© American College of Medical Genetics and Genomics

ARTICLE Genetics inMedicine

Diagnosing Cornelia de Lange syndrome and related neurodevelopmental disorders using RNA sequencing

Stefan Rentas, PhD[1], Komal S. Rathi, MSc[2], Maninder Kaur, MSc[3], Pichai Raman, PhD[2], Ian D. Krantz, MD[3], Mahdi Sarmady, PhD[1] and Ahmad Abou Tayoun, PhD[4,5]

Purpose: Neurodevelopmental disorders represent a frequent indication for clinical exome sequencing. Fifty percent of cases, however, remain undiagnosed even upon exome reanalysis. Here we show RNA sequencing (RNA-seq) on human B-lymphoblastoid cell lines (LCL) is highly suitable for neurodevelopmental Mendelian gene testing and demonstrate the utility of this approach in suspected cases of Cornelia de Lange syndrome (CdLS).

Methods: Genotype–Tissue Expression project transcriptome data for LCL, blood, and brain were assessed for neurodevelopmental Mendelian gene expression. Detection of abnormal splicing and pathogenic variants in these genes was performed with a novel RNA-seq diagnostic pipeline and using a validation CdLS-LCL cohort ($n = 10$) and test cohort of patients who carry a clinical diagnosis of CdLS but negative genetic testing ($n = 5$).

Results: LCLs share isoform diversity of brain tissue for a large subset of neurodevelopmental genes and express 1.8-fold more of these genes compared with blood (LCL, $n = 1706$; whole blood, $n = 917$). This enables testing of more than 1000 genetic syndromes. The RNA-seq pipeline had 90% sensitivity for detecting pathogenic events and revealed novel diagnoses such as abnormal splice products in *NIPBL* and pathogenic coding variants in *BRD4* and *ANKRD11*.

Conclusion: The LCL transcriptome enables robust frontline and/or reflexive diagnostic testing for neurodevelopmental disorders.

Genetics in Medicine (2020) 22:927–936; https://doi.org/10.1038/s41436-019-0741-5

Keywords: RNA-seq; LCLs; Mendelian disorders; CdLS

INTRODUCTION

Mendelian disorders causing multiple congenital anomalies and neurodevelopmental dysfunction have an approximately 30–40% chance of getting diagnosed by clinical exome sequencing (ES).[1,2] Performing ES data reanalysis 1–2 years after the first test can further increase diagnostic yield, but a substantial proportion of patients (50–60%) remain undiagnosed.[3,4] This is partly due to limitations in the interpretation of identified variants (e.g., variants of uncertain significance) and undiscovered gene–disease relationships. Limitations with ES technology also result in missing certain variant types, such as structural rearrangements, repeat expansions, and noncoding regulatory and deep-intronic variants impacting gene expression and splicing.[2] Characterizing intronic variants that disrupt splicing is especially critical as splice-region variants cause highly damaging effects on gene products and represent about 10% of all disease-causing variants.[5] Although genome sequencing (GS) can detect noncoding variants, this method alone cannot prove if there is a functional consequence to the nearest transcript. Alternative approaches are therefore necessary to capture pathogenic events causing Mendelian conditions.

RNA sequencing (RNA-seq) is the best approach currently available to detect genome-wide differences in transcript abundance and splicing.[6] Additionally, this method can identify germline exonic single-nucleotide variants (SNVs), indels, and allele specific expression.[7–11] RNA-seq has been used to support the diagnostic testing of neuromuscular[9,10] and mitochondrial disorders.[11] The decision to investigate these conditions likely reflects the relative accessibility of affected tissue (i.e., muscle biopsy and dermal fibroblasts for neuromuscular disease and mitochondriopathy, respectively). Complex protocols to derive muscle lineage cells from patient fibroblasts have also been shown to create a tissue source that recapitulates the transcriptional diversity of primary muscle tissue.[9] Effort to circumvent invasive biopsies and complex cell culture protocols for RNA-seq testing has also been done by testing whole blood of patients with predominately neurological phenotypes and a suspected Mendelian condition.[8] This approach, however, had limited efficacy with a 7.5% diagnostic yield across all patients.[8] Therefore, implementing RNA-seq into diagnostic practice is restricted by primary tissues that do not have an ideal transcriptional

[1]Division of Genomic Diagnostics, The Children's Hospital of Philadelphia, Philadelphia, PA, USA; [2]Department of Biomedical and Health Informatics, The Children's Hospital of Philadelphia, Philadelphia, PA, USA; [3]Roberts Individualized Medical Genetics Center, Division of Human Genetics, The Children's Hospital of Philadelphia, Philadelphia, PA, USA; [4]Genomics Center, Al Jalila Children's Specialty Hospital, Dubai, UAE; [5]Genetics Department, Mohammed Bin Rashid University of Medicine and Health Sciences, Dubai, UAE. Correspondence: Mahdi Sarmady (sarmadym@chop.edu) or Ahmad Abou Tayoun (ahmad.tayoun@ajch.ae)
These authors contributed equally: Stefan Rentas, Komal S. Rathi.

Submitted 2 October 2019; accepted: 19 December 2019
Published online: 8 January 2020

landscape to test for genes causing phenotypically variable, rare disorders.

Here we demonstrate the utility of RNA-seq on patient-derived B-lymphoblastoid cell lines (LCLs) for the diagnosis of patients presenting with multiple congenital anomalies and neurodevelopmental phenotypes. LCLs are made by Epstein–Barr virus (EBV) transformation of mature B lymphocytes and have been broadly used in human genetics research for decades.[12] LCL transcriptome analysis revealed biologically relevant isoforms of known or predicted candidates of Mendelian genes causing neurodevelopmental phenotypes and/or multiple congenital anomalies (neurodevelopmental Mendelian genes, NMGs). Upon establishing a resource of expressed and testable NMGs by LCL RNA-seq, we optimized RNA-seq bioinformatic parameters to enable detection of abnormal splicing events and coding and noncoding pathogenic variants from messenger RNA (mRNA). The efficacy of this approach was demonstrated on LCLs from patients with a clinical diagnosis of Cornelia de Lange syndrome (CdLS [MIM 122470]), a rare autosomal dominant, multisystem disorder caused by pathogenic variants in cohesin complex proteins and characterized by growth and developmental delay, facial dysmorphism, microcephaly, hirsutism, limb anomalies, and congenital heart defects.[13] Our work establishes LCL RNA-seq as a viable frontline diagnostic tool in neurodevelopmental disorders and as a vital reflexive test when DNA testing is negative.

