Supplementary Materials and Methods

Exome/genome sequencing and bioinformatics

Representative exome and genome sequencing procedures and bioinformatic analyses:

Individual 1: Exome sequencing was performed on genomic DNA from a peripheral blood sample. Libraries were prepared using Agilent SureSelect Human All Exon v5 (Agilent Technologies, Santa Clara, CA) and sequenced on a HiSeq 2000 instrument (Illumina, San Diego, CA) according to the manufacturer's recommendations for paired-end 101-bp reads. A mean depth of 75.69 x was reached and 93.7% of the RefSeq exons were covered by at least by 10 reads. Variants were identified using a computational platform of the FHU Translad, hosted by the University of Burgundy Europe Computing Cluster (CCuB). Raw data quality was evaluated by FastQC software (v0.11.4 – http://www.bioinformatics.babraham.ac.uk/projects/fastqc). Reads were aligned to the GRCh37/hg19 human genome reference sequence using the Burrows–Wheeler Aligner (v0.7.15) (1). Aligned read data underwent the following steps: (a) duplicate paired-end reads were removed by Picard software (v2.4.1 – http://broadinstitute.github.io/picard/), and (b) base quality score was recalibrated by the Genome Analysis Toolkit (GATK v3.8) base recalibrator (2,3). Using GATK Haplotype Caller, Single Nucleotide Variants with a quality score >30 and an alignment quality score >20 were annotated with SnpEff (v4.3) (4). Rare variants were identified by focusing on nonsynonymous changes present at a frequency of less than 1% in the GNOMAD database (5). Copy Number Variants were detected using xHMM (v1.0) (6) and annotated using in-house python scripts. They were filtered by frequency in public

databases (DGV, ISCA, DDD). The analysis of OMIM morbid genes failed to identify pathogenic or likely pathogenic variants accounting for the clinical presentation of Individual 1. A *de novo* start-loss variant was identified in *PTBP1* (7).

Individuals 2, 3, and 19: Exome sequencing was performed following routine diagnostic procedures as described previously (8). Essentially, DNA was sequenced on an Illumina system after exome enrichment with either the Agilent SureSelect All Exon V4 or the Twist Human Core Exome + RefSeq Panel kit. Reads were aligned to the Hg19 reference genome with BWA, and variants were called using GATK and annotated using an in-house developed pipeline.

Individuals 5, 6, and 14: Using genomic DNA from the proband and parents, the exonic regions and flanking splice junctions of the genome were captured using the IDT xGen Exome Research Panel v1.0 (Integrated DNA Technologies, Coralville, IA). Massively parallel (NextGen) sequencing was done on an Illumina system with 100 bp or greater paired-end reads. Reads were aligned to human genome build GRCh37/UCSC hg19, and analyzed for sequence variants using a custom-developed analysis tool. Reported variants were confirmed, if necessary, by an appropriate orthogonal method in the proband and, if submitted, in selected relatives. The additional sequencing technology and variant interpretation protocol have been previously described (9). The general assertion criteria for variant classification are publicly available on the GeneDx ClinVar submission page (http://www.ncbi.nlm.nih.gov/clinvar/submitters/26957/).

Individual 10: Genomic DNA was isolated from amniotic fluid or peripheral blood lymphocytes (parents) using standard techniques. Target sequences were enriched with the KAPA HyperExome (Roche) and sequenced 2x150 bp paired-end by Illumina sequencing (NextSeq 2000). Data were mapped against the reference genome (GRCh37/hg19) using the Burrows–Wheeler Aligner (BWA v0.7.10) and variant calling was done using the Genome Analysis Toolkit (GATK v3.3-0). For analysis, Alissa Interpret v5 was used and variants with an allele frequency >1% in the Genome Aggregation Database (gnomAD, https://gnomad.broadinstitute.org/), but not labeled as likely pathogenic or pathogenic in ClinVar or HGMD, were excluded. Only variants in exons +/- 6 nt were analyzed. Variants were visualized with the Integrative Genomics Viewer (IGV). A *de novo* start-loss variant was identified in the candidate gene *PTBP1* (7).

Individuals 11–13: Genome sequencing (GS) was performed on genomic DNA from peripheral blood samples of an affected mother, her healthy parents and her two affected sons. Libraries for sequencing on either Illumina HiSeq X or NovaSeq 6000 S4 (Illumina Inc, San Diego, CA, USA) were prepared from genomic DNA using the Illumina TruSeq PCR-free kit with a mean insert size of >350 base pairs (bp), resulting in over 432 million (range 329–515M) mapped unique sequences per sample with a mean read depth of 34× (range 29–43×). Alignment of reads to human reference genome (GRCh37/hg19) and variant calling was performed by the National Genomics Infrastructure at Science for Life Laboratory (SciLifeLab, Stockholm, Sweden). The variants were identified using GATK best practices (2), functionally annotated using Variant Effect Predictor (VEP; version 91) (10), loaded into a database using GEMINI

(v0.20.0) (11), explored in GEMINI using built-in tools, and visualized in Integrative Genomics Viewer (IGV) (12). Structural variants were detected using the FindSV pipeline (https://github.com/J35P312/ FindSV) merging calls from CNVnator v0.3.2 and TIDDIT (13,14).

Individuals 17 and 22: Proband and parental peripheral blood samples were submitted for clinical exome sequencing to The Steve and Cindy Rasmussen Institute for Genomic Medicine at Nationwide Children's Hospital. Genomic DNA extraction from peripheral blood and genotyping assay using a custom Agena MassArray panel (Agena) to ensure sample provenance and verify familial relationships was performed as described by Miller et al. (2020) (15). Libraries underwent target capture using SureSelect Human All Exon V6 (Agilent) followed by paired-end 151-bp sequencing to at least 114× mean depth on a HiSeq 4000 (Illumina), with 94.6% of targeted bases at 20× or greater. Secondary analysis was performed using Churchill 3.0 (16) and variant annotation using VarHouse (an in-house tool). Human Phenotype Ontology (HPO) terms (17) inferred from clinician-provided phenotypes by Mr. Phene (an in-house tool) were used to prioritize variants in phenotypically relevant genes for clinical variant interpretation.

Written informed consent was obtained for Individuals 17 and 22. Individuals and their parents were consented for clinical exome sequencing as well as research studies under a protocol approved by the Institutional Review Board at Nationwide Children's Hospital (IRB18-00662, Gene Discovery in Clinical Genomic Patients).

Individuals 24 and 25: Whole genome sequencing was performed following the recommandations of France's Genomic Medicine Plan. Whole blood-extracted genomic DNA was sequenced according to standard procedures for a PCR-Free genome on a NovaSeq6000 instrument (Illumina). Sequencing data were aligned to the GRCh38p13 full assembly using bwa 0.7+. Variants were called by several algorithms including GATK4+, Bcftools1.10+, Manta1.6+, and CNVnator0.4+, and annotated using the variant effect predictor. Detected variants were prioritized using in-house procedures. Further details are available on request from http://www.auragen.fr.

Individual 29: The patient was referred to our pediatric genetics clinic for diagnostic work-up for autism spectrum disorder. Since there was no specific clinical diagnosis, trio whole exome sequencing (WES) on genomic DNA extracted from EDTA blood from the patient and both parents was performed using the Agilent SureSelectXT Kit v6 capture kit with paired-end sequencing (HiSeq SBS Kit v4, 125 Fwd-125 Rev) on a HiSeq2500 sequencing platform (Illumina Inc) (Q30-value: 92.79). NextGene v2.4.2.2 (Softgenetics) was used for data processing and analysis. Since no apparently causative variant was detected, the patient was included in our research study to unravel novel causes of intellectual disability (ethical approval by the ethics committee of the canton of Zurich (StV 11/09 and PB_2016-02520)) after informed consent was obtained. The study was funded by the Swiss National Science Foundation (SNSF) grant 320020_179547 (A.R.).

Computer-assisted facial composite

Facial composite of the PTBP1 cohort was generated using the Facer toolkit (https://github.com/johnwmillr/Facer), which is available under the MIT Licence.

RNA sequencing, data alignment and gene reads counting

RNA libraries were prepared by IntegraGen (Evry, France) following the NEBNext UltraTM II mRNA-seq kit from a minimum of 500 ng. Polyadenylated mRNAs were sequenced paired-end (2 × 100 bp) with an average of 80 M clusters on a NovaSeq 6000 (Integragen, Evry, France). Sequencing data were aligned against the reference genome (GrCh37/Hg19) with STAR aligner (version 2.5.2) using the two-pass mode. Gene annotation was done with the RefSeq database (2022-10-28 version) (18,19). To estimate gene expression, reads per gene were counted with the STAR quantMode option set as GeneCounts.

