Genomics 114 (2022) 110468

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Pathogenic variants detected by RNA sequencing in Cornelia de Lange syndrome

Rie Seyama [a,b], Yuri Uchiyama [a,c], José Ricard Magliocco Ceroni [d], Veronica Eun Hue Kim [d], Isabel Furquim [d], Rachel Sayuri Honjo [d], Matheus Augusto Araujo Castro [d], Lucas Vieira Lacerda Pires [d], Hiromi Aoi [b], Kazuhiro Iwama [a,e], Kohei Hamanaka [a], Atsushi Fujita [a], Naomi Tsuchida [a,c], Eriko Koshimizu [a], Kazuharu Misawa [a], Satoko Miyatake [a,f], Takeshi Mizuguchi [a], Shintaro Makino [g], Atsuo Itakura [b], Débora R. Bertola [d], Chong Ae Kim [d], Naomichi Matsumoto [a,*]

[a] Department of Human Genetics, Yokohama City University Graduate School of Medicine, Yokohama, Japan
[b] Department of Obstetrics and Gynecology, Juntendo University, Tokyo, Japan
[c] Department of Rare Disease Genomics, Yokohama City University Hospital, Yokohama, Japan
[d] Genetics Unit, Instituto da Crianca, Faculdade de Medicina, Universidade de Sao Paulo, Brazil
[e] Department of Neonatal Medicine, Yokohama City University Medical Center, Yokohama, Japan
[f] Department of Clinical Genetics, Yokohama City University Hospital, Yokohama, Japan
[g] Department of Obstetrics and Gynecology, Juntendo University Urayasu Hospital, Urayasu, Japan

ARTICLE INFO

Keywords:
Aberrant splicing
Cornelia de Lange syndrome
Genetic analysis
RNA sequencing
Whole exome sequencing

ABSTRACT

Recent studies suggest that transcript isoforms significantly overlap (approximately 60%) between brain tissue and Epstein–Barr virus-transformed lymphoblastoid cell lines (LCLs). Interestingly, 14 cohesion-related genes with variants that cause Cornelia de Lange Syndrome (CdLS) are highly expressed in the brain and LCLs. In this context, we first performed RNA sequencing of LCLs from 22 solved (with pathogenic variants) and 19 unsolved (with no confirmed variants) CdLS cases. Next, an RNA sequencing pipeline was developed using solved cases with two different methods: short variant analysis (for single-nucleotide and indel variants) and aberrant splicing detection analysis. Then, 19 unsolved cases were subsequently applied to our pipeline, and four pathogenic variants in *NIPBL* (one inframe deletion and three intronic variants) were newly identified. Two of three intronic variants were located at *Alu* elements in deep-intronic regions, creating cryptic exons. RNA sequencing with LCLs was useful for identifying hidden variants in exome-negative cases.

1. Introduction

Exome sequencing (ES) is a standard method for identifying the genetic causes of Mendelian disorders [1,2]; however, because its genetic solution rate is only 30%–40% [3], other approaches for finding pathological variants should be considered. RNA sequencing (RNA-seq) is widely used to determine the sequences of all transcripts derived from target tissues, and it enables the identification of differentially expressed genes and alternative transcripts [4,5]. Furthermore, RNA-seq pipelines for detecting aberrant splicing events can identify deep-intronic variants that create a cryptic exon [6]. Thus, one advantage of RNA-seq analysis is that it allows the investigation of patient-specific (pathogenic)

Abbreviations: CdLS, Cornelia de Lange Syndrome; dbGaP, Database of Genotypes and Phenotypes; ES, Exome sequencing; FPKM, Fragments Per Kilobase of exon per Million mapped reads; GATK, Genome Analysis Toolkit; GTEx, Genotype-Tissue Expression; hnRNP, Heterogeneous nuclear ribonucleoprotein; IGV, Integrated Genome Viewer; LCLs, Human B-lymphoblastoid cell lines; MBLN1, Muscleblind-like Protein 1; NIPBL, NIPBL cohesin loading factor; NMD, Nonsense-mediated mRNA decay; NMG, Neurodevelopmental Mendelian gene; PCA, Principal component analysis; PCR, Polymerase chain reaction; RNA-seq, RNA sequencing; RT-PCR, Reverse transcription-polymerase chain reaction; Sam68, SRC associated in mitosis of 68 kDa; SD, Standard deviation; SLM-2, Sam68-like mammalian protein 2; SNV, Single-nucleotide variant; TIA-1, TIAT-cell intracellular antigen 1; TIAL-1, TIA1-related/like protein.

* Corresponding author at: Department of Human Genetics Yokohama City, University Graduate School of Medicine, Fukuura 3-9, Kanazawa-ku, Yokohama 236-0004, Japan.
E-mail address: naomat@yokohama-cu.ac.jp (N. Matsumoto).

https://doi.org/10.1016/j.ygeno.2022.110468
Received 23 June 2022; Received in revised form 11 August 2022; Accepted 26 August 2022
Available online 27 August 2022

transcripts. However, RNA-seq analysis of neurodevelopmental disorders is limited by difficulties in obtaining appropriate RNA derived from target neural tissues, such as brain. Fresard et al. showed that approximately 70.6% of disease genes in the Online Mendelian Inheritance in Man (https://omim.org/) were expressed in whole blood, and their subsequent RNA-seq analysis for rare diseases using total RNA derived from patients' whole blood yielded a diagnostic rate of approximately 7.5% [7]. Recently, Rentas et al. showed that the expression of transcripts in brain tissue and human B-lymphoblastoid cell lines (LCLs) was positively correlated with 63% (4182/6628) shared isoforms [8]. Cornelia de Lange Syndrome (CdLS; MIM#122470) is a rare neurodevelopmental disorder with dysmorphic features, and genes related to CdLS are highly expressed in brain tissue and LCLs [8]. Focusing on this feature, Rentas et al. reported the RNA-seq of a CdLS cohort and newly identified pathogenic variants associated with aberrant splicing events [8].

In this study, we performed RNA-seq of 41 CdLS probands previously analyzed by trio-based ES (22 solved and 19 unsolved). We developed a new RNA-seq pipeline for rare diseases using 22 solved CdLS cases as positive controls and demonstrated its advantages by identifying four new pathogenic variants from 19 CdLS cases that were previously unsolved by ES.

2. Methods

2.1. Subjects

Our CdLS cohort consisted of 65 Brazilian patients and one Japanese patient. In total, 41 CdLS patients with available LCLs were selected for RNA-seq, including 22 cases that were genetically solved by ES and 19 that were unsolved (Fig. 1, Table 1, and Supplementary Methods) [9]. The blood samples of Brazilian patients suspected of having CdLS by clinical geneticists based on clinical features were sent to us for genetic diagnosis through the cooperation of the Brazilian Association of Cornelia de Lange Syndrome (CdLS Brazil). One Japanese patient was recruited as a part of the Initiative on Rare and Undiagnosed Diseases project in Japan [10]. The clinical classification of CdLS [11] for the 19 unsolved cases are presented in Table S1. The 22 solved cases harbored three nonsense, six frameshift, three inframe, five missense, and five possible splicing-related variants of genes previously reported as pathogenic for CdLS and other syndromes resembling CdLS (Table 1). We also used the RNA-seq data of LCLs from 105 controls and whole blood from 10 controls in the Genotype-Tissue Expression (GTEx) Project [12]. We downloaded fastq files of LCLs ($n = 105$) and whole blood ($n = 10$) paired-end RNA-seq reads from GTEx via the Database of Genotypes and Phenotypes (dbGaP) (https://www.ncbi.nlm.nih.gov/gap/), accession phs000424.v8.p2.c1, that were sequenced on an Illumina HiSeq 2000 system (Illumina, San Diego, CA, USA) with 76 bp paired-end reads. Permission for downloading the GTEx data was registered with dbGaP as OMB control number 0925–0670.

This study was approved by the Institutional Review Boards of Yokohama City University, Faculty of Medicine, and the University of Sao Paulo, Faculty of Medicine. Written informed consent was obtained from patients or their guardians.

2.2. ES

ES was performed as previously described [13]. The flowchart of the ES analysis is shown in Fig. 1 and detailed methods are provided in Supplementary Methods. Briefly, proband-based exomes for single-nucleotide variants (SNVs) and small indels, and copy-number variation analysis were used to detect genetic causes for known Mendelian disorders, including CdLS. When we could not detect any causative variants, we carried out trio-based ES [14].

2.3. RNA-seq

The overview of our RNA-seq pipeline is presented in Fig. S1. Briefly, the pipeline consists of short variant analysis, aberrant splicing detection analysis, and differential gene expression analysis.

2.3.1. RNA-seq and creating BAM files of the genome and transcriptome

Total RNA was isolated from LCLs derived from all 41 probands using an RNEasy plus mini kit (Qiagen N.V., Hilden, Germany) according to the manufacturer's protocol. All RNA samples satisfied the RNA integrity number cut-off of ≥ 9.5. After poly-A selection capture, a unique

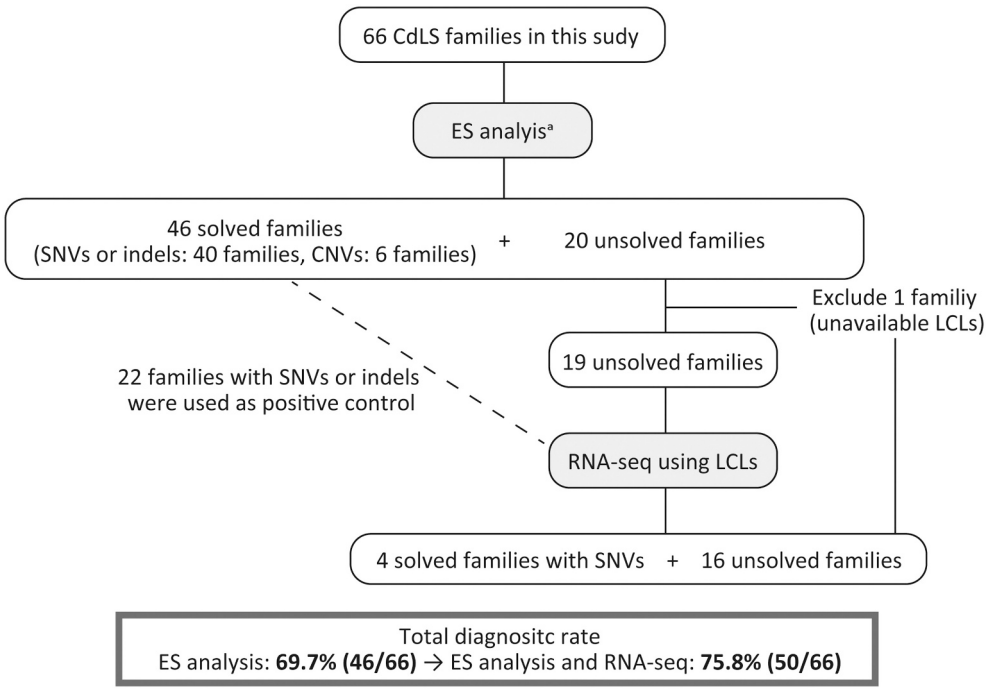

Fig. 1. Flowchart of the exome sequencing and subsequent RNA sequencing analysis pipeline developed in this study.
A total of 66 Cornelia de Lange Syndrome (CdLS) families participated in this study. Proband- or Trio-based exome sequencing (ES) analyses identified 40 pathogenic variants [single-nucleotide variants (SNVs) or small indels] and six copy-number variations (CNVs). Subsequent RNA sequencing (RNA-seq) was performed on the remaining 19 probands that were unsolved by ES but with lymphoblastoid cell lines (LCLs) and 22 probands solved by ES analysis as positive controls, resulting in the identification of four new pathogenic variants in the 19 unsolved CdLS cases. As a result, the total diagnostic rate was increased from 69.7% (46/66) to 75.8% (50/66). [a]Aoi H et al. *J. Hum. Genet.* 64 (2019) 967–978.