MATERIALS AND METHODS

Patient samples

All patient samples were obtained with informed consent and with the approval of the Children's Hospital of Philadelphia Institutional Review Board (IRB). Permission and consent were obtained to publish photos of individuals with CdLS in family 1 (Fig. **3d**). Patient samples utilized in this study had a clinical genetics evaluation resulting in a diagnosis of CdLS or related neurodevelopmental disorder.

Gene expression analysis for control LCL, blood and brain tissues, and CdLS-LCL patient samples

Processed gene and isoform expression data was downloaded from the University of California–Santa Cruz (UCSC) Computational Genomics Laboratory using STAR alignment and RSEM normalization using hg38 as reference genome and GENCODE v23 gene annotation.[14,15] To harmonize expression data, we processed CdLS-LCL patient samples using the same pipeline: STAR aligner v2.5.4b to align reads to hg38 reference genome and RSEM v1.2.28 to quantify gene expression in terms of fragments per kilobase of transcript per million mapped reads (FPKM) and isoform expression in terms of transcripts per million (TPM). To determine expression of genes across all samples, we first filtered our gene list to only protein-coding genes and used a cut-off of mean FPKM >1 and mean coverage >10× to identify expressed protein-coding genes. A curated list of 2541 NMGs (gene list downloaded from publicly available GeneDx Autism/ID Xpanded Panel) were annotated with Online Mendelian Inheritance in Man (OMIM) phenotype identifiers (api.omim.org) and overlapped with the set of expressed protein-coding genes in each data set to identify expressed NMGs in CdLS-LCL ($n = 1745$), Genotype–Tissue Expression sequencing project (GTEx)[16] LCLs ($n = 1706$), and GTEx whole blood ($n = 917$). Scripts used to perform principal component analysis, expression scatterplots, bar plots, and gene list intersections are available at github.com/komalsrathi/MendelianRNA-seq. Additional plots were generated with GraphPad Prism v5.

RNA-seq data processing for detecting abnormal splice events

We used paired-end RNA-sequencing data from 15 CdLS-LCL samples to detect abnormal transcript splicing. In addition, raw fastq files containing paired-end RNA-sequencing reads were downloaded from dbGaP accession: phs000424.v7.p2 for GTEx LCL ($n = 106$) and whole blood ($n = 336$). Paired-end fastq files from all three data sets were aligned to the hg19 genome using STAR v2.5.4b aligner in two-pass mode. Splice junctions identified from the first-pass alignment were filtered for mitochondrial junctions and for any unannotated junctions that were supported with fewer than five reads. The filtered splice junctions were concatenated across all samples per data set and were used as input to realign the reads with STAR second-pass alignment. Following the second-pass alignment, duplicated reads were tagged using Picard MarkDuplicates utility.

The first step in abnormal splice junction discovery was to use the sorted, de-duplicated bam files to extract all splice junctions supported by uniquely mapped reads for each sample. Next, read support normalization was done on the resulting splice junctions to correct for variability in gene expression and library size. This step was done to transform the raw read counts into proportion of reads that support a splice junction compared with all other overlapping junctions. Using the above approach, we found a total of 3534 splice junctions in CdLS-LCL samples ($n = 15$) and a total of 13,980 splice junctions in GTEx samples ($n = 442$) corresponding to the 14 tested CdLS-related genes.

The next step was to filter the junctions to identify potential deleterious splice events. Abnormal splice events were identified by multiple criteria: (1) those that were seen at a level of at least 5% of canonical junctions, (2) number of reads that support the junction in the entire data set ≥10, (3) not seen in GTEx normal samples, and (4) identified in at most one patient (patient-specific events). The resulting patient-specific abnormal transcript splice events were visualized in the form of sashimi plots that were generated using the R package ggsashimi.

Variant calling pipeline

Raw RNA-seq reads were aligned to GRCh37/hg19 reference genome using STAR version v2.5.4b.[15] Picard MarkDuplicate version 2.18.14–0 command was used to remove polymerase chain reaction (PCR) duplicates. We used Genome Analysis Toolkit (GATK) HaplotyperCaller v3.6 instead of v4 to achieve higher precision.[9,17] Variants were called according to GATK best practices for RNA-seq variant calling with minor modifications.[18] Applying SplitNCigarReads to the bam files reduced overall variant calling sensitivity, therefore we excluded this step from the final pipeline. Instead, Filter_reads_with_N_cigar parameter was applied in HaplotypeCaller to remove reads with CIGAR strings containing N operator. Variant calling region was set to exons of the genes (exons of all transcripts) listed in Table S6 and surrounding 10 bp flanking intronic regions. Impact prediction of variants was performed using SnpEff v4.3.[19] The following filters were applied to variants (condition to keep variants listed): (1) quality filter: QD ≥2 and FS <30; (2) variant impact filter: synonymous, missense, or any variants marked as impact "HIGH" by SnpEff; (3) common variants filter: gnomAD maximum subpopulation frequency (AF_PopMax) <0.2%. Coverage statistics were calculated using GATK DepthOfCoverage v3.6 over two bed files: (1) exons of the genes used for variant calling and (2) 10 bp flanking intronic regions of those genes. Variants from BAM files were displayed using Alamut (v2.11) and IGV (v2.6.3).

RESULTS

Using RNA-seq as a diagnostic tool for Mendelian disorders requires extracting mRNA from accessible patient tissue. Determining which tissue to assay is critical as the disease-causing gene could be silenced or display tissue specific splicing that obscures data interpretation. To determine if blood is an appropriate specimen type for RNA-seq, we compared its gene expression profile to LCLs using data from GTEx.[16] Choosing to compare blood with LCL was done since EBV transformation of B lymphocytes is a relatively easy and reproducible method to generate patient cell lines with very low variant rates and stable karyotypes.[12,20–22] We set an expressed gene threshold of >1 FPKM and read depth of >10× across exonic regions. Overall, nearly twofold more genes were expressed above this threshold in LCLs ($n = 10{,}612$) compared with blood ($n = 5617$), and of those genes expressed in blood, 93% (5243/5617) were found in LCL, indicating LCLs provide a similar complement of expressed genes and over 5000 more testable genes (Fig. **1a**). We next explored the expression of 2541 curated NMGs (Table S1). Sixty-seven percent of NMGs (1706/2541 genes) were expressed in LCLs above the expressed gene threshold compared with 36% in blood (917/2541 genes) (Table S1). This equaled 1.8× more NMGs that are testable in LCL compared with blood. Additionally, the relative expression of NMGs was higher in LCLs compared with blood (Fig. **1b**). These analyses indicate LCLs provide a comprehensive transcriptional landscape to perform RNA-seq based diagnostic testing.