Assessment of exon skipping events on PTBP1 targets using cDNA amplicon sequencing

Total RNA was extracted from cells using trizol, and reverse-transcribed into cDNA using QuantiTect Reverse Transcription Kit (Qiagen). Targeted regions were amplified by PCR using Primestar GXL DNA Polymerase (Takara) and specific primers (listed below) designed to span exon–exon junctions. PCR products were purified with AMPure XP beads (Beckman Coulter) and subjected to library preparation with Nextera XT library preparation kit (Illumina) and MiSeq sequencing (Illumina).

The qualities of paired-end reads and bases were assessed using FastQC (v0.12.1) from the FASTQ files, and showed a mean phred quality score above thirty on all read positions for all the samples. Reads from the fastq files were then aligned to the

GRCh38 human reference genome using TopHat2 (v2.1.1) (20) with the default options. Transcript isoforms were assembled *de novo* from the resulting bam files using the cufflinks R package (v2.2.1) with the default configuration to identify and quantify aberrant exon inclusion/skipping events. Transcripts from the cufflinks output were annotated using R packages TxDb.Hsapiens.UCSC.hg38.knownGene_3.18.0 and org.Hs.eg.db_3.19.1. The R package ggtranscript (v1.0.0)(21) was used to help visualize alternative splicing events.

PBX1_cDNA_ex1-9F AGCAGGACATTGGAGACATTTTA

PBX1_cDNA_ex1-9R CACTGATGAAGGGGTAGTAGCAT

DLG4_cDNA_ex2-20F ATACCGCTACCAAGATGAAGACA

DLG4_cDNA_ex2-20R TGTGGTAGATCTCCTCAAAGCTG

PTBP1_cDNA_ex3-15F TGACGAGCTTTTCTCTACTTGTG

PTBP1_cDNA_ex3-15R CATCTGGATCAGTGCCATCTT

Splicing and RIP-seq data analysis

Splicing anomalies were detected with the *rMATs* tool with default parameters (4.0.2 version)(22). Sashimi plots were generated with *ggsashimi* and illustrate the aberrant events found in all biological replicates of PTBP1 siRNA-treated conditions and absent from all other conditions (23). The Y-axis was fixed for each event and scaled according to the lowest read count among all samples. Indicated values were averaged per group. Raw gene counts normalization using the median of ratios method and differential gene expression analysis based on the negative binomial regression model were performed with the *DESeq2* R package (24). Genes with a read count below ten across all samples were excluded from the analysis. To identify

PTBP1's targets, the count table was restricted to the WT and IgG samples (three Input, three IPs and one IgG) and genes were identified as associated with PTBP1 when their expression levels in the PTBP1 IP fraction were significantly 1.5 times higher than the input fraction and the IgG fraction (adjusted *p*-value ≤ 0.05 — Benjamini–Hochberg correction). To compare gene expression between WT and start-loss samples, the count table included six WT samples (three inputs and three IPs) and six start-loss samples (three inputs and three IPs). For data exploration and visualization, a variance-stabilizing transformation was first applied to raw counts to ensure homoscedasticity. Principal component analysis (PCA) and hierarchical clustering were then performed respectively with the *prcomp* and *hclust* R functions following the complete linkage method on the 5% most variable expressed genes. Genes were identified as up or down-regulated when their fold change (FC) exceeded $\log_2(1.5)$ and their adjusted *p*-value ≤ 0.05. Venn diagrams were generated with the *nVennR* R package (25).

Gene and protein set enrichment analysis (G/PSEA)

Genes or proteins were ranked according to the logarithm of their adjusted *p*-value combined with the sign of their FC. GSEA was performed with the *clusterProfiler* R package(26) by interrogating either the WikiPathways (2024-10-10 version) or Gene Ontology database (2023-01-01 version) (27–29). Biological pathways with an adjusted *p*-value ≤ 0.05 were reported. GSEA was performed three times with a ranked list generated from the input samples comparison, the IP comparison and from a likelihood ratio test to identify differently enriched genes in the IP fraction depending on the genotype (WT or start-loss). For PSEA, 0.25% of the gene list included protein

isoforms (5/1942) and therefore corresponding genes were counted twice. GO terms of interest were selected manually before removing redundant terms by keeping those with the lowest adjusted *p*-value among terms with a semantic similarity score > 0.4 (Wang method) using the *GOSemSim* R package(30,31). Circular plots were generated with the *circlize* R package(32).

Nano LC-MS/MS analysis

The analysis was performed using nanoACQUITY Ultra-Performance-LC (UPLC, Waters). Peptide digests were trapped on a Symmetry C18 pre-column (C18, 180 µm x 20 mm, 5 µm particle size, Waters) and the peptides were separated on an ACQUITY UPLC® BEH130 C18 separation column (C18, 75 µm × 250 mm, 1.7 µm particle size, Waters). The solvent system consisted of 0.1% formic acid in water (solvent A) and 0.1% formic acid in acetonitrile (solvent B). Peptide trapping was performed for 3 minutes at a flow rate of 5 µL/min with 99% A and 1% B and elution was performed at 60°C at a flowrate of 350 nL/min from 8% to 35% of B in 38 minutes. The mass spectrometer was operated in positive mode, with the following settings: spray voltage 1800 V and capillary temperature 250°C. The MS scan had a resolution of 70000, the AGC target was $3x10^6$, and the maximum IT was 50 ms on the m/z [300–1800] range. The MS/MS scans had a resolution of 17500, the AGC target was $1x10^5$, and the maximum IT was 100 ms with a fixed first mass of 100 m/z and Isolation window of 2 m/z. Top-10 HCD was selected with an MS2 identity threshold of $2x10^5$ and dynamic exclusion of 60 s. The normalized collision energy (NCE) was fixed at 27 V. The complete system was fully controlled by Thermo Fisher Scientific™ Xcalibur™

software. The raw data collected were processed and converted with MSConvert into .mgf peak list format.

Site-directed mutagenesis and cloning

Site-directed mutagenesis was performed on commercially available plasmids containing either the open reading frame of PTBP1-4 (NM_002819) or PTBP1-1 (NM_031991) in-frame with a C-terminal turbo-GFP under the control of a cytomegalovirus (CMV) promoter (pCMV6-PTBP1-4/PTNP1-1-tGFP OriGene CAT#: RG201779 and RG202564 respectively). Variants NM_002819.5:c.2T>C (p.Met1? or Met1 start-loss/M1-sl), NM_002819.5:c.41G>A (p.Lys14Gln), NM_002819.5:c.137A>G (p.Lys46Arg), NM_002819.5:c.137A>C (p.Lys46Thr), and NM_002819.5:c.144A>T (p.Lys48Asn) were introduced by PCR using the following reaction mix in a final volume of 10 µL: 20 ng of plasmid, 1 µL of reaction buffer (600380, Agilent), 1 µL of each primer (10 µM), 2 µL of dNTP (2.5 mM), 0.5 µL of DMSO, and 0.2 µL of PfuUltra High-fidelity (600380, Agilent). Mutagenesis was performed using the following steps: denaturation 1 minute at 95°C; cycling (18 X) 50 seconds at 95°C, 50 seconds at 60 °C, 1 min/kb at 68°C, and a final elongation 7 minutes at 68°C. HPLC-purified primers were used (IDT, Iowa, USA). The sequence of primers used for site-directed mutagenesis are indicated below:

PTBP1 _M1_sl_FW CTGCCGCCGCGATCGCCACGGACGGCATTGTCCCAG

PTBP1 _M1_Rv CTGGGACAATCGCGTCCGTGGCGATCGCGGCGGCAG

PTBP1_Lys14Gln_Fw GCCGTTGGTACAAAGCAGGGATCTGACGAGCTTTTCTC

PTBP1_Lys14Gln_Rv GAGAAAAGCTCGTCAGATCCCTGCTTTGTACCAACGGC

PTBP1_Lys46Thr_Fw CGGAAATGACAGCAAGACGTTCAAAGGTGACAGCCG

PTBP1_Lys46Thr_Rv CGGCTGTCACCTTTGAACGTCTTGCTGTCATTTCCG

PTBP1_Lys46Arg_Fw CGGAAATGACAGCAAGAGGTTCAAAGGTGACAGCCG

PTBP1_Lys46Arg_Rv CGGCTGTCACCTTTGAACCTCTTGCTGTCATTTCCG

PTBP1_Lys48Asn_Fw GACAGCAAGAAGTTCAATGGTGACAGCCGAAGTGC

PTBP1_Lys48Asn_Rv GCACTTCGGCTGTCACCATTGAACTTCTTGCTGTC

PCR products were digested with *Dpn*1 (R0176S, NEB) to remove the native plasmid and used for Top-10 competent bacteria transformation (C404010, Invitrogen).

pCMV6-PTBP2-tGFP (NM_021190) was purchased from OriGene (CAT#: RG202163). Variants NM_021190.4:c.2T>C (p.Met1? or M1 start-loss/M1-sl) and NM_021190.4:c.41G>C (PTBP2_p.Arg14Thr) were synthesized and subcloned by Eurofins Genomics.