Table 1
Details of pathogenic variants in CdLS cases solved by exome sequencing and results of RNA sequencing analysis.

	Family ID	Gene (accession number)	Variant description		Coordinates (hg19)	Detection method and status		Allele depth	Total depth	Clinical classification of CdLS[a]	ACMG/AMP guidelines[b]
			cDNA	Amino acid change		Short variant	Aberrant splicing				
Nonsense	CdLS43	*NIPBL* (NM_133433.4)	c.826C > T	p.Gln276*	chr5:36972101	Detected	–	27	70	Classic	Pathogenic (PVS1, PS1, PS2, PM2, PP4)
	CdLS44	*NIPBL* (NM_133433.4)	c.190C > T	p.Gln64*	chr5:36955699	Detected	–	25	52	Classic	Pathogenic (PVS1, PS1, PS2, PM2, PP4)
	CdLS49	*ANKRD11* (NM_013275.6)	c.5434C > T	p.Gln1812*	chr16:89347516	Detected	–	23	53	Non-classic	Pathogenic (PVS1, PS2, PM2)
Frameshift	CdLS3	*NIPBL* (NM_133433.4)	c.6179dup	p.His2060Glnfs*4	chr5:37044518	Detected	–	5	39	Classic	Pathogenic (PVS1, PS2, PM2, PP4)
	CdLS10	*NIPBL* (NM_133433.4)	c.5174delA	p.Lys1725Serfs*17	chr5:37020719	Undetected	–	0	21	Classic	Pathogenic (PVS1, PM2, PP4)
	CdLS15	*NIPBL* (NM_133433.4)	c.2479_2480del	p.Arg827Glyfs*2	chr5:36985761–36,985,762	Detected	–	10	45	Non-classic	Pathogenic (PVS1, PS1, PS2, PM2, PP3)
	CdLS19	*NIPBL* (NM_133433.4)	c.1903_1904insA	p.Ser635Tyrfs*3	chr5:36985185	Detected	–	7	61	Classic	Pathogenic (PVS1, PS2, PM2, PP3, PP4)
	CdLS23	*ANKRD11* (NM_013275.6)	c.3255_3256del	p.Lys1086Glufs*15	chr16:89349694–89,349,695	Detected	–	14	42	Molecular test	Pathogenic (PVS1, PS1, PM2)
	CdLS27	*NIPBL* (NM_133433.4)	c.5030_5031del	p.Ile1677Serfs*21	chr5:37020576–37,020,577	Undetected	–	1	65	Classic	Pathogenic (PVS1, PS2, PM2, PP3, PP4)
Inframe	CdLS6	*EP300* (NM_001429.4)	c.7014_7028del	p.His2338_Pro2342del	chr22:41574724–41,574,738	Detected	–	29	104	Non-classic	Variant uncertain significance (PM2, PM4)
	CdLS35	*NIPBL* (NM_133433.4)	c.6653_6655del	p.Asn2218del	chr5:37048661–37,048,663	Detected	–	27	39	Classic	Pathogenic (PS2, PM1, PM2, PM4, PP4)
	CdLS36	*SMARCA4* (NM_001128849.3)	c.2519_2542delinsGGA	p.Ala840_Leu848delinsGlyIle	chr19:11130280–11,130,303	Undetected	–	0	130	Molecular test	Pathogenic (PS2, PM1, PM2, PM4)
Missense	CdLS8	*NIPBL* (NM_133433.4)	c.6620 T > C	p.Met2207Thr	chr5:37048634	Detected	–	43	102	Classic	Likely pathogenic (PS2, PM2, PP3, PP4)
	CdLS48	*SMC1A* (NM_006306.4)	c.1487G > A	p.Arg496His	chrX:53436051	Detected	–	132	352	Classic	Pathogenic (PS1, PS2, PM1, PM2, PP3, PP4)
	CdLS52	*NIPBL* (NM_133433.4)	c.6027G > C	p.Leu2009Phe	chr5:37038759	Detected	–	43	103	Non-classic	Likely pathogenic (PS2, PM2, PP3)
	CdLS57	*NIPBL* (NM_133433.4)	c.6448C > G	p.Leu2150Val	chr5:37045649	Detected	–	54	120	Non-classic	Pathogenic (PS2, PM1, PM2, PM5, PP3)
	CdLS58	*NIPBL* (NM_133433.4)	c.6893G > A	p.Arg2298His	chr5:37049342	Detected	–	57	104	Molecular test	Pathogenic (PS1, PS2, PM1, PM2, PP3)
Splicing	CdLS17	*NIPBL* (NM_133433.4)	c.3121 + 1G > A	p.Asp499_Lys1040del	chr5:36986404	Detected	Detected	2	2	Classic	

(*continued on next page*)

Table 1 (*continued*)

Family ID	Gene (accession number)	Variant description		Coordinates (hg19)	Detection method and status		Allele depth	Total depth	Clinical classification of CdLS[a]	ACMG/AMP guidelines[b]
		cDNA	Amino acid change		Short variant	Aberrant splicing				
CdLS32	*NIPBL* (NM_133433.4)	c.7410 + 4A > G	p.Lys2422_Glu2470del	chr5:37057438	Undetected	Detected	0	0	Classic	Pathogenic (PVS1, PS1, PS2, PM2, PP4)
CdLS42	*SMC1A* (NM_006306.4)	c.2497-13C > G	p.Lys854_Glu855insCysTrpAspGln	chrX:53423550	Undetected	Detected	0	1	Classic	Pathogenic (PS1, PS2, PM1, PM2, PM4, PP4)
CdLS47	*NIPBL* (NM_133433.4)	c.6343G > T	p.Val2085Profs*5 p.Gly2115Cys	chr5:37044831	Undetected	Detected	23	73	Non-classic	Pathogenic (PVS1, PS2, PM2, PP3)
CdLS60	*NIPBL* (NM_133433.4)	c.5329-15A > G	p.Ile1777_Arg1809del	chr5:37022138	Undetected	Detected	0	0	Molecular test	Pathogenic(PS1, PS2, PM2, PM4, PP3,)

CdLS, Cornelia de Lange syndrome; hg19, human reference genome.

[a] Kline AD et al. *Nat Rev Genet.* 2018;19(10):649–666.

[b] Richards S, et al. *Genet Med.* 2015;17(5):405–424.

dual indexed library was created with an Illumina Truseq Stranded mRNA library kit (Illumina) and sequenced on an Illumina NovaSeq 6000 system (Illumina) using 100 bp paired-end reads. The mean RNA-seq read pairs was 33.4 M. Our stranded mRNA library preparation protocol was similar to that of the GTEx project (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000424.v8.p2). After trimming with Trimmomatic-0.39, all the fastq files of the 41 CdLS samples and reference controls from GTEx registered on dbGAP were aligned to the human reference genome (GRCh37/hg19; https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.13/) and annotated with GENCODE v19 (https://www.gencodegenes.org/human/release_19.html) gene annotations using the STAR aligner (v2.5.4b). Two types of BAM files, one for the reference genome and one for the transcriptome, were created during this process. In the BAM files for the reference genome, each index was created using SAMtools 1.9 [15].

2.3.2. Principal component and correlation analyses of the RNA-seq data

We analyzed BAM files with RSEM v1.3.3 [16] referring to the hg19 reference genome to quantify gene expression [Fragments Per Kilobase of exon per Million mapped reads (FPKM)]. Principal component analysis (PCA) was performed to compare gene expression patterns among GTEx-whole blood, GTEx-LCLs, and CdLS-LCLs. We focused on all the genes ($n = 57{,}783$), the neurodevelopmental Mendelian genes (NMGs) reported by Rentas et al. with a mean FPKM cut-off > 1 among all the samples ($n = 1695$) [8], and 53 CdLS genes (described in section 2.3.3). PCA was performed with prcomp in R v4.1.1. Correlations among the expressed genes (NMGs and 53 CdLS-related genes) were determined between GTEx-LCLs and CdLS-LCLs. The PCA, Pearson correlation coefficient analysis, and scatterplot drawing were performed with R v4.1.1. We performed differential gene expression analysis between CdLS and controls using the RSEM results as described in Supplementary Methods.

2.3.3. Calculation of coverage in coding regions in the RNA-seq data

Sequencing coverage for CdLS cases was calculated using DepthOfCoverage (v3.8.1.0) in the Genome Analysis Toolkit (GATK). The mean RNA-seq read depths in the coding regions of CdLS-related sequences in RefSeq (https://www.ncbi.nlm.nih.gov/refseq/) were from × 34.89 to × 56.75 (Table S2).

2.3.4. Creating a CdLS-related gene list

A CdLS-related gene list was created for the short variant analysis as described in section 2.3.5. The list contains three groups of genes: 1) 14 cohesion-related genes with pathogenic variants found in CdLS and highly expressed in the brain and LCLs [8]; 2) 16 cohesin-related genes [17] with no report of disease-causing variants; and 3) 23 mSWI/SNF complex-related genes [18] (Table S3). These 53 genes had sufficient read depths ($> \times 10$) in each sample (Table S4). In particular, the read depths of 14 cohesin-related genes (the first group) were reasonable, ranging from × 16.18 to × 212.58 (Table S5).

2.3.5. Short variant (SNVs and indels) analysis of the RNA-seq data

According to the GATK best practices for RNA-seq variant calling for the GATK (v4.0.4.0) pipeline, variant calls and subsequent annotation were run against the BAM files aligned to the reference genome by HaplotypeCaller (v4.0.4.0) and annotated with ANNOVAR (http://annovar.openbioinformatics.org/en/latest/), respectively [19]. Briefly, the BAM files were evaluated for their sequencing qualities using Markduplicate and SplitNCigarReads and recalibrated using BaseRecalibrator and ApplyBQSR with the default settings. Variant calling was performed using HaplotypeCaller. The called variants were filtered out by their quality scores with default settings and subsequently annotated with ANNOVAR together with our in-house exome data of 575 healthy Japanese controls [20]. Detected variants were evaluated using our filtering criteria (see section 3.2.1), and candidate variant lists were created for each individual. After the filtration, each variant was confirmed in the

patients together with the normal controls by manual inspection using the Integrated Genome Viewer (IGV) v2.11.4.

The following variants were excluded from our list of 53 genes using a filtering process similar to the one we used for our ES analysis in an autosomal dominant fashion: 1) variants registered in the public databases and/or our in-house 575 Japanese healthy control database; 2) synonymous variants; 3) homozygous variants; 4) variants > 30 bp in length; and 5) splicing variants (Fig. S2). The remaining variants were evaluated for pathogenicity with our ES analysis pipeline (see Supplementary Methods) [13].