Expression of biologically relevant isoforms was determined by comparing with GTEx brain tissue. A total of 17,894 transcript isoforms were found in GTEx for the 1706 LCL expressed NMGs (Table S2). Expression of all 17,894 transcripts correlated between LCL and brain ($r = 0.74$, $p < 2.2\times10^{-16}$) (Fig. **1c**), but only 37% exceeded the expressed gene threshold (6628/17,894), which translates to an observed 3–4 expressed isoforms per gene for LCL. In brain we found similar results with expression of 29% of annotated isoforms (5278/17,894) (Fig. **1c**). LCL expressed isoforms had 63% overlap with brain (4182/6628) (Fig. **1d**), and between these two groups there was 65% overlap for the max expressed transcript (1114/1706). Altogether, there is significant positive correlation in isoform expression between brain and LCLs and overlap in expression of the most biologically relevant transcripts.

Next we investigated what types of genes were represented in the list of 1706 LCL NMGs. Gene Ontology term (GO) analysis for molecular function found significant enrichment for genes associated with catalytic activity ($p = 2.86\times10^{-65}$), nucleotide binding ($p = 7.3\times10^{-34}$), and chromatin binding ($p = 3.03\times10^{-12}$) (Table S3). These gene functions were reflected in biological processes that showed enrichment for metabolic processes ($p = 5.49\times10^{-35}$), mitochondrion organization (1.07×10^{-22}), chromosome organization ($p = 8.64\times10^{-24}$), and nervous system development ($p = 2.18\times10^{-21}$) (Table S3). Examining GO term enrichment for 835 NMGs that did not meet the expressed gene threshold in LCLs showed enrichment for terms such as ion channel ($p = 3.72\times10^{-42}$) and neurotransmitter receptor activities ($p = 1.40\times10^{-21}$) (Table S4). Thus, LCLs are appropriate specimens for diagnostic testing for a large variety of NMGs participating in many core cell processes, but display limitations for testing transcripts with highly restricted neural lineage expression.

A clinical diagnosis for a Mendelian disorder is based on a recognizable pattern of phenotypic features and is typically due to a pathogenic variant in a single gene. Phenotype–gene relationships are curated into phenotypic supersets (PS) in OMIM[23] (Figure S1). From the 1706 NMGs expressed in LCLs, 1361 were involved in 2008 OMIM PS. After removing duplicate PS, there was a total of 1085 unique PS, of which 788 were represented by one gene and 297 by more than one gene (Table S5). RNA-seq from LCLs detected expression of all genes in 61/297 PS (20.5%) that had more than one disease-causing gene (Fig. **1e**). Examples of PS with full gene representation included congenital disorders of glycosylation, type II (PS212066, 17 genes); mitochondrial complex III deficiency (PS124000, 9 genes); Aicardi–Goutieres syndrome (PS225750, 7 genes); and CdLS (PS122470, 5 genes) (Table S5). The cumulative number of PS with 50–100% of genes represented by RNA-seq was 177/297 (60%) (Fig. **1e**). Combining the 788 PS represented by a single gene with 61 PS that have all genes expressed results in complete gene representation in 78% (849/1085) of OMIM PS (Fig. **1f**). Overall, LCLs express hundreds of single genes that cause neurodevelopmental disorders as well as coexpress many complete sets of genes that comprise OMIM phenotypic series.