After bacterial transformation, clones were selected on agar plates containing 100 µg/mL ampicillin (11593027, Thermo Fisher Scientific). Ten clones per plasmid preparation were amplified in liquid medium (A5306, Sigma) and extracted using QIAprep Spin Miniprep Kit (27106, Qiagen). The presence of the variants was verified by Sanger sequencing.

Cell culture

NIH-3T3 cells and human primary fibroblasts from healthy and affected individuals were cultured in DMEM High Glucose Medium 4.5 g/L (11965-092, DMEM Gibco, Thermo Fisher Scientific, Waltham, MA, USA) supplemented with 10% Fetal Bovine Serum (FBS, 10270-106, Thermo Fisher Scientific) and 1% ZellShield (commercial mixture of antibiotics and anti-mycoplasma reagents) (130050 Minerva, Biovalley, France). Cells were cultured at 37°C in a humidified 5% CO_2 atmosphere. The

following cell lines/DNA samples were obtained from the NIGMS Human Genetic Cell Repository at the Coriell Institute for Medical Research: GM05565 and GM05399.

Plasmid transfection

Plasmid transfections were performed with jetOPTIMUS according to the supplier's recommendations (101000025, Polyplus) using 2 µg of plasmid. NIH-3T3 cells were seeded 24 hours prior to transfection in 6-well plates at $1.5x10^5$ cells per well. Analysis was performed 24 hours after transfection.

Protein extraction and western blot

For whole cell lysate, cells were harvested by trypsinization and lysed in 1x RIPA buffer (9806S, Cell Signaling) supplemented with PMSF (1 µM) and a protease inhibitor cocktail (93482, Sigma). After 15 minutes on ice and centrifugation at 13000 g at 4°C for 15 minutes, the total lysate was recovered. For nuclear and cytoplasmic compartment isolation, the NE-PER Nuclear and Cytoplasmic Extraction Reagent kit (78833, Thermo Fisher Scientific) was used according to the manufacturer's instructions. Proteins were dosed using the BCA kit (23225, Thermo Fisher Scientific). 40 µg of proteins from cell lysates were run on 10% SDS-PAGE gel and migration performed in 25 mM Tris buffer, 192 mM Glycine, 0.10% SDS at 80 V for 30 minutes and then 120 V for 3 hours. The proteins were transferred for 1 hour at 120 V in Tris-Boric buffer (Tris 49 mM, Boric acid 48.5 mM) onto a Millipore PVDF membrane (IPVH00010, Immobilon-P) previously activated with methanol. The membrane was saturated with a solution of TBS, 0.05% Tween 20, and 5% BSA for 30 minutes. Primary antibodies against PTBP1 (Abcam, Ab133734, 1:2000 dilution), PTBP2

(Millipore, SAB1405229, 1:500 dilution), TurboGFP (Thermo Fisher Scientific, PA5-22688, 1:2000 dilution), Vinculin (Millipore, V9131, 1:5000 dilution), or Lamin A/C (Cell Signaling, 477S, 1:2000) were added overnight at 4°C. After 3 washes, HRP-conjugated secondary antibody was added for 1 hour. The labeled proteins were detected using ECL Clarity™ substrate (170-5061, Bio-Rad Laboratories, Inc) and bands on western blots were visualized using ChemiDoc Imaging System (Bio-Rad). The molecular weights and intensity of the bands corresponding to proteins of interest were determined using Image Lab software (Bio-Rad). The values obtained for band quantification were normalized according to the intensity of the loading control of the corresponding protein fraction.

To ensure the accuracy and reliability of results, experiments were performed at least in triplicate.

Cycloheximide Chase Assay

Control or patient-derived fibroblasts were seeded at 2x105 cells per well in 6-well plates. The cells were treated with 50 ug/mL cycloheximide (Sigma), and the proteins were extracted after 0, 8, and 16 hours of treatment. Protein fractions were analyzed by immunoblot. The abundance of PTBP1 was then measured and compared with the loading control.

To ensure the accuracy and reliability of results, experiments were performed at least in triplicate.

Immunofluorescence staining and Duolink studies on cells

Cells were seeded for 24 hours before labeling in 6-well plates containing coverslips, $2x10^5$ fibroblasts, or $1.5x10^5$ NIH-3T3 per well. Cells were washed with cold PBS and fixed with 4% formaldehyde for 20 minutes at room temperature. Cells were then incubated with NH_4Cl (50 mM) solution for 10 minutes, followed by permeabilization with 0.3% PBS-Triton solution for 7 minutes, still at room temperature. Saturation was performed with 1% PBS-BSA, 0.1% Tween 20, for 15 minutes at room temperature. Primary antibodies were incubated at 4°C overnight, diluted in blocking buffer. The next day, after 3 washes with cold PBS, secondary antibodies (1:1000 dilution) were added for 45 minutes. After three final washes, nuclei were stained using Hoechst. Actin was stained using ActinGreen 488 ReadyProbes reagent (Invitrogen), following the manufacturer's instructions. The following antibodies were used: anti-PTBP1 (MABE986, Millipore, 1:500 dilution), anti-DCP1a (WH0055802M6, Sigma, 1:500 dilution), anti-G3BP1 (05-1938, Millipore, 1:500 dilution), anti-EEA1 (GTX634169, GeneTex, 1:500 dilution), and anti-PIST (GTX64526, GeneTex, 1:200 dilution).

For more advanced colocalization studies, the Duolink In Situ Detection Reagents Orange kit was used, in combination with PLA Probe Anti-Rabbit PLUS (DUO92002, Sigma) and Anti-Mouse MINUS (DUO92004, Sigma). After Phansalkar thresholding and focusing on the cytoplasmic area, images were analyzed using the Analysis Particles tools from Image J software.

Microscopic studies were conducted with a Zeiss fluorescence AX10 microscope and a confocal Leica DMi8 microscope at the DImaCell imaging INRAE platform in Dijon. Quantification was obtained by calculating a Corrected Total Cell Fluorescence (CTCF) parameter with the following equation: CTCF = integrated density - (area of selected cell x mean fluorescence of background (33,34). For colocalization analysis

between DCP1A and PTBP1 or PTBP2, in the cytoplasmic compartment (region of interest, ROI), the Pearson's correlation coefficient (R^2) was obtained using the Jacop plugin in ImageJ software, after subtracting the nuclear area from the total cell area. Quantification was performed on thirty cells per condition. A uniform intensity threshold was applied prior to measuring colocalization to ensure standardized comparison across experimental groups(35–38).

To ensure the accuracy and reliability of results, experiments were performed at least in triplicate.

siRNA transfection

$1.5x10^5$ fibroblasts per well in 6-well plates were seeded and transfected for 48 hours: 5 µL of transfection agent Lipofectamine RNAimax (Thermo Fisher Scientific) in a final volume of 250 µL of Opti-MEM (Gibco) was combined with 250 µL of Opti-MEM containing 100 pmol of control siRNA (sc-37007-SantaCruz or SIC001-Sigma-Aldrich) or siRNA specific to PTBP1 transcript (sc-38280-SantaCruz). 500 µL of the transfection mixture was added to 2 mL of complete medium (DMEM 4.5 g/L, FBS 10%, ZellShield 1%) with a cell concentration of $5x10^4$ cells/mL. Knockdown of PTBP1 was verified by Western blot and RNAs were subsequently isolated.

RNA immunoprecipitation and isolation

Cells were harvested by trypsinization, counted, and rinsed with cold PBS. 8 µL per million cells of a complete lysis buffer containing protease inhibitor cocktail and RNase inhibitor provided by the EZ-MAGNA RIP RNA-binding protein immunoprecipitation kit (Millipore) were used to lyse cells on ice before freezing at -80°C. Cell lysates were

incubated overnight at 4°C with the RIP immunoprecipitation buffer containing magnetic beads coated with 250 ng of either the PTBP1 antibody (Millipore – MABE986) or a normal mouse IgG included in the kit. 10% of cell lysate was kept aside as input. For the IgG IP fraction, cell lysates of three WT lines were pooled before IP because low material quantity was expected. Beads were then washed with RIP wash buffer before RNAs were released from the bead/antibody/protein complexes with heat and protein digestion. RNAs were isolated with a phenol/chloroform/isoamyl alcohol extraction (25:24:1, v/v) followed by an ethanol precipitation. Purified RNAs were stored at -80°C until sequencing.