2.3.6. *Aberrant splicing detection analysis of the RNA-seq data*

LeafCutterMD 0.2.9 (https://github.com/davidaknowles/leafcutter/) was used to explore aberrant splicing events with BAM files for the reference genome of CdLS and controls [21,22]. The minimum coverage threshold and the maximum cluster length were used to detect all true positive aberrant splicing events (see section 3.4.2). The detected variants were evaluated using SpliceAI [23] and confirmed by manual inspection of Sashimi plots with IGV v2.11.4. Changes in the binding of proteins to transcripts caused by SNVs involved in aberrant splicing events were evaluated using SpliceAid2 [24].

2.4. *Variant confirmation with both genomic DNA and cDNA*

Detected candidate variants were confirmed at genomic DNA and cDNA levels. For the genomic DNA, trio-based Sanger sequencing was performed to clarify familial segregations. Candidate variants involved in aberrant splicing events were confirmed by reverse transcription-polymerase chain reaction (RT-PCR) using total RNA extracted from proband-derived lymphoblastoid cells. RT-PCR products were Sanger sequenced to confirm the aberrant transcripts. For complicated splicing events, TA-cloning and subsequent Sanger sequencing were performed to obtain the precise alternative transcript sequences. The biological parentages were confirmed by trio-based ES analysis in all 19 ES-unsolved families. The detailed RNA extraction conditions, RT-PCR, and TA-cloning are described in Supplementary Methods. Primer sequences are available on request. The locations of the primer sets on the DNA sequences are shown in Fig. S3.

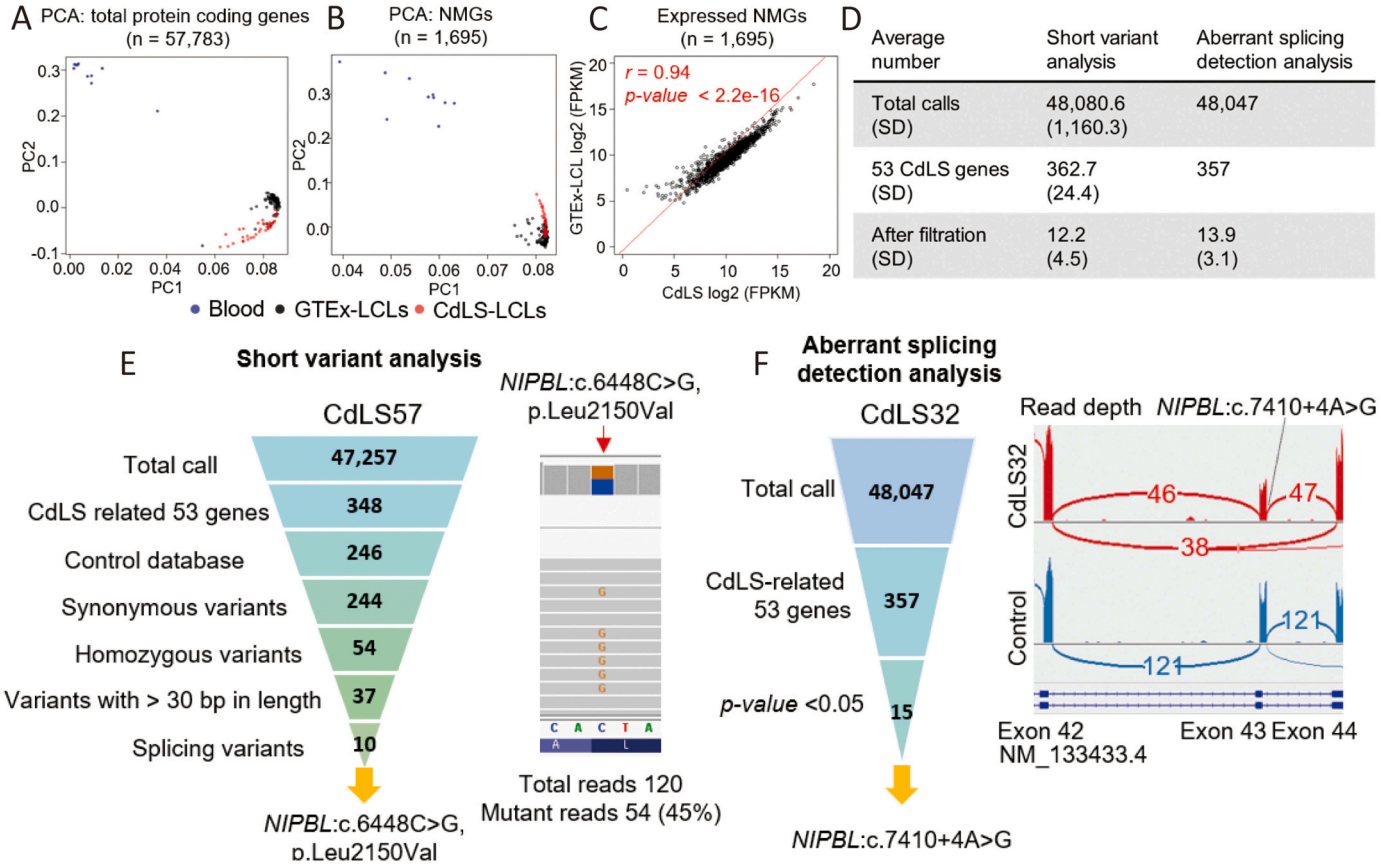

Fig. 2. Quality control and RNA sequencing analysis pipelines.
(A, B) Principal component analysis (PCA) of gene expression profiles among Genotype-Tissue Expression sequencing project (GTEx)-blood, GTEx-EBV-transformed lymphoblastoid B cell lines (LCLs), and Cornelia de Lange Syndrome (CdLS)-LCLs. (A) Total gene expression profile (n = 57,783) and (B) neurodevelopmental Mendelian gene (NMG) expression profile (n = 1695). (C) Correlation between CdLS-LCLs and GTEx-LCLs for NMG expression (n = 1645). (D) Average number of variants for each filtering process in short variant and aberrant splicing detection analyses. Numbers in brackets indicate standard deviations (SDs). (E, F) Number of variants in each filtering step for CdLS57 (short variant analysis) (E, left panel) and for CdLS32 (aberrant splicing detection analysis) (F, left panel). Variants were narrowed down by including and excluding various variants. (E, right panel) Pathogenic variants of CdLS57 in RNA sequencing (RNA-seq) reads captured with the Integrated Genome Viewer (IGV). Red arrow indicates the variant position. The number of reads is given at the bottom (45%, 54/120). (F, right panel) Sashimi plot of aberrant splicing events producing premature termination of *NIPBL* in CdLS32 (red) and the same region in a healthy control (blue). Numbers indicate supporting exon junction reads. The black line indicates the splicing variant position. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2
RNA sequencing analysis detected three variants that cause aberrant splicing and a 3 bp deletion in four CdLS cases.

Family ID	Sex	Age	Gene	Accession number of cDNA	cDNA	Amino acid change	Coordinates (hg19)	Detection method	Clinical classification of CdLS[a]	SpliceAI (A loss/D loss/A gain/D gain)	ACMG/AMP guidelines[b]
CdLS40	Female	9 y	*NIPBL*	NM_133433.4	c.359-1508 T > G	p.Gly120Alafs*5	chr5:36960078	Aberrant splicing detection	Classic	0/0/0.22(−65 bp)/0.25(+24 bp)	Pathogenic (PVS1, PS2, PM2, PP4)
CdLS53	Female	6 m	*NIPBL*	NM_133433.4	c.5863-21 T > G	p.Leu1954Asnfs*8	chr5:37036460	Aberrant splicing detection	Classic	0.15(+21 bp)/0/0.91(+1 bp)/0	Pathogenic (PVS1, PS2, PM2, PP4)
CdLS61	Female	22 y	*NIPBL*	NM_133433.4	c.6653_6655del	p.Asn2218del	chr5:37048661–37,048,663	Short variant analysis	Classic	–	Pathogenic (PS1, PS2, PM2, PP4)
CdLS65	Male	2 y 4 m	*NIPBL*	NM_133433.4	c.4560 + 1965G > T	p.Ile1521Glyfs*13	chr5:37012292	Aberrant splicing detection	Non-classic	0/0/0.05(−14 bp)/0	Pathogenic (PVS1, PS2, PM2, PP4)

CdLS, Cornelia de Lange Syndrome; hg19, human reference genome; y, year; m, month; A, acceptor; D, donor.
[a] Kline AD et al. *Nat Rev Genet.* 2018;19(10):649–666.
[b] Richards S, et al. *Genet Med.* 2015;17(5):405–424.

3. Results

3.1. Correlation of gene expression among GTEx-blood, GTEx-LCLs, and CdLS-LCLs

PCA was carried out among GTEx-blood, GTEx-LCLs, and CdLS-LCLs [12], and similarities in expression profiles of total genes ($n = 57{,}783$) and NMGs ($n = 1695$) were compared. Total gene and NMG expression profiles of GTEx-LCLs and CdLS-LCLs were similarly clustered in the PCA, indicating their similarity (Fig. 2A, B). The expression of NMGs (n = 1695) and the 53 CdLS-related genes showed a strong correlation between CdLS-LCLs and GTEx-LCLs ($r = 0.935$, $p < 2.2 \times 10^{-16}$ and $r = 0.969$, $p < 2.2 \times 10^{-16}$, respectively) (Figs. 2C and S4). The expression levels of NMGs in LCLs have been shown to have a positive correlation with those in brain tissue [8]. Thus, it seems reasonable to analyze our CdLS-LCL data compared with GTEx-LCL data for CdLS-related genes. The diagnosis of CdLS by RNA-seq using LCLs of patients has also been tested in a previous study [8].

3.2. Short variant analysis of RNA-seq data in 22 positive control cases

Approximately 4.8×10^4 short variants found in one or more reads in one individual were detected by short variant analysis (Fig. 2D and Table S6). Most of these variants were extremely low in prevalence (found only in one or a few reads). It was difficult to determine whether these variants were errors or reduced reads due to variant effects, such as nonsense-mediated mRNA decay (NMD) (Table S5). Thus, we focused on 53 CdLS-related genes to exclude error calls and increase the accuracy of short variant candidates. Moreover, frameshift variants > 30 bp in length were excluded from the candidates because these variants were undetected by previous ES analyses and manual ES inspection of IGV images (Fig. S5).

3.2.1. Filtering criteria of short variant analysis with positive controls

We evaluated whether our filtering criteria could effectively identify previously detected pathogenic variants in our positive controls. For example, 47,257 variants were detected in CdLS57, and the number of variants in CdLS-related genes was 348. Because the trio-based ES analyses were conducted previously, we searched for extremely rare de novo variants by RNA-seq analysis. Thus, the number of candidate variants was reduced to 246 (by excluding variants found in the control database), 244 (by excluding synonymous variants), 54 (by excluding homozygous variants), 37 (by excluding variants > 30 bp in length), and 10 (by excluding splicing variants) (Fig. 2E). The pathogenic variant (*NIPBL*, NM_133433.4:c.6448C > G p.Leu2150Val) was easily found in the remaining 10 short variants. These variants were evaluated for their pathogenicity in our ES analysis pipeline [13] (see Supplementary Methods for details). The average number of variants in each inclusion and exclusion process in 22 positive controls was similar to that in CdLS57 (Table S6 and Fig. 2D, E).