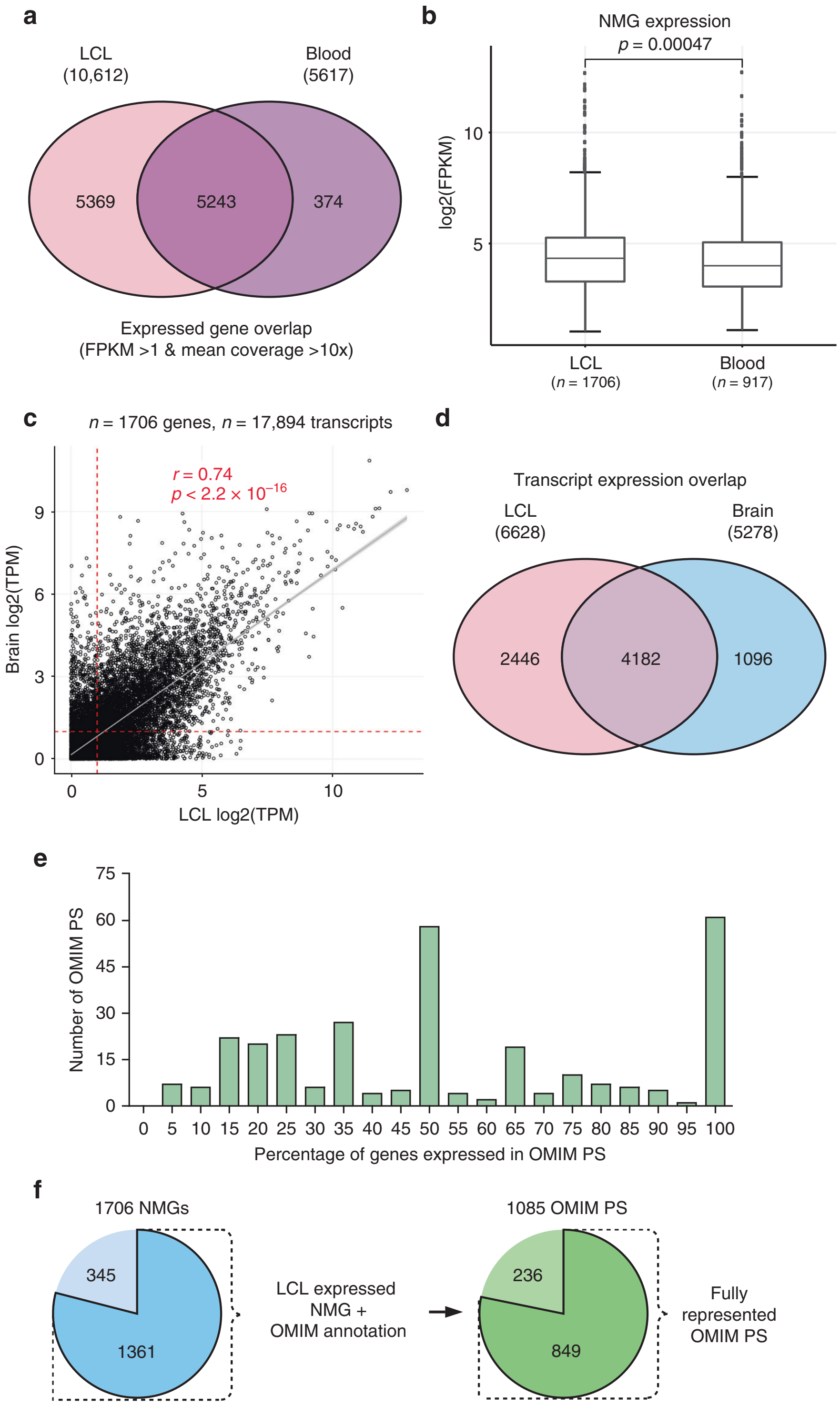

Fig. 1 Genes involved in neurodevelopmental Mendelian disorders are expressed in B-lymphoblastoid cell lines (LCLs). (**a**) Total number and overlap of expressed transcripts in Genotype–Tissue Expression sequencing project (GTEx) LCL and blood samples meeting the expressed gene threshold of fragments per kilobase of transcript per million mapped reads (FPKM) >1 and mean coverage >10×. (**b**) Mean expression of neurodevelopmental Mendelian genes (NMGs) in LCL ($n = 1706$ genes) and blood ($n = 917$ genes). Data shown as mean ± SEM. (**c**) Correlation of isoform expression between GTEx LCL and brain. (**d**) Overlap of transcript isoforms that meet the expressed gene threshold in brain and LCL. (**e**) Histogram depicting number of OMIM phenotypic supersets (PS) with partial or complete gene coverage in LCLs from a total of 297 PS. (**f**) Charts depicting the number of LCL expressed NMGs ($n = 1706$) that have a documented disease–gene relationship in OMIM (dark blue) and the subsequent breakdown of these genes into OMIM PS ($n = 1085$) with complete (dark green) or partial representation (light green).

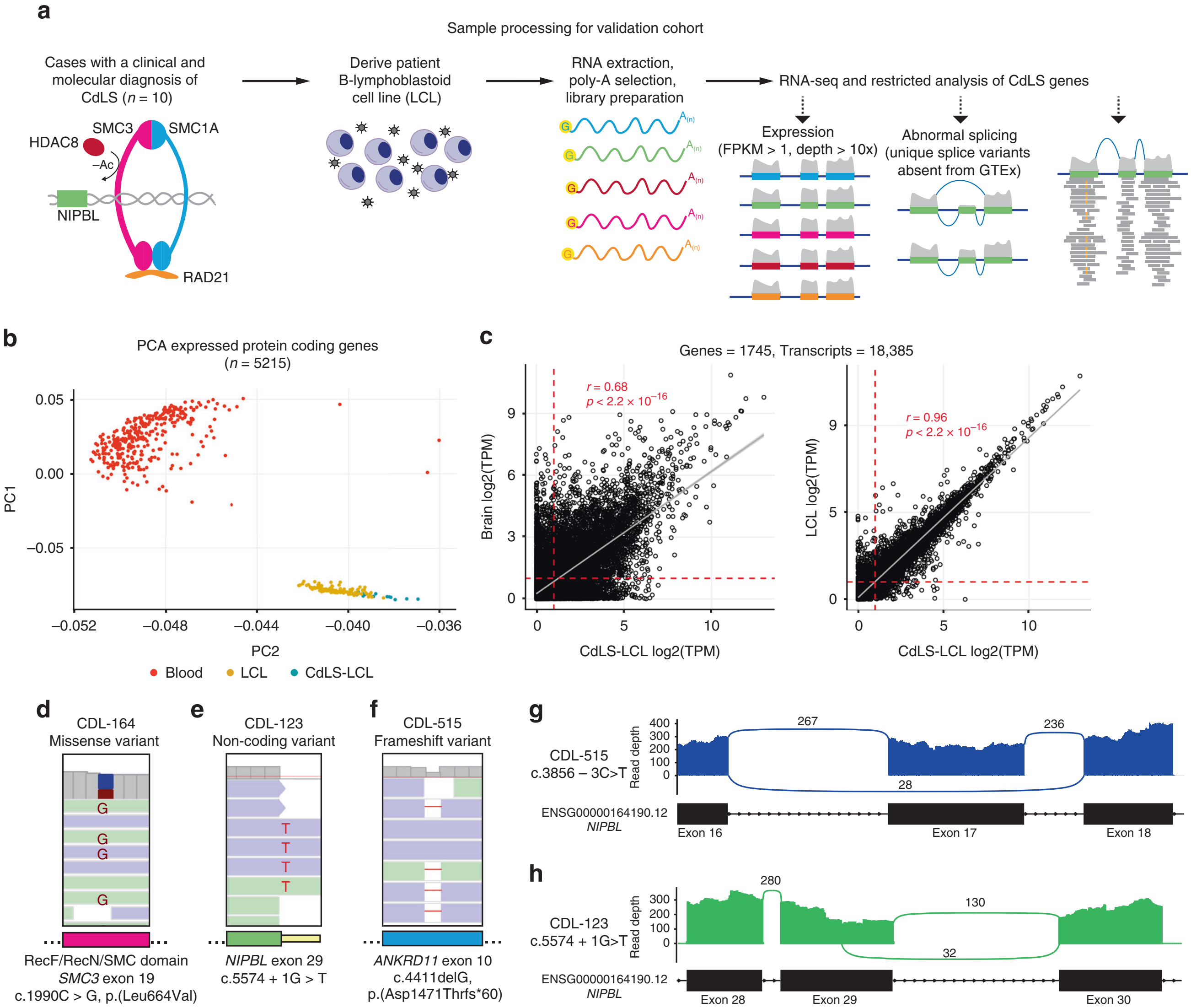

Fig. 2 B-lymphoblastoid cell lines (LCL) RNA-seq identifies pathogenic variants and abnormal splicing in a validation cohort of Cornelia de Lange syndrome (CdLS) patients. (**a**) Schematic of sample processing and RNA-seq analysis performed on the validation cohort of patient LCLs. (**b**) Principal component analysis (PCA) comparing similarity in gene expression profiles between Genotype–Tissue Expression sequencing project (GTEx) blood, GTEx LCL, and CdLS-LCL. (**c**) Correlation between CdLS-LCL and GTEx brain (left panel) and CdLS-LCL and GTEx LCL (right panel) for all annotated transcript isoforms belonging to 1745 neurodevelopmental Mendelian genes (NMGs). (**d**–**f**) Representative images of RNA-seq reads capturing pathogenic *SMC3* missense variant (**d**), *NIPBL* splice-site variant (**e**), and *ANKRD11* frameshift variant (**f**). (**g**,**h**) Sashimi plots of abnormal splicing events causing premature termination of *NIPBL*, numbers indicate supporting exon junction reads.