DNA methylation data generation and processing

The PTBP1 cohort consists of eleven individuals: four males and seven females with ages ranging from one to 27 years old (median = seven), with confirmed variants of interest in PTBP1 (two missense and nine start-loss variants). DNA from sampled peripheral blood of all individuals in the cohort was subjected to bisulfite conversion using the Illumina Infinium MethylationEPIC v1 BeadChip array according to the manufacturer's protocols (Illumina, San Diego, CA, USA). Data preprocessing and methylation analysis were performed based on previously published standard protocols (39,40). All computations were implemented using R statistical software (R Core Team. R: A Language and Environment for Statistical Computing. Published online in 2022. https://www.R-project.org/, version 2.4.1) and its associated libraries. Intensity data files containing methylated and unmethylated signal intensities were subjected to quality control using the SeSAMe package (41). Standard preprocessing includes dye bias correction, background subtraction, quality probe masking for

Illumina microarrays (42), and detection *p*-value probe masking based on Infinium I out-of-band signal calibration (p=0.05).

Episignature discovery

Discovery analysis for signature identification was performed using the nine case samples with confirmed start-loss mutations. Prior to analysis, probes meeting any one of the following conditions were also removed: probes in the X or Y chromosome, probes targeting CpG sites overlapping known SNPs, cross-reactive probes, high-variability probes from prior Illumina DNAm array products after manufacturing change, and EPICv1 probes excluded in the new EPICv2 array. This resulted in the removal of 169,708 probes. Matched controls were selected for the discovery analysis using the MatchIt package (43), and matching was based on age, sex, and array type used to generate the methylation data. Existing samples in the EpiSign™ Knowledge Database (EKD) previously identified to cause batch effects and samples with more than 5% failed probes were excluded in the matching process. A total of 54 matched controls were selected for the analysis (match ratio 1:6), and principal component analysis (PCA) was used to investigate data structure and outliers.

Differential analysis was performed on the discovery cohort using the limma package (44). Methylation beta values were first converted to M-values using logit transformation, then data was fitted using multivariate linear regression. In the model, methylation values were used as predictors, the case/control labels as responses, and estimated blood cell compositions as covariates. Empirical Bayes was used to compute moderated statistics, and adjusted *p*-values were computed to control for false discovery rates using the Benjamini–Hochberg method. Probes defining the

episignature were selected by inspecting parameters derived from ranking probes based on their feature importance. First, probes were rank selected based on scores computed using the absolute mean methylation difference and the negative log of their adjusted *p*-value. Then, the top subsets ranked based on their scores from a receiver operating characteristic (ROC) curve analysis implemented using the caret package (45) were retained. Finally, highly correlated pairwise probes were removed. Consequent probe lists dependent on the rank divisions were assessed using unsupervised clustering and visualization of Euclidean clustering using Ward's method and multidimensional scaling. Plots were generated using R ggplot2 (Wickham H. Ggplot2: Elegant Graphics for Data Analysis. Springer–Verlag New York; 2016. https://ggplot2.tidyverse.org) and gplots (Warnes GR, Bolker B, Bonebakker L, et al. gplots: Various R Programming Tools for Plotting Data. Published online in 2022. R package version 3.1.3) packages, and the probe set with the optimal clustering results was selected to define the signature. Leave-one-out cross-validation on the discovery case samples was also performed to check the reproducibility and robustness of the identified profile. Binary prediction models were developed and trained using support vector machines (SVMs) with the R e1071 package (Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F. e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. Published online in 2023. https://CRAN.R-project.org/package=e1071). Selected signature probes were used as features and discovery samples were used for training, together with 75% of the control and EpiSign™ cohort samples from the EKD, while the remaining 25% was used for testing the classifier performance. This 75%/25% train/test split was repeated four times, leading to a fourfold cross-validation of non-discovery EKD samples. SVM

hyperparameters such as class weights, regularization parameters, and kernel functions were determined using grid search.

Functional annotation and methylome comparison with EpiSign cohorts

Functional annotation and genome-wide methylation profile comparison with existing EpiSign™ disorder cohorts were performed within the EpiSign Knowledge Database ((EKD), following previously published protocols (46). In brief, the pipeline includes matching the cohort cases to healthy and EpiSign™-negative controls, identifying differentially methylated probes (DMPs) and regions (DMRs), annotating results, and investigating the similarity and overlap of the PTBP1 signature with 56 signatures in EpiSign™ version 32. Differentially methylated sites were identified using linear regression analysis and probes present in both 450 K and EPICv1 arrays with at least a 5% mean methylation difference and an adjusted *p*-value of less than 0.01 were selected. DMRs were identified using the DMRcate package (47). Candidate regions were defined such that there are at least five CpG probes with a maximum 1kb distance between any two contiguous probes, while differentially methylated regions were selected as regions with statistically significant mean methylation differences between cases and controls: at least a 5% mean difference and a combined *p*-value using Stouffer transformation of <0.01. Probes in the DMP and DMR results were annotated with CpG- and gene-based annotations using the annotatr (48) and AnnotationHub (Morgan M, Shepherd L., AnnotationHub: Client to access AnnotationHub resources. Published online in 2022.) packages. Gene set enrichment analysis was also performed using the missMethyl package (49). Similarities among signature cohorts based on the intersection between DMP sets were visualized using

a heatmap (pheatmap, Kolde R., pheatmap: Pretty Heatmaps. Published online in 2019. https://CRAN.R-project.org/package=pheatmap package), scatter plot, and circos plot (circlize package) (32). The comparison of signatures using the top 500 DMPs of each cohort was also illustrated using a tree-and-leaf plot (Kume LW, Rizzardi LEA, Cardoso MA, Castro MAA., TreeAndLeaf: Displaying binary trees with focus on dendrogram leaves. R package version 180. Published online in 2022) where the leaf node sizes represent the relative number of total DMPs, and the node colors represent the global methylation profile of the corresponding signature. All differential probes were used for cohorts with fewer than 500 DMPs. Samples for each signature cohort were aggregated, and median probe values were calculated prior to hierarchical clustering using Ward's method and tree-and-leaf plot generation.

Zebrafish husbandry

Danio rerio (AB strain) was raised and bred at a temperature of 28°C with a 14-hour light/10-hour dark cycle. Animal care was provied in strict accordance with protocols approved by the Italian Ministry of Public Health and the University of Pisa Ethics Committee (authorization 99/2012-A, 19.04.2012), in compliance with EU legislation (Directive 2010/63/EU). Zebrafish embryos were obtained by natural spawning, staged according to the hours post-fertilization and raised at 28°C in 1X E3 medium (5.0 mM NaCl, 0.17 mM KCl, 0.33 mM $CaCl_2$, 0.33 mM $MgSO_4$, 0.1% methylene blue) in Petri dishes.

mRNA preparation and injection in zebrafish embryos

WT PTBP1, PTBP1 M1-sl, PTBP1_Lys46Thr, and PTBP1_Lys46Arg variant's full-length cDNAs obtained from site-directed mutagenesis of pCMV6-PTBP1-tGFP (NM_002819) were subcloned into the pCS2+ vector. Capped mRNAs encoding EGFP, wild-type PTBP1, and the previously described variants were synthesized using mMESSAGE mMACHINE SP6 transcription kit (Thermo Fisher Scientific), following the manufacturer's instructions, and injected into the yolk sac of one-cell stage embryos (200 pg per embryo). 200 pg of EGFP-capped mRNA were co-injected to verify successful injections using a fluorescence stereomicroscope and as controls. In all experiments, embryos injected with PTBP1, PTBP1 M1-sl, PTBP1 Lys46Thr, or PTBP1 Lys46Arg mRNAs were compared with embryos injected with EGFP mRNA only at the same developmental stage. Microinjections were performed using a FemtoJet microinjector (Eppendorf).