Through short variant filtering (Fig. S2), 14 variants (82.4%, 14/17) were detected in 17 positive controls harboring pathogenic variants (not splicing variants) (Table 1). The detection rates of variant types (nonsense, frameshift, inframe, and missense), were 100% (3/3), 66.7% (4/6), 66.7% (2/3), and 100% (5/5), respectively.

3.2.2. Characteristics of positive controls undetected by RNA-seq

Our RNA-seq pipeline did not detect three variants that were previously confirmed by ES analysis: [CdLS10 (*NIPBL*, NM_133433.4: c.5174delA p.Lys1725Serfs*17), CdLS27 (*NIPBL*, NM_133433.4: c.5030_5031del p.Ile1677Serfs*21), and CdLS36 (*SMARCA4*, NM_001128849.3:c.2519_2542delinsGGA p.Ala840_Leu848delinsGlyIle)]. The number of mutant reads was zero or one in two of the three undetected variants in the RNA-seq data (CdLS10 and CdLS27), suggesting possible NMD effects (Fig. S6). ES analysis identified a 24 bp indel (inframe) in *SMARCA4* in CdLS36 (Fig. S7A). However, short

variant analysis of RNA-seq data did not detect this variant even though it was a nontruncating variant. Furthermore, no reads containing this change were detected by manual inspection of IGV images without prior knowledge of this variant (Fig. S7B). Sequencing was performed on a NovaSeq 6000 system in both analyses, but the settings of read lengths were different: 150 bp paired-end reads in ES and 100 bp paired-end reads in RNA-seq. Thus, we assumed that the soft clip reads containing this inframe indel were not effectively detected in the RNA-seq data, although slight coverage reduction in this region was recognized retrospectively (Fig. S7B).

Only one splicing variant (CdLS17) was detected in the five positive controls with the short variant analysis (Table 1 and Fig. S8A). A de novo nonsynonymous variant (c.6343G > T) in *NIPBL* was detected in CdLS47 by ES (Fig. S8B). This variant is located at the 3′ end of exon 36, possibly leading to a missense variant (p.Gly2115Cys, 23/73 reads) and exon 36 skipping (p.Val2085Profs*5, 14/73 reads) (Fig. S8B). Exon 36 skipping likely contributed to the disease pathogenesis in CdLS47. Interestingly, short variant analysis did not detect this variant even though sufficient

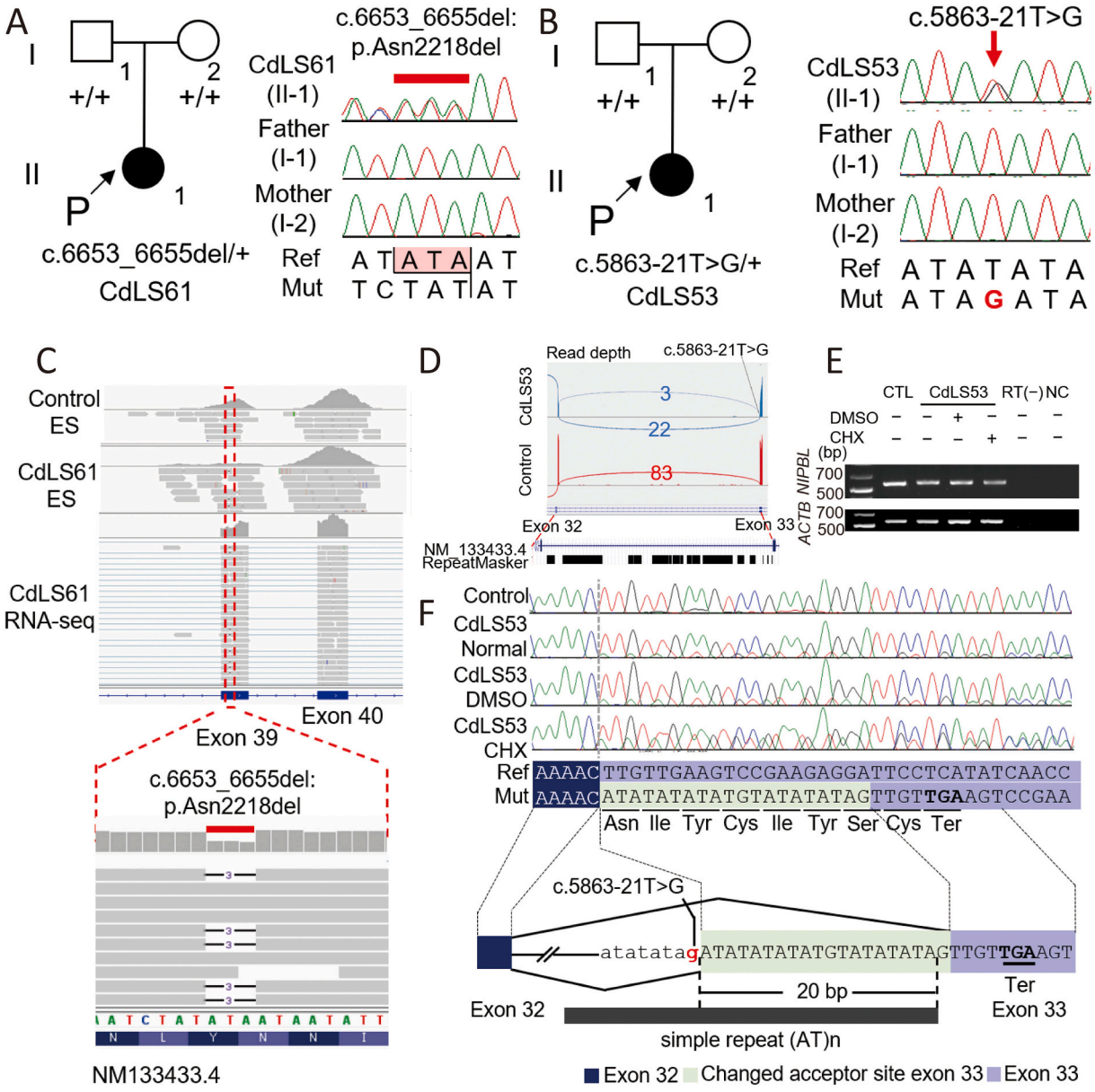

Fig. 3. RNA sequencing identified pathogenic variants in *NIPBL* missed by exome sequencing analysis.
(A, B) Familial pedigrees and Sanger sequencing electropherograms of trio-based polymerase chain reaction (PCR) amplification with genomic DNA in CdLS61 and CdLS53. (C) Integrated Genome Viewer (IGV) image of sequencing reads showing the 3 bp deletion of CdLS61 and the same region of the control. From top to bottom, exome sequencing (ES) of a control, ES of CdLS61, RNA sequencing (RNA-seq) of CdLS61, and an enlarged view of RNA-seq of CdLS61. (A, C) Red horizontal bar indicates the 3 bp deletion in CdLS61. (D) Sashimi plot of exons 32 and 33 of *NIPBL* in CdLS53 (blue) and the same region in a healthy control (red) and in the UCSC genome browser. Numbers indicate supporting exon junction reads. The black line indicates the variant position. (E) Gel electrophoresis and (F) electropherograms of reverse transcription-polymerase chain reaction (RT-PCR) products using lymphoblastoid B cell lines (LCLs) derived from a normal control (CTL) and CdLS53 cultured in media with or without dimethyl sulfoxide (DMSO) or cycloheximide (CHX). *ACTB* was used as an internal control. RT (−), no reverse transcription; Ref, reference allele; Mut, mutant allele; NC, negative control. (F) Light green, light purple, and dark blue squares indicate the newly identified region (21 bp) extending from the 5′ end of exon 33, original exon 33, and exon 32, respectively. The dark blue and dark gray bars indicate the original exon 32 and the newly identified region (21 bp) as a part of aberrant exon 33 and a simple repeat region of (AT)n, respectively. The base marked in red font indicates the pathogenic variant of CdLS53. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

mutant reads (23/73) were mapped. The variant could result in either p.Gly2115Cys (normal splicing) or p.Val2085Profs*5 (exon 36 skipping), but abnormal skipping may have interfered with the appropriate mapping of aberrant splicing reads to the reference genome. In contrast, aberrant splicing detection analysis detected all five pathogenic splicing variants, including CdLS47, in positive controls (see section 3.3). Thus, we did not focus on splicing events in the short variant analysis.

3.3. Investigating the precise settings of LeafCutterMD with positive CdLS controls

To clarify the precise coverage threshold, we investigated the number of mutant reads by manual inspection of IGV images in each positive control (Fig. S8). Among the five positive controls, CdLS60 had the smallest number of mutant reads, which was two (Fig. S8E). Therefore, the minimum coverage threshold was set to 1. To accommodate genes with long intron lengths, the distance to detect splicing changes was set to 500,000 bp, the upper limit. When clusters were analyzed using these settings, 48,047 clusters were found. Focusing on the 53 CdLS-related