We next used this information to develop a diagnostic RNA-seq pipeline for LCL specimens from patients with Mendelian disorders (Fig. 2a). Our protocol uses transcriptome data and bioinformatic filters for expressed transcripts containing deleterious coding and noncoding variants (+/−10 bp intron) and abnormal splicing events. Pipeline validation was performed on LCLs made from patients with CdLS, since all genes comprising the CdLS PS are expressed in normal LCLs (Table S5). The validation cohort consisted of ten CdLS-LCL specimens with various pathogenic variants in four CdLS genes (*NIPBL*, *SMC1A*, *SMC3*, and *HDAC8*) (Table 1). Gene expression profiles of CdLS-LCLs were compared with GTEx blood, brain, and LCL samples. Principal component analysis showed GTEx LCL and CdLS-LCL groups cluster together relative to blood indicating similarity in their global expression profiles (Fig. 2b; Figure S2a). CdLS-LCL expression of 18,385 transcript isoforms from 1745 NMGs meeting the expressed gene threshold (1745/2541 NMGs) showed positive correlation with brain ($r = 0.68$, $p < 2.2\times10^{-16}$) and very strong correlation with GTEx LCL ($r = 0.96$, $p < 2.2\times10^{-16}$), with 89.9% of genes having the same highest expressed isoform (Fig. 2c; Figure S2b). In line with correlated expression with GTEx LCL, there was 98% similarity in the number of genes captured per OMIM PS (1050/1071) (Figure S2c; Table S5). Thus, the measurable expression landscape for diagnostic

Table 1 RNA-seq results for CdLS-LCL validation cohort

Patient and variant data							RNA-seq results	
Patient	Sex	Ethnicity	Phenotype	Pathogenic variant	Variant type	Method	Abnormal mRNA detection	Pathogenic variant detection
CDL-27571	Female	Caucasian	Mild	*NIPBL* c.358+5G>C	Splice site	Sanger	*NIPBL* exon 4 skipping	–[a]
CDL-1682	Male	Caucasian	Moderate	*NIPBL* c.7861–1G>C	Splice site	CSGE/ Sanger	*NIPBL* new splice acceptor site exon 46	+
CDL-075	Male	Hispanic—Puerto Rican	Severe	*NIPBL* c.868+1G>A	Splice site	Sanger	*NIPBL* new splice donor site exon 8	+
CDL-123	Female	Caucasian	Moderate	*NIPBL* c.5574+1G>T	Splice site	CSGE/ Sanger	*NIPBL* new splice donor site exon 29	+
CDL-217	Female	Hispanic—Puerto Rican	Moderate	*NIPBL* c.64+1G>A	Splice site	CSGE/ Sanger	*NIPBL* exon 2 skipping	+
CDL-515	Female	N.R.	N.R.	*NIPBL* c.3856–3C>T	Splice site	Sanger	*NIPBL* exon 17 skipping and *ANRKD11* c.4411delG (p.Asp1471Thrfs*60)	+
CDL-223	Female	Hispanic—Puerto Rican	Severe	rsa[hg19] 5p13.2 (36,953,825–37,006,605)×1	*NIPBL* 52-kb deletion including exons 2–17	MLPA	*NIPBL* exons 2–17 skipping	–
CDL-164	Male	N.R.	N.R.	*SMC3* c.1990C>G, p.(Leu664Val)	Missense	Sanger	–	+
CDL-634	Female	Caucasian	Moderate	*SMC1A* c.2394dupA, p.(Arg799Thrfs*4)	Frameshift	Exome	–	+
CDL-418	Female	N.R.	N.R.	*HDAC8* c.881G>A, p.(Trp294*)	Nonsense	Sanger	–	–

CdLS Cornelia de Lange syndrome, *CSGE* conformation sensitive gel electrophoresis, *LCL* B-lymphoblastoid cell line, *MLPA* mixed ligation-dependent probe amplification, *mRNA* messenger RNA, *N.R.* not reported

[a] Insufficient coverage

testing did not greatly differ between control GTEx LCL and CdLS-LCL groups.

Variant calling in CdLS-LCLs was restricted to an RNA-seq panel of 14 genes causing cohesinopathy-related disorders (Table S6; Figure S3). For the highest expressed transcripts, read coverage of >5× was found across 92% of coding base pairs and 18% of bases at intron–exon boundaries (+/−10 bp) (Table S7, S8, S9, S10). This allowed detection of pathogenic coding variants in 2/3 samples and pathogenic splice-site variants in 5/6 samples (Fig. **2d, e**, Table **1**). We found that our capacity to identify pathogenic intronic splice-site variants was due to capturing a minor fraction of transcripts that encoded whole intron read-throughs (Figure S4a, b). An additional 20 unique variant calls required manual review including 16 splice artifacts (e.g., splice-region insertion of sequence from the adjacent exon), 2 false positive coding artifacts (e.g., variant with low variant allele fraction in multiple samples), and 2 true positive coding variants. Interestingly, one of these true positive variants is a frameshift in *ANKRD11* that causes KBG syndrome[24] (Fig. **2f**, Table **1**). Thus, patient CDL-515 has a pathogenic splice-region variant in *NIPBL* (c.3856–3C>T) and an *ANKRD11* loss-of-function variant expanding their diagnosis to two syndromes with overlapping clinical features.[25] In patient CDL-418, we were unable to detect the *HDAC8* p.Trp294* nonsense variant likely due to nonsense-mediated decay.[26] This represents a limitation of RNA-seq variant calling.

The other key feature of our pipeline is identifying abnormal splicing events. Fourteen cohesinopathy-related genes were filtered for unique events by comparing within the CdLS-LCL cohort and to GTEx blood and LCL ($n = 442$) to remove batch effects and normal splicing variation (Figure S5). Abnormal splice products were detected in 7/7 patient LCL with known *NIPBL* splice-region variants or multiexon deletion with no false positive splice events called (Figure S6, Table **1**). Major outcomes included exon skipping and utilization of new donor and acceptor splice sites (Fig. **2g, h**, Figures S6). Overall, our RNA-seq pipeline showed 90% sensitivity for calling pathogenic events (9/10 pathogenic variants and/or abnormal splice products).