Zebrafish embryos whole-mount in situ hybridization

Due to teleost's whole-genome duplication, the Zebrafish *ptbp1* is duplicated (*ptbp1a* Gene Bank: NM_001020477.1 and *ptbp1b* Gene Bank: NM_001122654.1). Given the high nucleotide sequence similarity (75% of nucleotide identity) between ptbp1a and ptbp1b transcripts, the RNA antisense probe designed for *in situ* hybridization experiments annealed to both paralogs. The ptbp1 cDNA was amplified from 48-hour post-fertilization wild-type zebrafish cDNA (F: TCCGACGAACTCTTTTCCTCC, R: TTGCTTCTTCCACAGAGGCCA) and cloned into pGEM®-T vector. Digoxigenin-UTP labeled antisense RNA probes were generated by in-vitro transcription with T7 RNA polymerase, according to the manufacturer's instructions (Roche). Wild-type zebrafish embryos at 24 and 48 hours post-fertilization were fixed in 4% paraformaldehyde in

1X PBS overnight at 4°C and stored in methanol at -20°C. Whole-mount in situ hybridization was performed as previously described (50).

Immunofluorescence

Zebrafish embryos injected with the previously described mRNAs were raised till 120 hpf and fixed in 4% paraformaldehyde in 1X PBS, then cryoprotected with 30% sucrose in 1X PBS, and embedded in Tissue-Tek OCT compound (Sakura). Samples were cryosectioned at 12 µm and processed for immunohistochemistry with standard protocols. MABE986 primary antibody was used at 1:500 and incubated at 4°C overnight. 1:500 Rhodamine Red goat anti-mouse-X secondary antibody (Thermo Fisher Scientific) was added for 12 hours at 4°C. After washes, an incubation with Hoechst (Thermo Fisher Scientific) was performed at 1:2000 for 5 minutes in PBS at room temperature. Confocal imaging showing the intracellular localization of human PTBP1 in zebrafish embryo sections was obtained with a Nikon A1 two-photon confocal microscope.

Alcian Blue staining on zebrafish larvae

Cartilage staining with Alcian Blue was performed on zebrafish larvae at five days post-fertilization (dpf). Larvae were fixed in 4% paraformaldehyde in 1X PBS overnight at 4°C. After washing several times in PBS-T (phosphate-buffered saline with 0.1% Tween-20), they were bleached in a mixture of 3% H_2O_2/1% KOH for 30 minutes or until the eyes were sufficiently translucent. Larvae were rinsed twice in PBS-T, transferred into a filtered Alcian Blue solution (1% HCl, 70% ethanol, 0.1% Alcian Blue), and stained overnight. Specimens were cleared in acidic ethanol (5% HCl, 70%

ethanol) for four hours and washed in PBT. Images of the embryos were acquired using a stereomicroscope (SMZ1500, Nikon) equipped with a digital camera (Nikon). The same magnification was always maintained within each control and treated image pair.

Supplemental results

Genetic findings in Individuals 11–13

A novel heterozygous *de novo* start-loss variant in *PTBP1* (NM_002819.4:c.1A>C) was identified in Individual I11, the mother of Individuals I12 and I13, who also harbor the same variant.

DNA methylation analysis: a predominantly hypermethylated profile delineates patients with PTBP1 variants.

We performed discovery analysis using nine samples carrying a PTBP1 start-loss mutation as cases and 54 matched controls. With our exhaustive iterative feature selection process, we identified an episignature by selecting the top 600 scoring samples based on methylation differences and *p*-values, and then narrowing down our selection to the top 300 probes using ROC analysis. No highly correlated probes were removed, and the 300 probes were used to define the signature. The mean methylation difference between cases and controls ranged from 17.38% to -27.31% (median = 5.79%), where 57.33% of the signature probes were hypermethylated (**Table S4**). Cluster analysis outputs using the selected probes and discovery samples validate our feature selection results by showing separate groupings of cases and controls in both heatmap and multidimensional scaling plots. Furthermore,

unsupervised leave-one-out cross-validation on the discovery cases showed each hold-out test case clustering with the remaining discovery cases for each iteration in the heatmaps, and closer to case group centroids in the MDS plots (**Figure S4A**), which supports the reproducibility and robustness of the identified signature.

We developed an SVM-based model and tested its diagnostic prediction utility for PTBP1-related disorder. The discovery samples together with 75% of the non-discovery EKD samples were used as training data and signature probes were used as features. SVM scores, henceforth referred to as methylation variant pathogenicity (MVP) scores, were computed and averaged over the fourfold cross-validation sets to assess the similarity of the test samples' methylation profiles to the signature profile. Here, higher scores are indicative of more similar profiles to cases. Using this metric together with cluster analysis, we were able to classify samples with high confidence and found the model to be highly specific given that almost all EKD test samples obtained scores close to 0 (**Figure S4B**). The model was also sensitive to PTBP1 test samples with missense mutations, with two of the test missense cases obtaining an MVP score > 0.1, clustering with discovery cases in the MDS plot, and showing similar profiles to discovery cases in the heatmap (**Figures 3A–B**).

Methylome comparison and annotation of the PTBP1 cohort with EpiSign disorders

We performed functional correlation analysis using all the PTBP1 samples that share the signature profile (eleven cases: nine start-loss and two missense samples). Matched controls were selected from EpiSign^{TM}-negative samples and healthy controls using a case-control ratio of 1:5. Our differential analysis together with the

previously outlined DMP selection criteria resulted in 2,831 differential probes. The global methylation profile was observed to be slightly hypermethylated, which is also concordant with the episignature discovery, as 52% of the DMPs have a positive mean methylation difference between cases and controls (**Table S5**), with percentage differences ranging from 5%–20% (**Figure S5A**). Gene set enrichment analysis of the DMPs did not return any statistically significant terms (data not shown). Next, we annotated the genomic location of the differential probes. We found that 21% of the selected probes are in CpG islands, 30% in Shores, 10% in Shelves, and 39% are in all other regions (**Figure S5B**). In relation to genes, 42% are in gene-coding sequence regions, 24% in promoter regions, 15% in 1–5 kb upstream of transcription start sites, and 19% are intergenic (**Figure S5B**). Using the Chi-square goodness-of-fit test, these annotation distributions were found to be significantly different from the respective background probe annotation distributions (p<0.001 for both CpG-based and genic annotations). Subsequent inspection of the correlation with 56 EpiSign™ cohorts revealed that PTBP1 shares the most DMPs, indifferent of methylation status, with Sotos (NSD1; 35%), RMNS (Rahman Syndrome, HIST1H1E; 33%), and TBRS (Tatton-Brown–Rahman Syndrome, DNMT3A; 29%) (**Figures S6A–B**). We further investigated the similarity between disorder cohorts by considering only the top 500 DMPs for each group and visualized the clustering results of the aggregated median values using a tree-and-leaf plot. This agglomerative approach revealed that the PTBP1 cohort was most similar to KDVS (Koolen–De Vries Syndrome, KANSL1), and also forms a sub-cluster with ARTHS (Arboleda–Tham Syndrome, KAT6A), SBBYS (Ohdo Syndrome, KAT6B), RSTS1 and 2 (Rubinstein–Taybi Syndrome, CREBBP,

EP300), HVDAS_C (Helsmoortel–Van der Aa Syndrome, ADNP), and GTPTS (Genitopatellar Syndrome, KAT6B) (**Figure S6C**).

PTBP1 differentially hypomethylated region

Using DMRcate, we found one differentially methylated region distinguishing PTBP1 cases from controls. The identified DMR (chr7:27169208-27171401) consists of nineteen probes and spans a CpG island in chr7, overlapping a group of genes: *HOXA-AS2*, *HOXA-AS3*, *HOXA3*, *HOXA4*, and *RP1-170O19.22*. Cases were hypomethylated compared to controls in said region, with a mean methylation difference of 11.2% (maximum difference of 17.3%) and a combined *p*-value Stouffer transformation of the FDRs of individual CpGs in the DMR equal to 9.65x10-26 (**Figure S6D**).

