disabilities Research Center (HD024064) from the Eunice Kennedy Shriver National Institute of Child Health & Human Development. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. German Bundesministerium für Bildung und Forschung (BMBF) supported the study through the e:Med Networking fonds AbCD-Net (FKZ 01ZX1706A to VAY and JG). MM was supported by the BMBF through the project MechML (01IS18053F). HTC also received funding support from The Robert and Janice McNair Foundation. See Supplemental Acknowledgments for consortium details.

Address correspondence to: David R. Murdock, Department of Molecular and Human Genetics, Baylor College of Medicine, One Baylor Plaza, MS225, Houston, Texas 77030. Phone: 713.798.3415; Email: drmurdoc@bcm.edu.

SC's present address is: Regeneron, Tarrytown, New York, USA.

MJ's present address is: Kennedy Krieger Institute, Baltimore, Maryland, USA.

PM's present address is: University of Utah and George E. Wahlen Department of Veterans Affairs Medical Center, Salt Lake City, Utah, USA.

1. Posey JE, et al. Molecular diagnostic experience of whole-exome sequencing in adult patients. *Genet Med*. 2016;18(7):678–685.
2. Yang Y, et al. Molecular findings among patients referred for clinical whole-exome sequencing. *JAMA*. 2014;312(18):1870–1879.
3. Splinter K, et al. Effect of genetic diagnosis on patients with previously undiagnosed disease. *N Engl J Med*. 2018;379(22):2131–2139.
4. Gonorazky HD, et al. Expanding the boundaries of RNA sequencing as a diagnostic tool for rare Mendelian disease. *Am J Hum Genet*. 2019;104(5):1007.
5. Kremer LS, et al. Genetic diagnosis of Mendelian disorders via RNA sequencing. *Nat Commun*. 2017;8:15824.
6. Cummings BB, et al. Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. *Sci Transl Med*. 2017;9(386):eaal5209.
7. Frésard L, et al. Identification of rare-disease genes using blood transcriptome sequencing and large control cohorts. *Nat Med*. 2019;25(6):911–919.
8. Lee H, et al. Diagnostic utility of transcriptome sequencing for rare Mendelian diseases. *Genet Med*. 2020;22(3):490–499.
9. Karam R, et al. Assessment of diagnostic outcomes of RNA genetic testing for hereditary cancer. *JAMA Netw Open*. 2019;2(10):e1913900.
10. Wai HA, et al. Blood RNA analysis can increase clinical diagnostic rate and resolve variants of uncertain significance. *Genet Med*. 2020;22(6):1005–1014.
11. Moles-Fernández A, et al. Computational tools for splicing defect prediction in breast/ovarian cancer genes: how efficient are they at predicting RNA alterations? *Front Genet*. 2018;9:366.
12. Robinson JT, et al. Variant review with the Integrative Genomics Viewer. *Cancer Res*. 2017;77(21):e31–e34.
13. Sobreira N, et al. GeneMatcher: a matching tool for connecting investigators with an interest in the same gene. *Hum Mutat*. 2015;36(10):928–930.
14. Machini K, et al. Analyzing and reanalyzing the genome: findings from the MedSeq project. *Am J Hum Genet*. 2019;105(1):177–188.
15. Lenski C, et al. Novel truncating mutations in the polyglutamine tract binding protein 1 gene (PQBP1) cause Renpenning syndrome and X-linked mental retardation in another family with microcephaly. *Am J Hum Genet*. 2004;74(4):777–780.
16. Wan D, et al. X chromosome-linked intellectual disability protein PQBP1 associates with and regulates the translation of specific mRNAs. *Hum Mol Genet*. 2015;24(16):4599–4614.
17. Germanaud D, et al. The Renpenning syndrome spectrum: new clinical insights supported by 13 new PQBP1-mutated males. *Clin Genet*. 2011;79(3):225–235.
18. Kalscheuer VM, et al. Mutations in the polyglutamine binding protein 1 gene cause X-linked mental retardation. *Nat Genet*. 2003;35(4):313–315.
19. Jaganathan K, et al. Predicting splicing from primary sequence with deep learning. *Cell*. 2019;176(3):535–548.e24.
20. DeMari J, et al. CLTC as a clinically novel gene associated with multiple malformations and developmental delay. *Am J Med Genet A*. 2016;170A(4):958–966.
21. Nabais Sá MJ, et al. De novo CLTC variants are associated with a variable phenotype from mild to severe intellectual disability, microcephaly, hypoplasia of the corpus callosum, and epilepsy. *Genet Med*. 2020;22(4):797–802.
22. Le C, et al. Infantile-onset multisystem neurologic, endocrine, and pancreatic disease: case and review. *Can J Neurol Sci*. 2019;46(4):459–463.
23. Koolen DA, et al. A new chromosome 17q21.31 microdeletion syndrome associated with a common inversion polymorphism. *Nat Genet*. 2006;38(9):999–1001.
24. Koolen DA, et al. Koolen-de Vries syndrome [updated 2019 Jun 13]. In: Adam MP, et al., eds. GeneReviews® [Internet]. University of Washington, Seattle; 1993–2020. https://www.ncbi.nlm.nih.gov/books/NBK24676/.
25. Koolen DA, et al. The Koolen-de Vries syndrome: a phenotypic comparison of patients with a 17q21.31 microdeletion versus a KANSL1 sequence variant. *Eur J Hum Genet*. 2016;24(5):652–659.
26. Firth HV, et al. DECIPHER: database of chromosomal imbalance and phenotype in humans using Ensembl resources. *Am J Hum Genet*. 2009;84(4):524–533.
27. Landrum MJ, et al. ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res*. 2016;44(D1):D862–D868.
28. Jiang Y, et al. De novo truncating variant in NSD2gene leading to atypical Wolf-Hirschhorn syndrome phenotype. *BMC Med Genet*. 2019;20(1):134.
29. Derar N, et al. De novo truncating variants in WHSC1 recapitulate the Wolf-Hirschhorn (4p16.3 microdeletion) syndrome phenotype. *Genet Med*. 2019;21(1):185–188.
30. Boczek NJ, et al. Developmental delay and failure to thrive associated with a loss-of-function variant in WHSC1 (NSD2). *Am J Med Genet A*. 2018;176(12):2798–2802.
31. Fishilevich S, et al. GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database (Oxford)*. 2017;2017:1–17.