To test the utility of our approach on unsolved cases, we performed RNA-seq on five patients with moderate to severe clinical presentations of a suspected cohesinopathy and nondiagnostic genetic testing (Table **2**). Assessment with our 14-gene panel revealed a positive result in 3/5 cases. Patient CDL-679 had severe CdLS presentation and partial skipping of *NIPBL* exons 33 and 34 with premature termination when exon 32 was linked to 34 and a truncated inframe product when spliced to exon 35 (Fig. **3a**). Patient CDL-022 had a c.1038G>C, p.Lys346Asn missense variant in exon 6 of *BRD4* (Fig. **3b**), a gene recently found to cause CdLS-like phenotypes.[27] Evidence to support variant pathogenicity includes amino acid conservation and deleterious computational predictions (ClinPred and SIFT),[28,29] absence from gnomAD (Karczewski et al., unpublished), and occurrence in the second bromodomain, which has previously been shown

Table 2 CdLS-LCL test cohort RNA-seq results for 14 cohesinopathy genes

Patient data and prior negative genetic testing									RNA-seq results		
Sample	Gender	Ethnicity	Phenotype	Cytogenetics	*NIPBL*	*SMC3*	*SMC1A*	*HDAC8*	Result	TP pass filter variants[a]	FP pass filter variants[a]
CDL-679	Male	N.R.	Severe	Karyotype	Seq	Seq	Seq	–	*NIPBL* exons 33 and 34 skipping	0	3
CDL-614	Male	Caucasian	Moderate	Chromosome microarray	Seq/ MLPA	Seq	Seq	Seq	*NIPBL* cryptic exon inclusion from intron 21	0	1
CDL-069	Male	African American	Severe	–	Seq/ MLPA	Seq	Seq	Seq	–	0	0
CDL-086	Male	Caucasian	Severe	–	Seq/ MLPA	Seq	Seq	–	–	0	3
CDL-022	Male	Caucasian	Moderate	–	Seq/ MLPA	Seq	Seq	Seq	*BRD4* c.1038G>C, p.(Lys346Asn)	1	1

CdLS Cornelia de Lange syndrome, *FP* false positive, *LCL* B-lymphoblastoid cell line, *MLPA* mixed ligation-dependent probe amplification, *N.R.* not reported, *Seq* gene sequencing, *TP* true positive

[a] Includes coding and splice-region variants

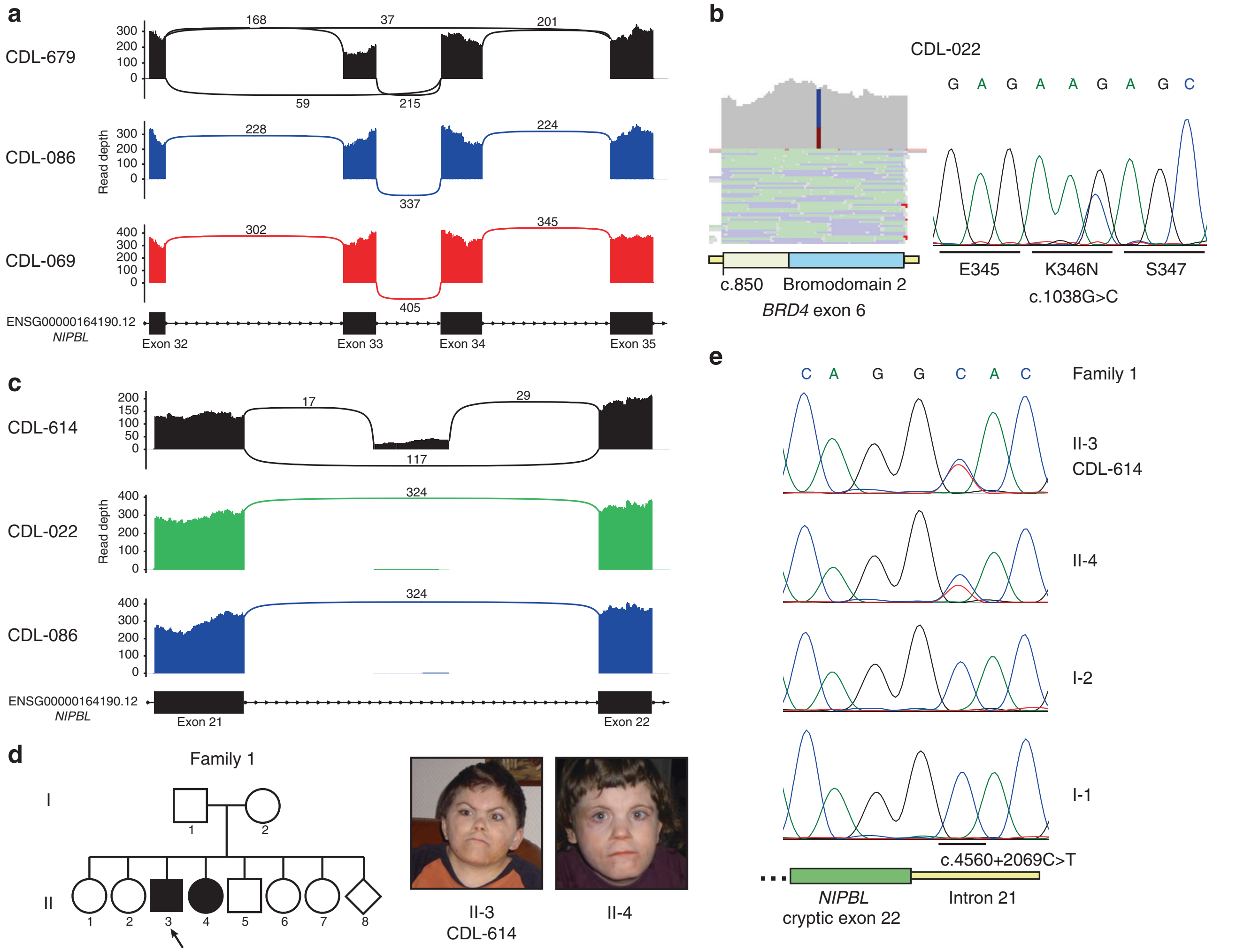

Fig. 3 B-lymphoblastoid cell line (LCL) RNA-seq reveals abnormal splicing and pathogenic variants in patients with only clinical diagnoses of cohesinopathy. (**a**) Sashimi plots from test cohort patients showing abnormal *NIPBL* splicing in patient CDL-679 relative to other samples. (**b**) RNA-seq reads (left) and Sanger confirmation of complementary DNA (cDNA) (right) of a likely damaging *BRD4* missense variant. (**c**) Sashimi plots from test cohort patients showing abnormal *NIPBL* splicing in patient CDL-614. (**d**) Left panel shows family pedigree for patient CDL-614 (proband indicated by arrow), right panel shows images of proband and affected sibling with Cornelia de Lange syndrome (CdLS). (**e**) Locus specific long-range polymerase chain reaction (LR-PCR) followed by Sanger sequencing of *NIPBL* intron 21 in proband, affected sibling, and parents of family 1.