Supplemental Figures

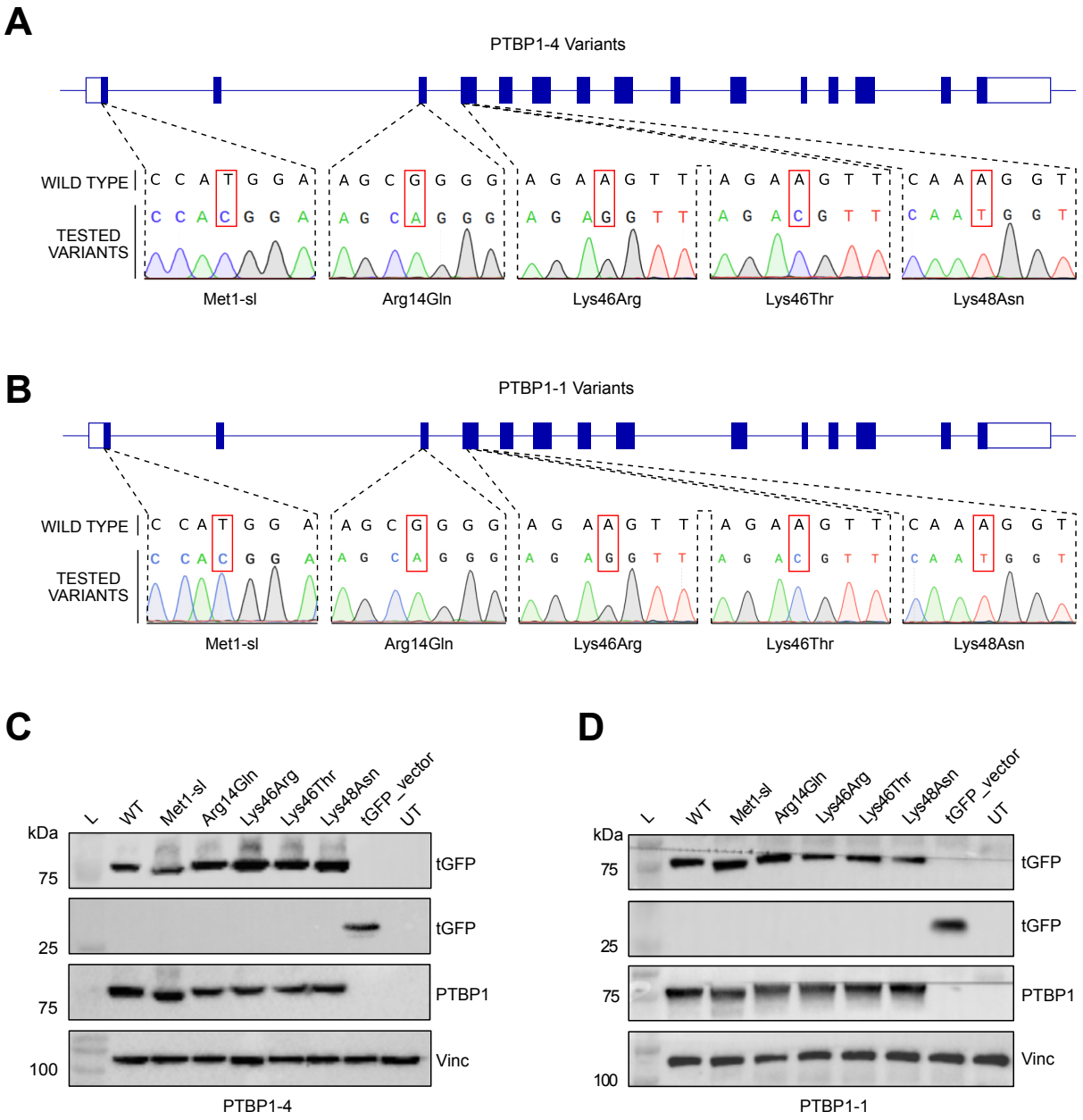
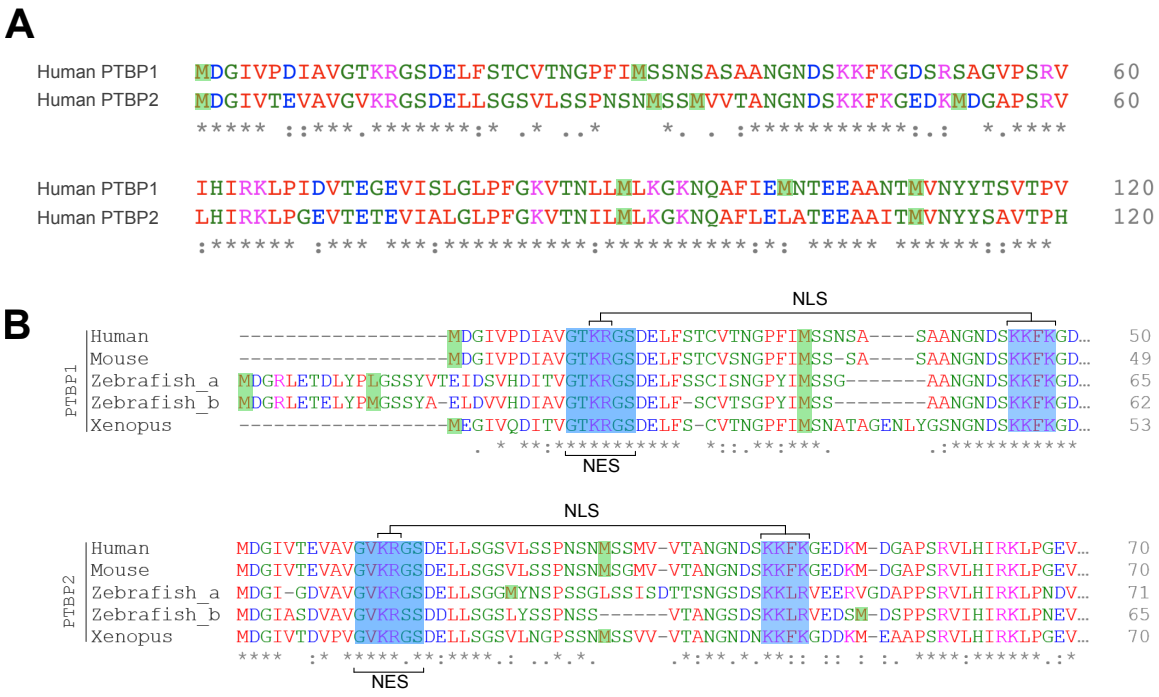


Figure S1. Site-directed mutagenesis of human PTBP1 variants. (A) Representation of variations generated in PTBP1-4 or **(B)** PTBP1-1 plasmids and verified by Sanger sequencing. Western blot performed on proteins extracted from NIH-3T3 cells transfected with **(C)** PTBP1-4 or **(D)** PTBP1-1 constructs and tGFP vector. PTBP1 and tGFP appear at 80 kDa (PTBP1 + tGFP). Note that the lower band represents proteins with a start-loss variant lacking the first thirty aa. The band at 28 kDa corresponds to the empty tGFP vector. Vinculin is used as a loading control. L, protein ladder. UT, untransfected.



Protein	Frame	Methionine	Reliability	Kozak motif A/GXXATGG	ORF Length (aa)
PTBP1	1	1	0.84	GXXATGG	557
	1	31	0.63	AXXATGa	527
PTBP2	1	1	0.34	GXXATGG	531
	1	32	0.15	AXXATGa	500
	1	53	0.21	AXXATGG	479

Figure S2. PTBP1 and PTBP2 characterization of translation initiation. (A) Alignment of the first 120 N-terminal amino acids of human PTBP1 and PTBP2 proteins showing multiple methionines within the same reading frame (highlighted in green). **(B)** Alignment of the most N-terminal amino acid sequences of PTBP1 and PTBP2 proteins showing the conserved cell localization signals across different species. NLS, Nuclear Localization Signal. NES, Nuclear Export Sequence. Methionines of the same reading frame are highlighted in green. **(C)** *In silico* Kozak consensus sequence prediction of the first methionines of PTBP1 and PTBP2 by the tool ATGpr. **(D)** Representation of the different mutated PTBP1 methionines (i.e., Met1Thr and Met31Thr), mutagenesis verification by Sanger sequencing, and protein

expression assessment by Western blot. Notice the shifted bands according to the specific methionine translation initiation. The PTBP1 epitope recognized by PTBP1 antibody is indicated and is located upstream of the alternative start codon at Met90.