32. Alfares A, et al. Whole-genome sequencing offers additional but limited clinical utility compared with reanalysis of whole-exome sequencing. *Genet Med*. 2018;20(11):1328–1333.
33. Richards S, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med*. 2015;17(5):405–424.
34. South ST, et al. ACMG Standards and Guidelines for constitutional cytogenomic microarray analysis, including postnatal and prenatal applications: revision 2013. *Genet Med*. 2013;15(11):901–909.
35. Tucker T, et al. Single exon-resolution targeted chromosomal microarray analysis of known and candidate intellectual disability genes. *Eur J Hum Genet*. 2014;22(6):792–800.
36. Retterer K, et al. Assessing copy number from exome sequencing and exome array CGH based on CNV spectrum in a large clinical cohort. *Genet Med*. 2015;17(8):623–629.
37. Jiang Y, et al. CODEX2: Full-spectrum copy number variation detection by high-throughput DNA sequencing. *Genome Biol*. 2018;19(1):1–13.
38. Aicher JK, et al. Mapping RNA splicing variations in clinically accessible and nonaccessible tissues to facilitate Mendelian disease diagnosis using RNA-seq. *Genet Med*. 2020;22(7):1181–1190.
39. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat Genet*. 2013;45(6):580–585.
40. Dobin A, et al. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29(1):15–21.
41. Picard. http://broadinstitute.github.io/picard.
42. Li B, Dewey CN. RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*. 2011;12:323.
43. Yépez VA, et al. Detection of aberrant events in RNA sequencing data [published online January 3, 2020]. *Protoc Exch*. https://doi.org/10.21203/rs.2.19080/v1.
44. Brechtmann F, et al. OUTRIDER: a statistical method for detecting aberrantly expressed genes in RNA sequencing data. *Am J Hum Genet*. 2018;103(6):907–917.
45. Mertes C, et al. Detection of aberrant splicing events in RNA-seq data with FRASER. *bioRxiv*. 2019. https://doi.org/10.1101/2019.12.18.866830.
46. Amberger JS, et al. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Res*. 2015;43(database issue):D789–D798.
47. Karczewski KJ, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*. 2020;581(7809):434–443.
48. Quinodoz M, et al. DOMINO: using machine learning to predict genes associated with dominant disorders. *Am J Hum Genet*. 2017;101(4):623–629.
49. Rehm HL, et al. ClinGen--the Clinical Genome Resource. *N Engl J Med*. 2015;372(23):2235–2242.
50. R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.r-project.org/.
51. Kohl M. MKinfer: Inferential statistics. R package version 0.5. 2020. http://www.stamats.de/>.
52. Johnston JJ, et al. Autosomal recessive Noonan syndrome associated with biallelic LZTR1 variants. *Genet Med*. 2018;20(10):1175–1185.