to harbor severe disease-causing variants.[27] Although the patient's mother did not carry this variant, the father was unavailable to confirm de novo status. The third positive patient, CDL-614, was diagnosed with moderate–severe CdLS and had abnormal *NIPBL* splicing with inclusion of a cryptic exon found in intron 21 that is expected to introduce a premature stop codon (Fig. 3c). The de novo exon sits within a segmental duplication that has nearly identical homology to two intergenic regions on chromosomes 17 and 18 (Figure S7a, b). Interestingly, this patient is one of eight children, of whom one sibling also has a clinical diagnosis of CdLS and the other six siblings and parents are healthy (Fig. 3d). This family had previously been evaluated in linkage studies nearly 20 years ago.[30] Locus-specific long-range PCR enabled sequencing of the region containing the cryptic exon. This showed a heterozygous c.4560+2069C>T (chr5:37012396C>T) deep-intronic variant in the proband and affected sibling but not the parents (Fig. 3e). This variant was absent in gnomAD, and splice-site prediction algorithms confirmed it strongly induces formation of a novel 5' splice donor site (Table S11). We suspect parental germline mosaicism for this variant given there are two affected siblings with the same de novo deep-intronic variant.

DISCUSSION

New approaches are needed to find pathogenic events in Mendelian disorders because conventional testing by gene panel and ES leaves many patients undiagnosed. In this study, we show LCLs made from patient blood share the transcriptional repertoire of brain tissue for a significant proportion of NMGs, enabling testing of more than 1000 genetic syndromes. This analysis allowed us to create an important new resource of testable genes by RNA-seq for clinical practice (see Tables S2 and S5). Our analysis shows the LCL transcriptome compared with blood displays elevated NMG expression and has nearly twofold more total genes expressed above our set threshold. We suspect the limited transcriptional landscape of terminally differentiated blood cells and the heterogeneity of sampled cell populations[31] limit the number of testable genes in whole blood. This altogether shows that LCLs are superior to blood for diagnostic testing. We further found a strong similarity between GTEx LCL control and CdLS-LCL gene expression data sets. This indicates the act of B-cell transformation normalizes gene expression and makes a fairly homogeneous testing background. Therefore, while many genes are known to be differentially expressed between control LCL and CdLS-LCL,[32] the milieu of testable transcripts above our expression threshold remained nearly identical. This buffering effect is likely a robust phenomenon given that cohesinopathy gene perturbation is known to affect global gene expression profiles.[32–34] Based on these findings, we hypothesize that LCL from other genetic disorders will have a nearly identical testing background to normal LCL, which is important for clinicians and laboratory directors who need to know which genes can be reliably detected by this technique. Despite these observed benefits, the drawback of using LCLs is the time and costs required to establish cell lines, which take about one month to produce sufficient quantities of cells for RNA-seq and long-term storage. Additionally, prolonged culturing can yield genomic artifacts,[35,36] thus all testing should be done on early passage cell lines, which show negligible genetic changes compared with primary tissue.[20,35]

Using a combination of (1) variant calling and (2) abnormal splice pattern detection, we obtained 90% sensitivity for all pathogenic events in the validation cohort. Breakdown of these two features shows all abnormal splice products were detected without calling false positives, whereas variant calling missed two SNVs (false negatives), including a nonsense variant in *HDAC8* and a +5 splice-region variant in *NIPBL*. Dropout of the *HDAC8* p.Trp294* variant on Xq13.1 is likely due to nonsense-mediated decay resulting in only reference allele expression.[26] Other studies have utilized allelic imbalance to identify pathogenic events across the transcriptome;[8,9] however, we find challenges with implementing this strategy in a clinical setting. One difficulty is constraint in variation in haploinsufficient genes,[37–39] which can result in an absence of heterozygous variants in clinically significant genes within individuals. We noticed this phenomenon with *HDAC8* where no coding region variants were detected in our validation cohort. Additionally, we found the *HDAC8* nonsense variant did not significantly alter gene expression (Figure S3). Thus, allele imbalances and expression changes can be useful as a screening strategy to prioritize genes carrying pathogenic variants across the transcriptome, but their interpretation should be treated with caution given the absence of these differences does not preclude an actual pathogenic variant is present. The second variant that was not detected was *NIPBL* c.358+5G>C. Our pipeline was designed to find pathogenic intronic splicing variants by retaining reads from exon read-throughs. This approach led to detection of low levels of abnormal transcripts that retained introns for sequencing. We believe detecting these low abundant transcripts that retained introns indicates the efficiency of exon skipping and cryptic splice-site activation is incomplete upon inactivating a native splice site. Although the +5 pathogenic splice variant in CDL-27571 was not observed due to lack of coverage at this intronic position, deeper sequencing could be one strategy to boost detection of these low abundant exon read-through events. Retaining intronic reads for analysis of pathogenic splice variants is especially important if RNA-seq is used before testing by exome or genome sequencing since it identifies the causal genomic variant.

LCL RNA-seq testing on five patients who had gone several years without genetic diagnosis for their clinical presentation of CdLS or related cohesinopathy led to identification of two abnormal *NIPBL* splice events and one missense variant in *BRD4*. One of the abnormal *NIPBL* splice events resulted in the inclusion of a cryptic exon and represents a novel mechanism of *NIPBL* gene inactivation. Due to the high homology of the region containing the de novo deep-intronic pathogenic variant, other approaches such as genome sequencing would not have been able to detect this variant due to low mapping quality reads. The similarity of phenotypes in patients with pathogenic variants in *BRD4* and *NIPBL* reflects work showing BRD4 directly interacts with NIPBL to bind enhancers of developmentally important genes.[27] The *BRD4* variant could have been detected by DNA-based methods had the gene been known at the time of diagnosis and testing; however, our work shows that performing RNA-seq variant calling in concert with abnormal splicing analysis several months to years after primary exome analysis could yield new diagnoses in similar fashion to exome reanalysis workflows (albeit with more limited scope). Further supporting the utility of RNA-seq variant calling is finding the *ANKRD11* frameshift resulting in a dual diagnosis for patient CDL-515. The two remaining patients who did not receive a diagnosis by analyzing 14 CdLS-related genes suggests unknown cohesinopathy genes could be involved. In conclusion, our work illustrates the large testable transcriptional landscape of patient LCLs for neurodevelopmental disorders and shows the specific application of LCL RNA-seq in providing new diagnoses in patients with CdLS.

SUPPLEMENTARY INFORMATION

The online version of this article (https://doi.org/10.1038/s41436-019-0741-5) contains supplementary material, which is available to authorized users.

ACKNOWLEDGEMENTS

We thank the families involved in this study and Nancy Spinner and Deborah McEldrew for their technical support. This work was supported by internal departmental funds.