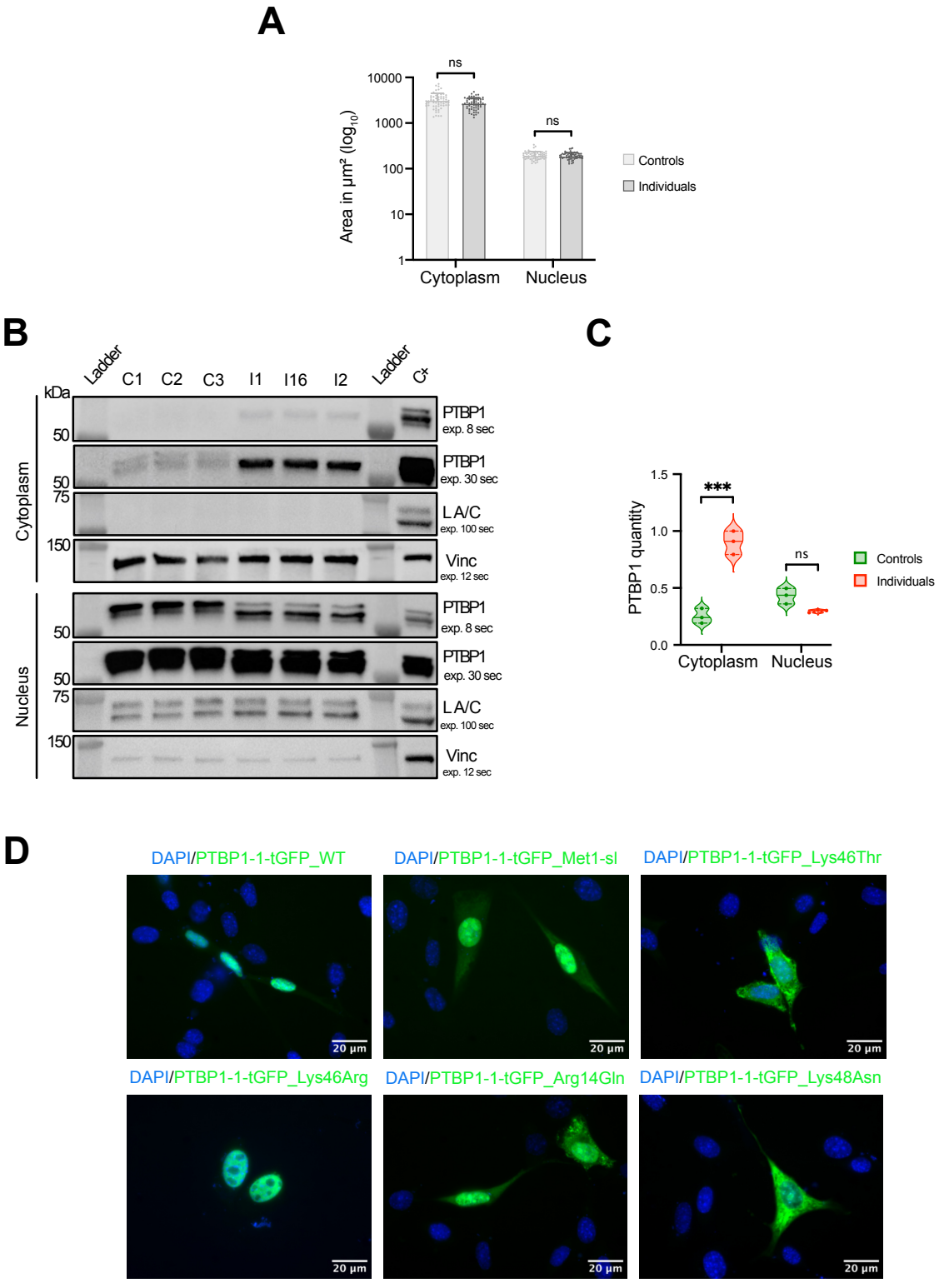


Figure S3. Assessment of PTBP1 nucleocytoplasmic distribution in fibroblasts. **(A)** Quantification of nuclear and cytoplasmic areas in fibroblasts derived from control and PTBP1 start-loss individuals. Measurements show no difference in either compartment in patient- versus control-derived fibroblasts. Sixty cells per condition were analyzed, with statistical analysis using the Mann–Whitney U test. **(B)** Western

blot analysis on cytosolic and nuclear protein fractions, extracted from control fibroblasts and fibroblasts from affected individuals with start-loss variants. Each blot was developed at two exposure times (eight and thirty seconds). Minimal residual expression of Vinculin in the nuclear fraction indicates minor cross-contamination. Positive control (C+) corresponds to the whole cell lysate. (**C**) Quantification of band intensities shown in (B). Statistical analysis was performed using the Mann–Whitney U test. **(D)** Fluorescence imaging of transfected NIH-3T3 cells expressing recombinant WT, start-loss, or missense PTBP1-1-tGFP plasmids (green) and stained with DAPI (blue). WT and control variant Lys46Arg-PTBP1 show nuclear localization. Start-loss and missense variants all show cytoplasmic accumulation. WT, wild-type. Scale bars, 20 μm.

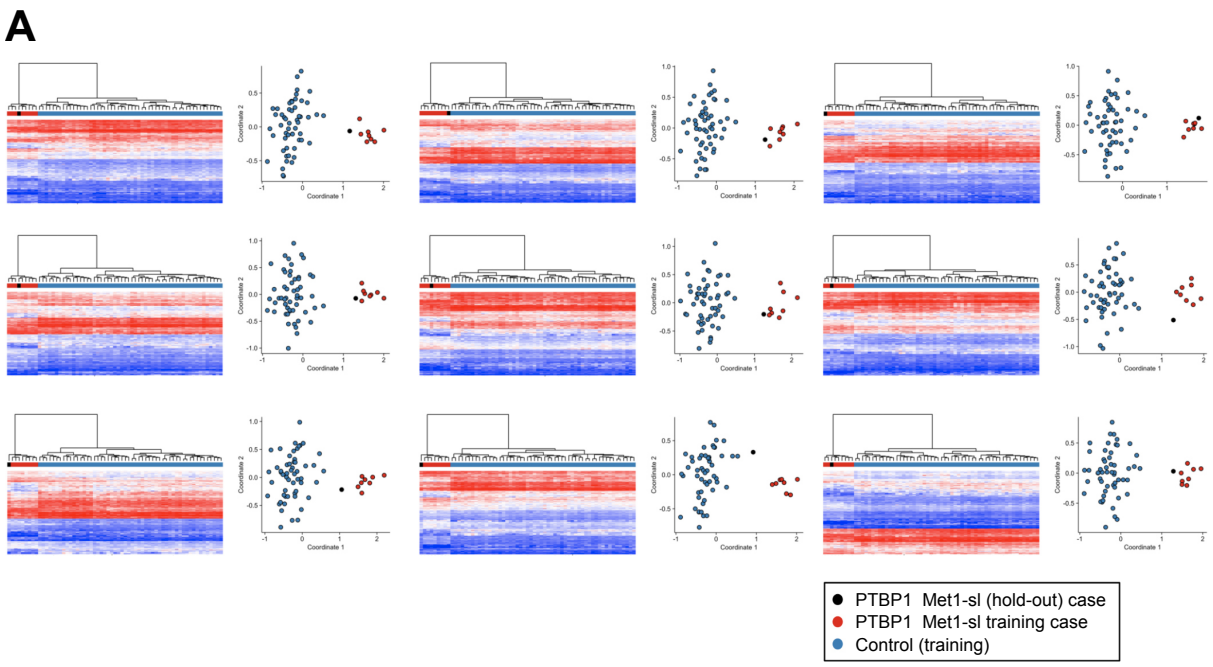


Figure S4. Unsupervised clustering results for leave-one-out cross-validation on PTBP1 discovery cases and episignature specificity. **(A)** In each cross-validation iteration, a single PTBP1 start-loss case was held-out from the discovery data as a test case and a sub-signature was generated using the same feature selection parameters of the identified signature. Reproducibility and robustness of the episignature is validated by consistent clustering of the hold-out case (black) with the remaining discovery cases (red) instead of the matched controls (blue). Each column represents either one PTBP1 individual or control, and each row represents one probe selected for this episignature.
(B) Methylation variant pathogenicity (MVP) scores of EKD test samples and PTBP1 test samples from the SVM classifier were computed and averaged over the fourfold cross-validation (gray). Methylation profiles more similar to the identified signature will have high MVP scores comparable to the SVM training scores (blue) of the discovery cases.