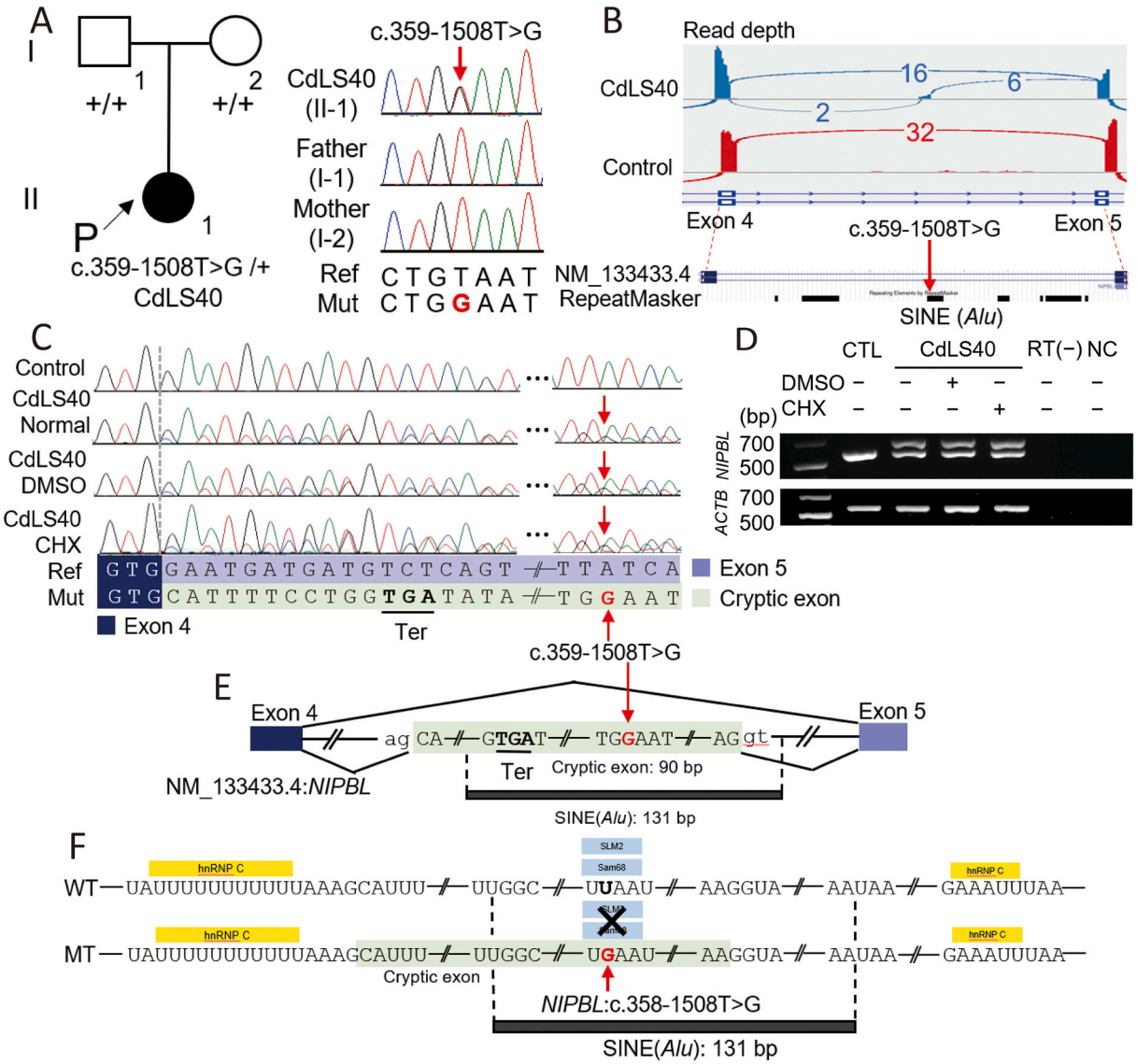

Fig. 4. Cryptic exon of *NIPBL* newly identified by RNA sequencing in CdLS40.
(A) Familial pedigree (left) and electropherograms (right) of trio-based Sanger sequencing. The red arrow and font indicate the pathogenic variant. (B) Sashimi plot of exons 4 and 5 of *NIPBL* in CdLS40 (blue) and a control (red) and the same region in the UCSC genome browser. Numbers indicate supporting exon junction reads. Gray bars indicate repetitive sequences detected by RepeatMasker. The red arrow indicates the variant position. (C) Electropherograms of Sanger sequencing and (D) gel electrophoresis of reverse transcription-polymerase chain reaction (RT-PCR) products of lymphoblastoid B cell lines (LCLs) derived from a normal control (CTL) and CdLS40 cultured in media with or without dimethyl sulfoxide (DMSO) or cycloheximide (CHX). *ACTB* was used as an internal control. RT (−), no reverse transcription. Ref, reference allele; Mut, mutant allele; NC, negative control. (C, E) Light green, dark blue, and light purple squares indicate a 90 bp cryptic exon in intron 4, and original exons 4 and 5, respectively. The dark gray bar indicates an *Alu* element. The red arrows and base marked in red font indicates the de novo pathogenic variant of CdLS40. (E) Schematic representation of exon 4, cryptic exon in intron 4, and exon 5. (F) Association between pre-mRNA and RNA-binding proteins located at intron 4, including an identified cryptic exon in CdLS40. The conditions between wild-type (WT) and mutant (MT) pre-mRNA associated with the RNA-binding proteins heterogeneous nuclear ribonucleoprotein (hnRNAP) C (yellow boxes) and Sam68-like mammalian protein 2 (SLM-2) and SRC associated in mitosis of 68 kDa (Sam68) (light pale blue boxes) are shown. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

genes, 357 clusters were found in 48,047 clusters. Furthermore, the average number of clusters with *p-values* < 0.05 was 13.9 [standard deviation (SD) 3.1] (Fig. 2D). For CdLS32, 15 out of 357 clusters met the conditions with *p-values* < 0.05 (Fig. 2F). Subsequent manual inspection of IGV images of each candidate cluster detected a disease-causing splicing event (*NIPBL*, NM_133433.4:c.7410 + 4A > G p.Lys2422_Glu2470del) (Fig. 2F). The remaining four positive cases were also confirmed by this threshold, and disease-causing splicing events were detected in all of them (Table S7 and Fig. S8). These settings were used for the 19 unsolved cases (Fig. S9).

3.4. Pathogenic variants detected from 19 unsolved probands in this study

Four pathogenic variants in *NIPBL* (one inframe deletion and three intronic variants) were detected by our pipeline in 19 unsolved probands (Table 2). Clinical information of four of these cases is presented in

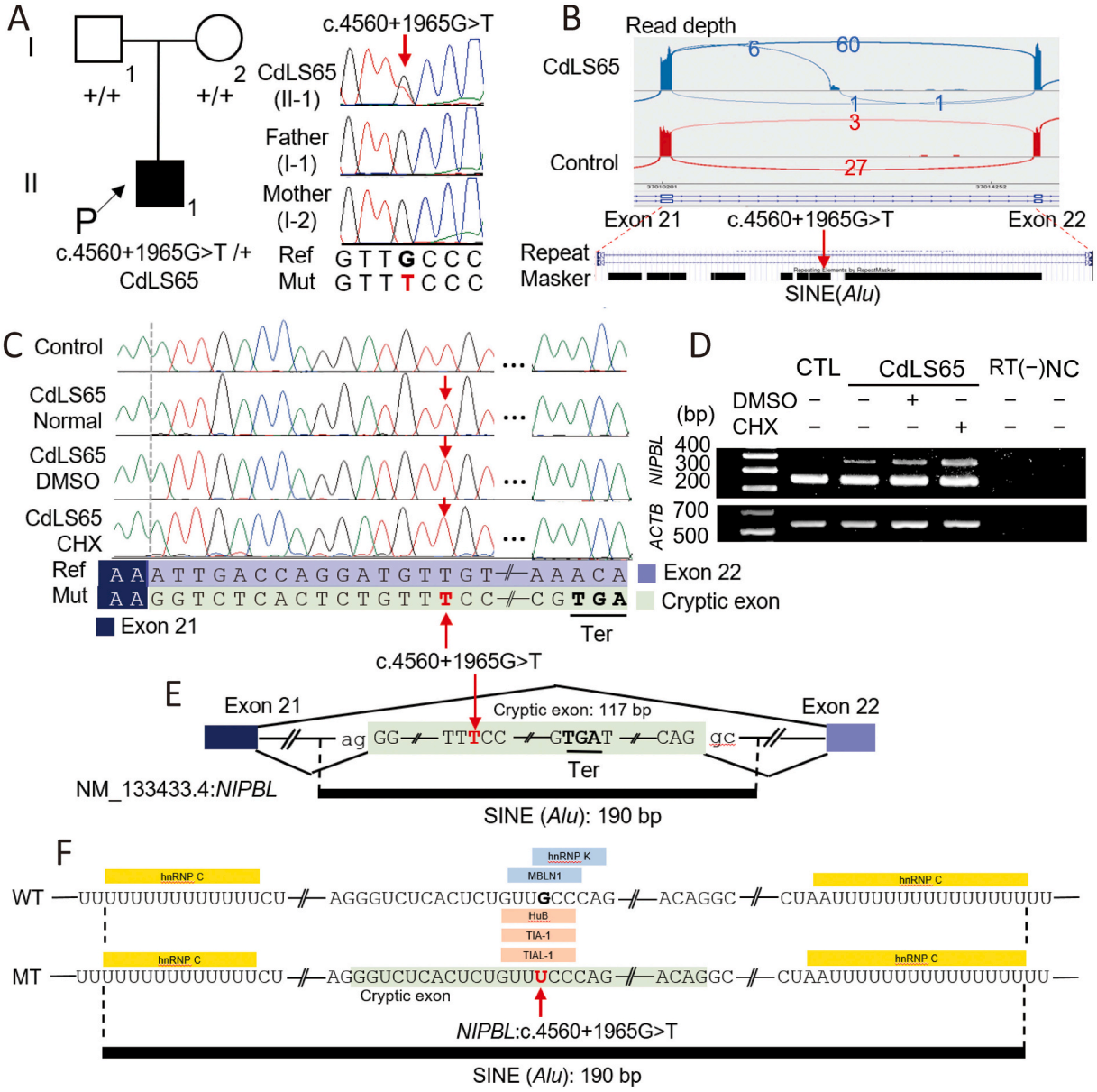

Fig. 5. Cryptic exon in *NIPBL* newly identified by RNA sequencing in CdLS65.
(A) Familial pedigree (left) and electropherograms (right) of trio-based Sanger sequencing. The red arrow and font indicate the pathogenic variant. (B) Sashimi plot at exons 21 and 22 of *NIPBL* in CdLS65 (blue) and a control (red) and the same region in the UCSC genome browser. Numbers indicate supporting exon junction reads. Gray bars indicate repetitive sequences detected by RepeatMasker. Red arrow indicates the pathogenic variant position. (C) Electropherograms of Sanger sequencing and (D) gel electrophoresis of reverse transcription-polymerase chain reaction (RT-PCR) products of lymphoblastoid B cell lines (LCLs) derived from a disease-free control (CTL) and CdLS65 cultured in media with or without dimethyl sulfoxide (DMSO) or cycloheximide (CHX). *ACTB* was used as an internal control. RT (−), no reverse transcription. Ref, reference allele; Mut, mutant allele; NC, negative control. (C, E) Light green, dark blue, and light purple squares indicate a 117 bp cryptic exon in intron 21, and original exons 21 and 22, respectively. The dark gray bar indicates an *Alu* element. The red arrows and base marked in red font indicate the de novo pathogenic variants in *NIPBL* for CdLS65. (E) Schematic representation of exon 21, a cryptic exon in intron 21, and exon 22. (F) Association between pre-mRNA and RNA-binding proteins located at intron 21. The conditions between wild-type (WT) and mutant (MT) pre-mRNA associated with the RNA-binding proteins heterogeneous nuclear ribonucleoprotein (hnRNP) K (yellow boxes), muscleblind-like Protein 1 (MBLN1) (light blue boxes), and Hu antigen B, TIAT-cell intracellular antigen 1 (TIA-1), and TIA1-related/like protein (TIAL-1) (light orange boxes) are shown. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Fig. S10 and Table S8.

3.4.1. Pathogenic variant detected by short variant analysis

One known inframe deletion [25] was detected by short variant analysis in one unsolved case (CdLS61, NM_133433.4:c.6653_6655del p.Asn2218del) (Fig. 3A, C).

3.4.2. Pathogenic variants detected by aberrant splicing detection analysis

The aberrant splicing detection analysis identified three disease-causing splicing variants in *NIPBL*. A de novo variant (NM_133433.4: c.5863-21 T > G p.Leu1954Asnfs*8) was identified 21 bp upstream of the 5′ end of canonical exon 33 in CdLS53, and subsequent RT-PCR confirmed a cryptic acceptor site creating new exon 33 (Fig. 3B, D–F). This variant (c.5863-21 T > G) was located at simple repeat (TA)n (by RepeatMasker), and an exon–intron junction was newly created because of an ATAT to AGAT change. Furthermore, the new exon 33 contained a 20 bp upstream intronic region (Fig. 3D, F). The remaining two variants were identified in the deep-intronic region of intron 4 (CdLS40, NM_133433.4:c.359-1508 T > G p.Gly120Alafs*5) and intron 21 (CdLS65, NM_133433.4:c.4560 + 1965G > T p.Ile1521Glyfs*13). RT-PCR revealed cryptic exon inclusions (a 90 bp region in intron 4 and a 117 bp region in intron 21) in each intron (Figs. 4 and 5). Those electropherograms showed moderate to slight recovery of the mutant reads under the inhibition of NMD by cycloheximide treatment. These results demonstrate that all three de novo variants induced a premature termination codon, leading to NMD (Figs. 3E–F, 4C–E, 5C–E, and Fig. S11).