DISCLOSURE

The authors declare no conflicts of interest.

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

REFERENCES

1. Retterer K, Juusola J, Cho MT, et al. Clinical application of whole-exome sequencing across clinical indications. Genet Med. 2016;18:696–704.
2. Boycott KM, Hartley T, Biesecker LG, et al. A diagnosis for all rare genetic diseases: the horizon and the next frontiers. Cell. 2019;177:32–37.
3. Ewans LJ, Schofield D, Shrestha R, et al. Whole-exome sequencing reanalysis at 12 months boosts diagnosis and is cost-effective when applied early in Mendelian disorders. Genet Med. 2018;20:1564–1574.
4. Baker SW, Murrell JR, Nesbitt AI, et al. Automated clinical exome reanalysis reveals novel diagnoses. J Mol Diagn. 2019;21:38–48.
5. Stenson PD, Mort M, Ball EV, et al. The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. Hum Genet. 2017;136:665–677.
6. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10:57–63.
7. Piskol R, Ramaswami G, Li JB. Reliable identification of genomic variants from RNA-seq data. Am J Hum Genet. 2013;93:641–651.
8. Fresard L, Smail C, Ferraro NM, et al. Identification of rare-disease genes using blood transcriptome sequencing and large control cohorts. Nat Med. 2019;25:911–919.
9. Gonorazky HD, Naumenko S, Ramani AK, et al. Expanding the boundaries of RNA sequencing as a diagnostic tool for rare Mendelian disease. Am J Hum Genet. 2019;104:466–483.
10. Cummings BB, Marshall JL, Tukiainen T, et al. Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. Sci Transl Med. 2017;9:eaal5209.
11. Kremer LS, Bader DM, Mertes C, et al. Genetic diagnosis of Mendelian disorders via RNA sequencing. Nat Commun. 2017;8:15824.
12. Sie L, Loong S, Tan EK. Utility of lymphoblastoid cell lines. J Neurosci Res. 2009;87:1953–1959.
13. Kline AD, Moss JF, Selicorni A, et al. Diagnosis and management of Cornelia de Lange syndrome: first international consensus statement. Nat Rev Genet. 2018;19:649–666.
14. Vivian J, Rao AA, Nothaft FA, et al. Toil enables reproducible, open source, big biomedical data analyses. Nat Biotechnol. 2017;35:314–316.
15. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15–21.
16. GT Consortium. The Genotype–Tissue Expression (GTEx) project. Nat Genet. 2013;45:580–585.
17. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20:1297–1303.
18. Van der Auwera GA, Carneiro MO, Hartl C, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics. 2013;43:11 10 11–33.
19. Cingolani P, Platts A, Wang le L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin). 2012;6:80–92.
20. Nickles D, Madireddy L, Yang S, et al. In depth comparison of an individual's DNA and its lymphoblastoid cell line using whole genome sequencing. BMC Genomics. 2012;13:477.
21. Herbeck JT, Gottlieb GS, Wong K, et al. Fidelity of SNP array genotyping using Epstein Barr virus-transformed B-lymphocyte cell lines: implications for genome-wide association studies. PLoS ONE 2009;4:e6915.
22. Londin ER, Keller MA, D'Andrea MR, et al. Whole-exome sequencing of DNA from peripheral blood mononuclear cells (PBMC) and EBV-transformed lymphocytes from the same donor. BMC Genomics. 2011; 12:464.
23. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res. 2005;33:D514–517.
24. Sirmaci A, Spiliopoulos M, Brancati F, et al. Mutations in ANKRD11 cause KBG syndrome, characterized by intellectual disability, skeletal malformations, and macrodontia. Am J Hum Genet. 2011;89:289–294.
25. Ansari M, Poke G, Ferry Q, et al. Genetic heterogeneity in Cornelia de Lange syndrome (CdLS) and CdLS-like phenotypes with observed and predicted levels of mosaicism. J Med Genet. 2014;51:659–668.
26. Kurosaki T, Popp MW, Maquat LE. Quality and quantity control of gene expression by nonsense-mediated mRNA decay. Nat Rev Mol Cell Biol. 2019;20:406–420.
27. Olley G, Ansari M, Bengani H, et al. BRD4 interacts with NIPBL and BRD4 is mutated in a Cornelia de Lange-like syndrome. Nat Genet. 2018;50:329–332.
28. Alirezaie N, Kernohan KD, Hartley T, Majewski J, Hocking TD. ClinPred: prediction tool to identify disease-relevant nonsynonymous single-nucleotide variants. Am J Hum Genet. 2018;103:474–483.
29. Sim NL, Kumar P, Hu J, Henikoff S, Schneider G, Ng PC. SIFT web server: predicting effects of amino acid substitutions on proteins. Nucleic Acids Res. 2012;40:W452–457.
30. Krantz ID, Tonkin E, Smith M, et al. Exclusion of linkage to the CDL1 gene region on chromosome 3q26.3 in some familial cases of Cornelia de Lange syndrome. Am J Med Genet. 2001;101:120–129.
31. Novershtern N, Subramanian A, Lawton LN, et al. Densely interconnected transcriptional circuits control cell states in human hematopoiesis. Cell. 2011;144:296–309.
32. Liu J, Zhang Z, Bando M, et al. Transcriptional dysregulation in NIPBL and cohesin mutant human cells. PLoS Biol. 2009;7:e1000119.
33. Newkirk DA, Chen YY, Chien R, et al. The effect of Nipped-B-like (Nipbl) haploinsufficiency on genome-wide cohesin binding and target gene expression: modeling Cornelia de Lange syndrome. Clin Epigenetics. 2017;9:89.
34. Kagey MH, Newman JJ, Bilodeau S, et al. Mediator and cohesin connect gene expression and chromatin architecture. Nature. 2010;467:430–435.
35. Oh JH, Kim YJ, Moon S, et al. Genotype instability during long-term subculture of lymphoblastoid cell lines. J Hum Genet. 2013;58:16–20.
36. Joesch-Cohen LM, Glusman G. Differences between the genomes of lymphoblastoid cell lines and blood-derived samples. Adv Genomics Genet. 2017;7:1–9.
37. Lek M, Karczewski KJ, Minikel EV, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536:285–291.
38. Amr SS, Al Turki SH, Lebo M, Sarmady M, Rehm HL, Abou Tayoun AN. Using large sequencing data sets to refine intragenic disease regions and prioritize clinical variant interpretation. Genet Med. 2017;19:496–504.
39. Havrilla JM, Pedersen BS, Layer RM, Quinlan AR. A map of constrained coding regions in the human genome. Nat Genet. 2019;51:88–95.