3.4.3. Changes in binding proteins around deep-intronic variants creating cryptic exons

De novo variants of CdLS40 and CdLS65 were not directly correlated to acceptor or donor site formation (Figs. 4E and 5E). Because these two variants were located in *Alu* elements (Figs. 4E–F and 5E–F), we first evaluated the loss of heterogeneous nuclear ribonucleoprotein (hnRNP) C-mediated prevention of *Alu* exonization using SpliceAid2 [26]. However, there were no differences in hnRNP C binding efficiency in these cryptic exon regions between wild-type and mutant alleles (Figs. 4F and 5F). The recognition sites of the following RNA-binding proteins were affected by the de novo variants in CdLS40 and CdLS65: the losses of SRC associated in mitosis of 68 kDa (Sam68) and Sam68-like mammalian protein 2 (SLM-2) in CdLS40, losses of muscleblind-like splicing regulator 1 (MBNL1) and hnRNP K in CdLS65, and gains of Hu antigen B, T-cell intracellular antigen 1 (TIA-1), and TIA1-related/like protein (TIAL-1) in CdLS65 (Figs. 4F and 5F).

Sam68 and SLM-2 are classified as STAR family proteins, which link signal transduction to post-transcriptional gene regulation [27]. In particular, Sam68 binds close to splice sites and regulates splicing by synergizing or competing with other splicing factors, such as U2AF2, hnRNP, and U170k [27,28]. TIAL-1 and TIA-1 were reported to act as regulators of transcription and pre-mRNA splicing and be involved in cell proliferation, apoptosis, embryogenesis, inflammation, and tumor suppression [29]. Although the types of predicted binding proteins altered by variants were different between CdLS40 and CdLS65, the variants were speculated to disrupt a protective mechanism preventing abnormal splicing of the *Alu* elements by affecting the binding of these proteins to the aberrant allele in these two cases. Thus, these data may suggest that the two de novo variants are strongly involved in the aberrant splicing.

3.5. Use of RNA-seq in short variant analysis

Regarding short variant analysis, trio-based ES analysis of CdLS61 did not identify a de novo 3 bp deletion in *NIPBL* (NM_133433.4: c.6653_6655del p.Asn2218del). Evaluation of the read depth at the inframe deletion (chr5:37048661–37,048,663) and entire exon 39 revealed an extremely decreased number of mapped reads in exon 39 compared with the numbers in the neighboring exons (38 and 40) regardless of the 37.5% GC ratio (not high) in exon 39 (Fig. 3C). Only six reads (the number of wild and aberrant alleles: 2 and 4, respectively) were mapped at the variant position. However, coverage of the *NIPBL* gene in the RNA-seq data was sufficient (32.41 ×) but less than that in ES (41.3 ×), suggesting the capture inefficiency of this particular region in the ES analysis (Fig. 3C).

We also compared the read depth of exon 39 and capture kits of ES among CdLS61, her parents, and disease-free controls (Fig. S12 and Table S9). Interestingly, the read depth of exon 39, entire *NIPBL* gene coverage, and mean coverage of ES for CdLS61, her parents, and a normal control sequenced with the same capture kit [SureSelect Human All Exon V6 (58 Mb) system; Agilent Technologies, Santa Clara, CA, USA], were similar (Fig. S12 and Table S9). However, the read depth of other normal controls sequenced with another capture kit (Twist Comprehensive Exome Panel; Twist Bioscience, South San Francisco, CA, USA) only showed a slight decrease in the read coverage of exon 39 compared with that of exon 40 (Fig. S12 and Table S9). Considering these results, the read depth reductions in exon 39 may be due to limitations of the capture kit.

Next, *NIPBL* expression (FPKM) was compared among CdLS groups based on the different types of variants and the control group (Fig. S13). *NIPBL* expression in CdLS patients with loss-of-function variants (nonsense, frameshift, and splicing related to NMD) [1779.1 (mean) ± 285.9 (SD) FPKM] was significantly reduced compared with *NIPBL* expression in the control group [2407.2 (mean) ± 546.0 (SD) FPKM] (*p-value* = $3.8e^{-0.6}$). Additionally, *NIPBL* expression was slightly different between the CdLS group with missense or inframe variants [2201.1 (mean) ± 223.3 (SD) FPKM] and the control group (*p-value* = 0.02). Because significant differences of gene expression were detected for different variant types, these results indicate that RNA-seq could detect short variants in the coding regions in cases in which expression was reasonably preserved.

4. Discussion

In this study, we first created an RNA-seq pipeline for determining disease-causing variants, including the detection of short coding variants and aberrant splicing events, in rare Mendelian disorders using the RNA-seq data of LCLs from our 22 CdLS cases with pathogenic variants solved by ES (as positive controls) and 106 healthy controls in the GTEx Biobank. RNA-seq data were also helpful for confirming the status of transcripts of candidate genes by screening cDNA reads using the IGV.

Short variant analysis was unable to detect splicing variants, including one seemingly missense but actual splicing variant in CdLS47 (1/5, Table 1). Thus, different methods were selected for 1) short variant analysis to detect SNVs and short indels and 2) aberrant splicing events. One of the most significant challenges in the short variant analysis was several error calls in the pipeline. To address this issue, a gene list compatible with the diseases of patients should be provided. With such a list, we could easily narrow down candidate variants. Moreover, NMD might prevent variant detection, potentially diminishing aberrant reads with truncating variants. In cases with pathogenic variants detected by either ES or RNA-seq, the technical limitation may weigh more on the biological experiment (such as DNA read length). For example, in CdLS36, the reason for failing to detect the variant was an insufficient DNA read length of the sequencer, and in CdLS61, it was the insufficient read coverage in exon 39 of *NIPBL* by the Agilent capture kit (Table S9). Detection was improved by using the Twist Comprehensive Exome Panel kit, which provided generally uniform read depths per gene and/or region (Table S9). Thus, some ES-negative cases were solved by subsequent genome-based analysis.

We next investigated 19 unsolved cases using our RNA-seq pipelines and found four previously unidentified disease-causing variants in four unsolved cases. Regarding CdLS61, short variant analysis detected a pathogenic variant even where low read coverage was found in ES.

Aberrant splicing detection analysis found de novo pathogenic variants associated with aberrant splicing events and or creating a cryptic exon in deep intronic regions. Because all four variants were located in the *NIPBL* gene, we re-investigated the remaining 52 CdLS-related genes to identify newly created cryptic exons in *NIPBL*, but no other variants were found. Some variants in *Alu* elements are known to be associated with cryptic exon formation in several genes [30,31]. Mendelian disease-related genes with many *Alu* elements within their gene bodies may be good targets for aberrant splicing detection in the future. Therefore, RNA-seq by trio-based ES provides a plausible approach in unsolved CdLS cases and other diseases.

In cancers with somatic variants, differentially expressed genes involved in development and metastasis have been identified by RNA-seq analyses [32–34]. However, Rentas et al. reported that they could not identify pathogenic variants in germline diseases if the expression of target genes was significantly different between the disease group and the control group [8]. As expected, we did not find any pathogenic variants in differentially expressed genes (such as *RAD21, SMC1A,* and *SMARCA4*) in our CdLS cohort (Fig. S14).

Recently, Pozojevic et al. reported that mosaic variants were identified using buccal swabs in CdLS cases in which pathogenic variants could not be detected by Sanger sequencing of DNA derived from peripheral blood leukocytes [35]. Thus, to identify pathogenic variants in the remaining 16 unsolved CdLS cases, ES analysis using buccal mucosal DNA or short- or long-read whole genome sequencing of the unanalyzed regions by ES should be considered.

5. Conclusion

In summary, by developing the RNA-seq pipeline, four pathogenic variants were newly identified by RNA-seq of 19 CdLS cases that were unsolved using ES analysis. As a result, the total diagnostic rate was increased from 69.7% (46/66) to 75.8% (50/66). Thus, we concluded that RNA-seq of LCLs was useful to determine hidden variants in ES-negative CdLS cases and is applicable to other Mendelian disorders.

Code availability

All computational tools/codes used in this study can be downloaded through the following websites.

STAR: https://github.com/alexdobin/STAR
GATK: https://github.com/broadinstitute/gatk
ANNOVAR: https://github.com/WGLab/doc-ANNOVAR
LeafCutterMD: http://davidaknowles.github.io/leafcutter
RSEM: https://github.com/deweylab/RSEM
DESeq2: https://github.com/mikelove/DESeq2
TCC-GUI: https://github.com/swsoyee/TCC-GUI

Funding

This study was supported by the Japan Agency for Medical Research and Development (AMED) [grant numbers JP22ek0109486, JP22ek0109549, JP22ek0109493]; Japan Society for the Promotion of Science (JSPS) KAKENHI [grant numbers JP20K07907, JP20K08164, JP21K15097, JP20K17428, JP21K07869, JP20K16932]; the Takeda Science Foundation; and the Ichiro Kanehara Foundation for the Promotion of Medical Science and Medical Care.

Author contributions

R.S., Y.U., and N.M. contributed to the conception and design of the study. R.S., Y.U., K.I., K.H., and N.M. were involved in the data analysis. J.C., V.K., I.F., R.H., D.B., and C.K. collected patients' samples and their clinical information. R.S., Y.U., and N.M. contributed to drafting the text. All authors critically read the manuscript, corrected it, and approved the final version of the article.

Ethics declarations

All genomic DNA from patients and their families were examined after obtaining informed consent. Experimental protocols were approved by the institutional review board of Yokohama City University under the number A170525011 (modified B211100023). Our research conformed to the principles of the Helsinki Declaration.

Declaration of Competing Interest

The authors declare that they have no conflicts of interest.

Data availability

The data that support the findings of this study are available on request to the corresponding author.

Acknowledgements

We thank all patients and their families for their participation in this study. We thank N. Watanabe, T. Miyama, M. Sato, S. Sugimoto, and K. Takabe for their technical assistance. We thank Melissa Crawford, PhD, and Margaret Biswas, PhD, from Edanz (https://jp.edanz.com/ac) for editing a draft of this manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ygeno.2022.110468.

References

[1] M.J. Bamshad, S.B. Ng, A.W. Bigham, H.K. Tabor, M.J. Emond, D.A. Nickerson, J. Shendure, Exome sequencing as a tool for Mendelian disease gene discovery, Nat. Rev. Genet. 12 (2011) 745–755, https://doi.org/10.1038/nrg3031.

[2] C.F. Wright, T.W. Fitzgerald, W.D. Jones, S. Clayton, J.F. McRae, M. Van Kogelenberg, D.A. King, K. Ambridge, D.M. Barrett, T. Bayzetinova, A.P. Bevan, E. Bragin, E.A. Chatzimichali, S. Gribble, P. Jones, N. Krishnappa, L.E. Mason, R. Miller, K.I. Morley, V. Parthiban, E. Prigmore, D. Rajan, A. Sifrim, G. J. Swaminathan, A.R. Tivey, A. Middleton, M. Parker, N.P. Carter, J.C. Barrett, M. E. Hurles, D.R. Fitzpatrick, H.V. Firth, Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data, Lancet. 385 (2015) 1305–1314, https://doi.org/10.1016/S0140-6736(14)61705-0.

[3] K. Retterer, J. Juusola, M.T. Cho, P. Vitazka, F. Millan, F. Gibellini, A. Vertino-Bell, N. Smaoui, J. Neidich, K.G. Monaghan, D. McKnight, R. Bai, S. Suchy, B. Friedman, J. Tahiliani, D. Pineda-Alvarez, G. Richard, T. Brandt, E. Haverfield, W.K. Chung, S. Bale, Clinical application of whole-exome sequencing across clinical indications, Genet. Med. 18 (2016) 696–704, https://doi.org/10.1038/gim.2015.148.

[4] B. Hwang, J.H. Lee, D. Bang, Single-cell RNA sequencing technologies and bioinformatics pipelines, Exp. Mol. Med. 50 (2018) 1–14, https://doi.org/10.1038/s12276-018-0071-8.

[5] Z. Wang, M. Gerstein, M. Snyder, Nrg2484-1, Nat. Rev. Genet. 10 (2009) 57–63.

[6] M. Sultan, M.H. Schulz, H. Richard, A. Magen, A. Klingenhoff, M. Scherf, M. Seifert, T. Borodina, A. Soldatov, D. Parkhomchuk, D. Schmidt, S.O. Keeffe, S. Haas, M. Vingron, H. Lehrach, M. Yaspo, A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome, Science (80-.) 685 (2008) 956–961.

[7] L. Frésard, C. Smail, N.M. Ferraro, N.A. Teran, X. Li, K.S. Smith, D. Bonner, K. D. Kernohan, S. Marwaha, Z. Zappala, B. Balliu, J.R. Davis, B. Liu, C.J. Prybol, J. N. Kohler, D.B. Zastrow, C.M. Reuter, D.G. Fisk, M.E. Grove, J.M. Davidson, T. Hartley, R. Joshi, B.J. Strober, S. Utiramerur, D.R. Adams, A. Aday, M. E. Alejandro, P. Allard, E.A. Ashley, M.S. Azamian, C.A. Bacino, E. Baker, A. Balasubramanyam, H. Barseghyan, G.F. Batzli, A.H. Beggs, B. Behnam, H. J. Bellen, J.A. Bernstein, G.T. Berry, A. Bican, D.P. Bick, C.L. Birch, D. Bonner, B. E. Boone, B.L. Bostwick, L.C. Briere, E. Brokamp, D.M. Brown, M. Brush, E. A. Burke, L.C. Burrage, M.J. Butte, S. Chen, G.D. Clark, T.R. Coakley, J.D. Cogan, H. A. Colley, C.M. Cooper, H. Cope, W.J. Craigen, P. D'Souza, M. Davids, J. M. Davidson, J.G. Dayal, E.C. Dell'Angelica, S.U. Dhar, K.M. Dipple, L.A. Donnell-Fink, N. Dorrani, D.C. Dorset, E.D. Douine, D.D. Draper, A.M. Dries, L. Duncan, D. J. Eckstein, L.T. Emrick, C.M. Eng, G.M. Enns, A. Eskin, C. Esteves, T. Estwick, L. Fernandez, C. Ferreira, E.L. Fieg, P.G. Fisher, B.L. Fogel, N.D. Friedman, W. A. Gahl, E. Glanton, R.A. Godfrey, A.M. Goldman, D.B. Goldstein, S.E. Gould, J.P. F. Gourdine, C.A. Groden, A.L. Gropman, M. Haendel, R. Hamid, N.A. Hanchard, F. High, I.A. Holm, J. Hom, E.M. Howerton, Y. Huang, F. Jamal, Y. Hui Jiang, J. M. Johnston, A.L. Jones, L. Karaviti, D.M. Koeller, I.S. Kohane, J.N. Kohler, D.

M. Krasnewich, S. Korrick, M. Koziura, J.B. Krier, J.E. Kyle, S.R. Lalani, C.C. Lau, J. Lazar, K. LeBlanc, B.H. Lee, H. Lee, S.E. Levy, R.A. Lewis, S.A. Lincoln, S.K. Loo, J. Loscalzo, R.L. Maas, E.F. Macnamara, C.A. MacRae, V.V. Maduro, M. M. Majcherska, M.C.V. Malicdan, L.A. Mamounas, T.A. Manolio, T.C. Markello, R. Marom, M.G. Martin, J.A. Martínez-Agosto, S. Marwaha, T. May, A. McConkie-Rosell, C.E. McCormack, A.T. McCray, J.D. Merker, T.O. Metz, M. Might, P. M. Moretti, M. Morimoto, J.J. Mulvihill, D.R. Murdock, J.L. Murphy, D.M. Muzny, M.E. Nehrebecky, S.F. Nelson, J.S. Newberry, J.H. Newman, S.K. Nicholas, D. Novacic, J.S. Orange, J.P. Orengo, J.C. Pallais, C.G. Palmer, J.C. Papp, N. H. Parker, L.D. Pena, J.A. Phillips, J.E. Posey, J.H. Postlethwait, L. Potocki, B. N. Pusey, G. Renteria, C.M. Reuter, L. Rives, A.K. Robertson, L.H. Rodan, J. A. Rosenfeld, J.B. Sampson, S.L. Samson, K. Schoch, D.A. Scott, L. Shakachite, P. Sharma, V. Shashi, R. Signer, E.K. Silverman, J.S. Sinsheimer, K.S. Smith, R. C. Spillmann, J.M. Stoler, N. Stong, J.A. Sullivan, D.A. Sweetser, Q.K.G. Tan, C. J. Tifft, C. Toro, A.A. Tran, T.K. Urv, E. Vilain, T.P. Vogel, D.M. Waggott, C.E. Wahl, N.M. Walley, C.A. Walsh, M. Walker, J. Wan, M.F. Wangler, P.A. Ward, K. M. Waters, B.J.M. Webb-Robertson, M. Westerfield, M.T. Wheeler, A.L. Wise, L. A. Wolfe, E.A. Worthey, S. Yamamoto, J. Yang, Y. Yang, A.J. Yoon, G. Yu, D. B. Zastrow, C. Zhao, A. Zheng, K. Boycott, A. MacKenzie, J. Majewski, M. Brudno, D. Bulman, D. Dyment, L. Lind, E. Ingelsson, A. Battle, G. Bejerano, J.A. Bernstein, E.A. Ashley, K.M. Boycott, J.D. Merker, M.T. Wheeler, S.B. Montgomery, Identification of rare-disease genes using blood transcriptome sequencing and large control cohorts, Nat. Med. 25 (2019) 911–919, https://doi.org/10.1038/s41591-019-0457-8.

[8] S. Rentas, K.S. Rathi, M. Kaur, P. Raman, I.D. Krantz, M. Sarmady, A.A. Tayoun, Diagnosing Cornelia de Lange syndrome and related neurodevelopmental disorders using RNA sequencing, Genet. Med. 22 (2020) 927–936, https://doi.org/10.1038/s41436-019-0741-5.

[9] H. Aoi, T. Mizuguchi, J.R. Ceroni, V.E.H. Kim, I. Furquim, R.S. Honjo, T. Iwaki, T. Suzuki, F. Sekiguchi, Y. Uchiyama, Y. Azuma, K. Hamanaka, E. Koshimizu, S. Miyatake, S. Mitsuhashi, A. Takata, N. Miyake, S. Takeda, A. Itakura, D. R. Bertola, C.A. Kim, N. Matsumoto, Comprehensive genetic analysis of 57 families with clinically suspected Cornelia de Lange syndrome, J. Hum. Genet. 64 (2019) 967–978, https://doi.org/10.1038/s10038-019-0643-z.

[10] Y. Takahashi, H. Date, H. Oi, T. Adachi, N. Imanishi, E. Kimura, H. Takizawa, S. Kosugi, N. Matsumoto, K. Kosaki, Y. Matsubara, Y. Ando, T. Anzai, T. Ariga, Y. Fukushima, Y. Furusawa, A. Ganaha, Y. Goto, K. Hata, M. Honda, K. Iijima, T. Ikka, I. Imoto, T. Kaname, M. Kobayashi, S. Kojima, H. Kurahashi, S. Kure, K. Kurosawa, Y. Maegaki, Y. Makita, T. Morio, I. Narita, F. Nomura, T. Ogata, K. Ozono, A. Oka, N. Okamoto, S. Saitoh, A. Sakurai, F. Takada, T. Takahashi, A. Tamaoka, A. Umezawa, A. Yachie, K. Yoshiura, Y. Chinen, M. Eguchi, K. Fujio, K. Hosoda, T. Ichikawa, T. Kawarai, T. Kosho, M. Masuno, A. Nakamura, T. Nakane, T. Ogi, S. Okada, Y. Sakata, T. Seto, Y. Takahashi, T. Takano, M. Ueda, H. Yagasaki, T. Yamamoto, A. Watanabe, Y. Hotta, A. Kubo, H. Maruyama, K. Moriyama, E. Nanba, N. Sakai, Y. Sekijima, T. Shimosegawa, T. Takeuchi, S. Usami, K. Yamamoto, H. Mizusawa, Six years' accomplishment of the initiative on rare and undiagnosed diseases: nationwide project in Japan to discover causes, mechanisms, and cures, J. Hum. Genet. (2022), https://doi.org/10.1038/S10038-022-01025-0.

[11] A.D. Kline, J.F. Moss, A. Selicorni, A.M. Bisgaard, M.A. Deardorff, P.M. Gillett, S. L. Ishman, L.M. Kerr, A.V. Levin, P.A. Mulder, F.J. Ramos, J. Wierzba, P.F. Ajmone, D. Axtell, N. Blagowidow, A. Cereda, A. Costantino, V. Cormier-Daire, D. FitzPatrick, M. Grados, L. Groves, W. Guthrie, S. Huisman, F.J. Kaiser, G. Koekkoek, M. Levis, M. Mariani, J.P. McCleery, L.A. Menke, A. Metrena, J. O'Connor, C. Oliver, J. Pie, S. Piening, C.J. Potter, A.L. Quaglio, E. Redeker, D. Richman, C. Rigamonti, A. Shi, Z. Tümer, I.D.C. Van Balkom, R.C. Hennekam, Diagnosis and management of Cornelia de Lange syndrome: first international consensus statement, Nat. Rev. Genet. 19 (2018) 649–666, https://doi.org/10.1038/s41576-018-0031-0.

[12] J. Lonsdale, J. Thomas, M. Salvatore, R. Phillips, E. Lo, S. Shad, R. Hasz, G. Walters, F. Garcia, N. Young, B. Foster, M. Moser, E. Karasik, B. Gillard, K. Ramsey, S. Sullivan, J. Bridge, H. Magazine, J. Syron, J. Fleming, L. Siminoff, H. Traino, M. Mosavel, L. Barker, S. Jewell, D. Rohrer, D. Maxim, D. Filkins, P. Harbach, E. Cortadillo, B. Berghuis, L. Turner, E. Hudson, K. Feenstra, L. Sobin, J. Robb, P. Branton, G. Korzeniewski, C. Shive, D. Tabor, L. Qi, K. Groch, S. Nampally, S. Buia, A. Zimmerman, A. Smith, R. Burges, K. Robinson, K. Valentino, D. Bradbury, M. Cosentino, N. Diaz-Mayoral, M. Kennedy, T. Engel, P. Williams, K. Erickson, K. Ardlie, W. Winckler, G. Getz, D. DeLuca, M. Daniel MacArthur, A. Kellis, T. Thomson, E. Young, M. Gelfand, Y. Donovan, G. Meng, D. Grant, Y. Mash, M. Marcus, J. Basile, J. Liu, Z. Zhu, N.J. Tu, D.L. Cox, E.R. Nicolae, H. K. Gamazon, A. Im, J. Konkashbaev, M. Pritchard, T. Stevens, X. Flutre, E.T. Wen, T. Dermitzakis, R. Lappalainen, J. Guigo, M. Monlong, D. Sammeth, A. Koller, S. Battle, M. Mostafavi, M. McCarthy, J. Rivas, I. Maller, A. Rusyn, F. Nobel, A. Wright, M. Shabalin, N. Feolo, A. Sharopova, J. Sturcke, J.M. Paschal, E. L. Anderson, L.K. Wilder, E.D. Derr, J.P. Green, G. Struewing, S. Temple, J.T. Volpi, E.J. Boyer, M.S. Thomson, C. Guyer, A. Ng, D. Abdallah, T.R. Colantuoni, S.E. Koester Insel, A. Roger Little, P.K. Bender, T. Lehner, Y. Yao, C.C. Compton, J. B. Vaught, S. Sawyer, N.C. Lockhart, J. Demchok, H.F. Moore, The genotype-tissue expression (GTEx) project, Nat. Genet. 45 (2013) 580–585, https://doi.org/10.1038/ng.2653.

[13] R. Seyama, N. Tsuchida, Y. Okada, S. Sakata, K. Hamada, Y. Azuma, K. Hamanaka, A. Fujita, E. Koshimizu, S. Miyatake, T. Mizuguchi, S. Makino, A. Itakura, S. Okada, N. Okamoto, K. Ogata, Y. Uchiyama, N. Matsumoto, Two families with TET3-related disorder showing neurodevelopmental delay with craniofacial dysmorphisms, J. Hum. Genet. (2021) 1–8, https://doi.org/10.1038/s10038-021-00986-y.

[14] Y. Uchiyama, D. Yamaguchi, K. Iwama, S. Miyatake, K. Hamanaka, N. Tsuchida, H. Aoi, Y. Azuma, T. Itai, K. Saida, H. Fukuda, F. Sekiguchi, T. Sakaguchi, M. Lei, S. Ohori, M. Sakamoto, M. Kato, T. Koike, Y. Takahashi, K. Tanda, Y. Hyodo, R. S. Honjo, D.R. Bertola, C.A. Kim, M. Goto, T. Okazaki, H. Yamada, Y. Maegaki, H. Osaka, L.H. Ngu, C.G. Siew, K.W. Teik, M. Akasaka, H. Doi, F. Tanaka, T. Goto, L. Guo, S. Ikegawa, K. Haginoya, M. Haniffa, N. Hiraishi, Y. Hiraki, S. Ikemoto, A. Daida, S. Ichiro Hamano, M. Miura, A. Ishiyama, O. Kawano, A. Kondo, H. Matsumoto, N. Okamoto, T. Okanishi, Y. Oyoshi, E. Takeshita, T. Suzuki, Y. Ogawa, H. Handa, Y. Miyazono, E. Koshimizu, A. Fujita, A. Takata, N. Miyake, T. Mizuguchi, N. Matsumoto, Efficient detection of copy-number variations using exome data: batch- and sex-based analyses, Hum. Mutat. 42 (2021) 50–65, https://doi.org/10.1002/humu.24129.

[15] P. Danecek, J.K. Bonfield, J. Liddle, J. Marshall, V. Ohan, M.O. Pollard, A. Whitwham, T. Keane, S.A. Mccarthy, R.M. Davies, H. Li, Twelve Years of SAMtools and BCFtools, Gigascience 10 (2022) 1–4, https://doi.org/10.1093/gigascience/giab008.

[16] B. Li, C.N. Dewey, RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome, BMC Bioinformatics. 12 (2011) 1–16, https://doi.org/10.1186/1471-2105-12-323/TABLES/6.

[17] G.D. Mehta, R. Kumar, S. Srivastava, S.K. Ghosh, Cohesin: functions beyond sister chromatid cohesion, FEBS Lett. 587 (2013) 2299–2312, https://doi.org/10.1016/j.febslet.2013.06.035.

[18] R.C. Centore, G.J. Sandoval, L.M.M. Soares, C. Kadoch, H.M. Chan, Mammalian SWI/SNF chromatin remodeling complexes: emerging mechanisms and therapeutic strategies, Trends Genet. 36 (2020) 936–950, https://doi.org/10.1016/j.tig.2020.07.011.

[19] G.A. Van der Auwera, M.O. Carneiro, C. Hartl, R. Poplin, G. del Angel, A. Levy-Moonshine, T. Jordan, K. Shakir, D. Roazen, J. Thibault, E. Banks, K.V. Garimella, D. Altshuler, S. Gabriel, M.A. DePristo, From fastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline, 2013, https://doi.org/10.1002/0471250953.bi1110s43.

[20] P. Cingolani, A. Platts, L.L. Wang, M. Coon, T. Nguyen, L. Wang, S.J. Land, X. Lu, D. M. Ruden, 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). 6 (2012) 80–92, https://doi.org/10.4161/fly.19695.

[21] Y.I. Li, D.A. Knowles, J. Humphrey, A.N. Barbeira, S.P. Dickinson, H.K. Im, J. K. Pritchard, Annotation-free quantification of RNA splicing using LeafCutter, Nat. Genet. 50 (2018) 151–158, https://doi.org/10.1038/s41588-017-0004-9.

[22] G. Jenkinson, Y.I. Li, S. Basu, M.A. Cousin, G.R. Oliver, E.W. Klee, LeafCutterMD: an algorithm for outlier splicing detection in rare diseases, Bioinformatics. 36 (2020) 4609–4615, https://doi.org/10.1093/bioinformatics/btaa259.

[23] K. Jaganathan, S. Kyriazopoulou Panagiotopoulou, J.F. McRae, S.F. Darbandi, D. Knowles, Y.I. Li, J.A. Kosmicki, J. Arbelaez, W. Cui, G.B. Schwartz, E.D. Chow, E. Kanterakis, H. Gao, A. Kia, S. Batzoglou, S.J. Sanders, K.K.H. Farh, Predicting splicing from primary sequence with deep learning, Cell. 176 (2019) 535–548.e24, https://doi.org/10.1016/j.cell.2018.12.015.

[24] F. Piva, M. Giulietti, A.B. Burini, G. Principato, SpliceAid 2: a database of human splicing factors expression data and RNA target motifs, Hum. Mutat. 33 (2012) 81–85, https://doi.org/10.1002/HUMU.21609.

[25] G. Borck, R. Redon, D. Sanlaville, M. Rio, M. Prieur, S. Lyonnet, M. Vekemans, N. P. Carter, A. Munnich, L. Colleaux, V. Cormier-Daire, NIPBL mutations and genetic heterogeneity in Cornelia de Lange syndrome, J. Med. Genet. 41 (2004) 1–6, https://doi.org/10.1136/jmg.2004.026666.

[26] K. Zarnack, J. König, M. Tajnik, I. Martincorena, S. Eustermann, I. Stévant, A. Reyes, S. Anders, N.M. Luscombe, J. Ule, Direct competition between hnRNP C and U2AF65 protects the transcriptome from the exonization of Alu elements, Cell. 152 (2013) 453–466, https://doi.org/10.1016/j.cell.2012.12.023.

[27] M. Feracci, J.N. Foot, S.N. Grellscheid, M. Danilenko, R. Stehle, O. Gonchar, H. S. Kang, C. Dalgliesh, N.H. Meyer, Y. Liu, A. Lahat, M. Sattler, I.C. Eperon, D. J. Elliott, C. Dominguez, Structural basis of RNA recognition and dimerization by the STAR proteins T-STAR and Sam68, Nat. Commun. 7 (2016), https://doi.org/10.1038/ncomms10355.

[28] S. Subramania, L.M. Gagné, S. Campagne, V. Fort, J. O'Sullivan, K. Mocaer, M. Feldmüller, J.Y. Masson, F.H.T. Allain, S.M. Hussein, M.É. Huot, SAM68 interaction with U1A modulates U1 snRNP recruitment and regulates mTor pre-mRNA splicing, Nucleic Acids Res. 47 (2019) 4181–4197, https://doi.org/10.1093/nar/gkz099.

[29] C. Sánchez-Jiménez, J.M. Izquierdo, T-cell intracellular antigens in health and disease, Cell Cycle 14 (2015) 2033–2043, https://doi.org/10.1080/15384101.2015.1053668.

[30] P.L. Deininger, M.A. Batzer, Alu repeats and human disease, Mol. Genet. Metab. 67 (1999) 183–193, https://doi.org/10.1006/mgme.1999.2864.

[31] X. Song, C.R. Beck, R. Du, I.M. Campbell, Z. Coban-Akdemir, S. Gu, A.M. Breman, P. Stankiewicz, G. Ira, C.A. Shaw, J.R. Lupski, Predicting human genes susceptible to genomic instability associated with Alu/Alu-mediated rearrangements, Genome Res. 28 (2018) 1228–1242, https://doi.org/10.1101/gr.229401.117.

[32] C. Bancells, O. Llorà-Batlle, A. Poran, C. Nötzel, N. Rovira-Graells, O. Elemento, B. F.C. Kafsack, A. Cortés, Revisiting the initial steps of sexual development in the malaria parasite plasmodium falciparum, Nat. Microbiol. 4 (2019) 144–154, https://doi.org/10.1038/s41564-018-0291-7.

[33] B. Wang, Y. Zhang, T. Qing, K. Xing, J. Li, T. Zhen, S. Zhu, X. Zhan, Comprehensive analysis of metastatic gastric cancer tumour cells using single-cell RNA-seq, Sci. Rep. 11 (2021) 1–10, https://doi.org/10.1038/s41598-020-80881-2.
[34] S.H. Yip, P.C. Sham, J. Wang, Evaluation of tools for highly variable gene discovery from single-cell RNA-seq data, Brief. Bioinform. 20 (2018) 1583–1589, https://doi.org/10.1093/bib/bby011.
[35] S.A. Huisman, E.J.W. Redeker, S.M. Maas, M.M. Mannens, R.C.M. Hennekam, High rate of mosaicism in individuals with Cornelia de Lange syndrome, J. Med. Genet. 50 (2013) 339–344, https://doi.org/10.1136/jmedgenet-2012-101477.