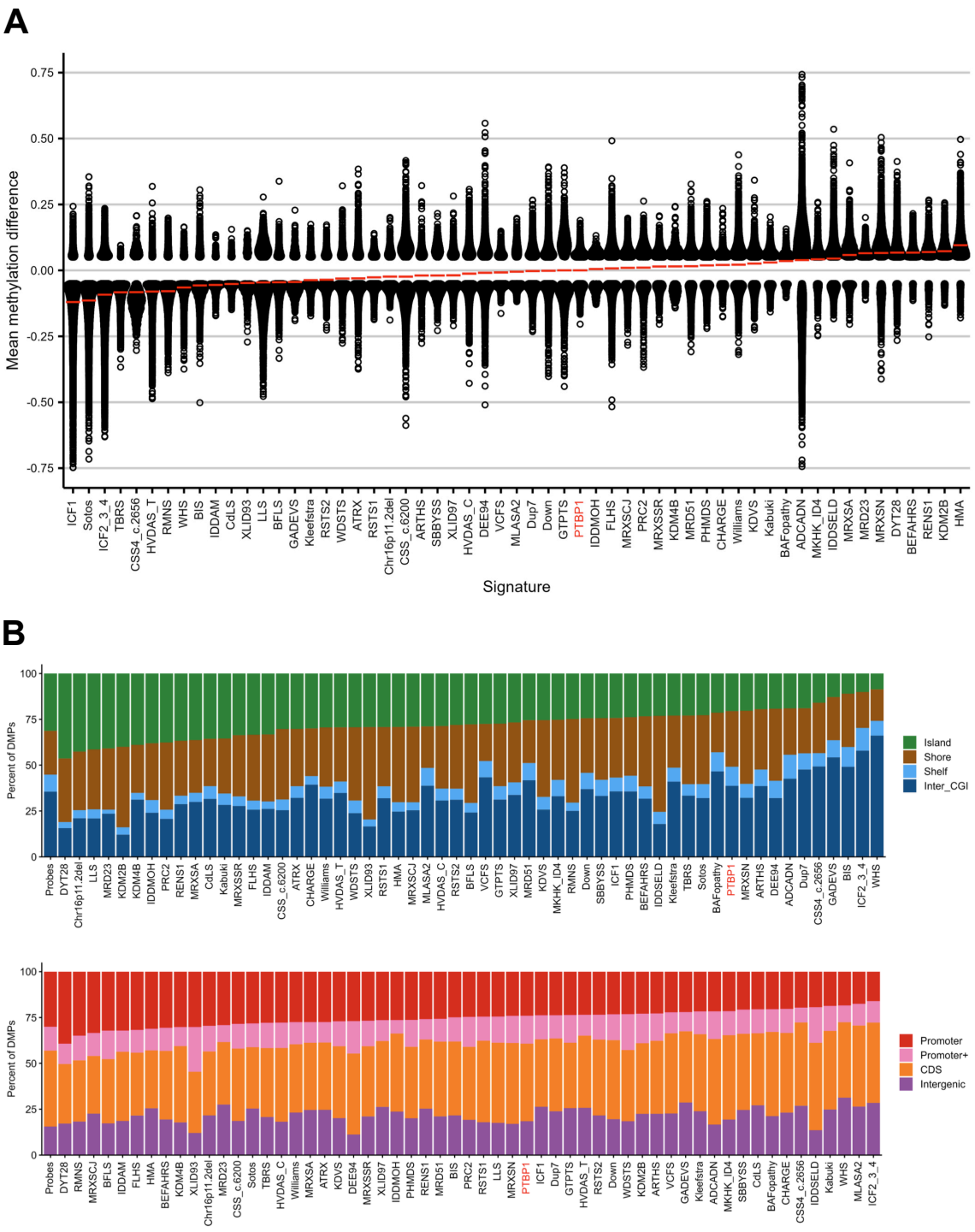


Figure S5. Differentially methylated probes (DMPs) of the PTBP1 cohort versus cohorts in EpiSign™ version 3. (A) Mean differences in the methylation of each signature's DMPs, sorted by the overall mean methylation differences (red line). **(B)** Annotation of the genomic location of the DMPs for all disorder cohorts in relation to CpG islands (top) and genes (bottom). Annotation labels are defined as follows: Island = CpG islands; Shore = within 0–2 kb of a CpG island boundary; Shelf = within 2–4 kb of a CpG island boundary; Inter_CGI = all other regions in the genome. For DMPs of all disorder cohorts in relation to genes, the annotation labels are defined as follows:

Promoter = 0–1 kb upstream of the transcription start site (TSS); Promoter+ = 1–5 kb upstream of the TSS; CDS = coding sequence region; Intergenic = all other regions of the genome. The first column labelled as "Probe" is the annotation distribution of the filtered probe set prior to differential analysis used as a reference background distribution.

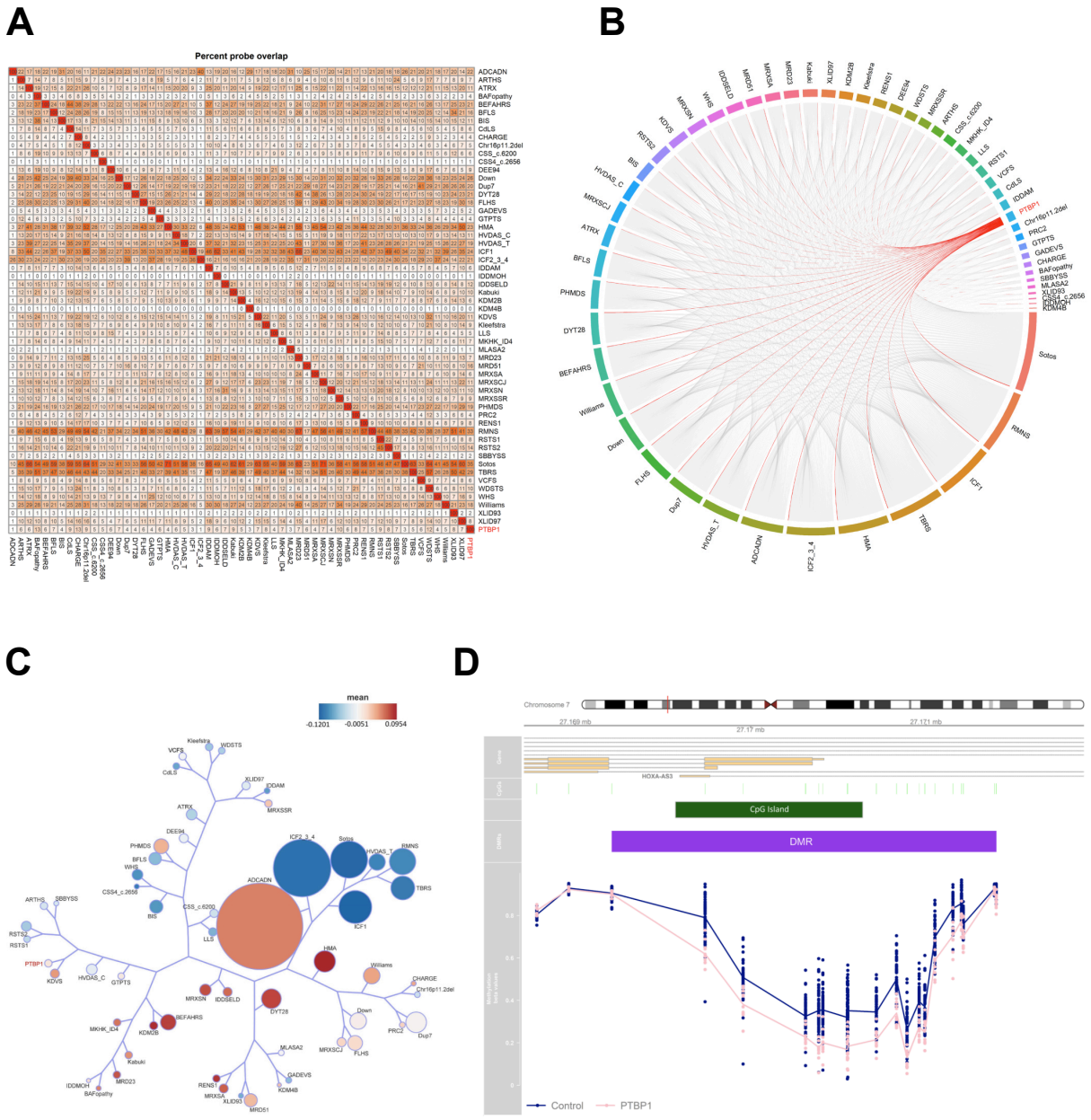


Figure S6. DMP overlaps and similarity between PTBP1 and EpiSign™ version 3 cohorts. (A) Heatmap showing the pairwise percentage of probe intersection between any two cohorts. Below the diagonal, cell colors and values indicate the percentage of DMPs in the cohorts along the columns that are also in the DMP list of cohorts along the rows. **(B)** Circos plot showing DMPs shared among signature cohorts. The thickness of connecting chords is indicative of the number of probes common to the cohort pairs. **(C)** Tree-and-leaf plot derived from Euclidean clustering of signature cohorts using the top 500 DMPs from each group. For each cohort, samples were aggregated and the median value for each probe was calculated. Each signature is represented by a leaf node, where the node size represents the relative DMP set size, and the node color correlates to the overall mean of the respective DMP set. **(D)** Hypomethylated region identified for *PTBP1* located at chr7:27169208-

27171401, spanning a CpG island. The region consists of nineteen CpG sites and overlaps the *HOXA-AS2*, *HOXA-AS3*, *HOXA3*, *HOXA4*, and *RP1-170O19.22* genes.

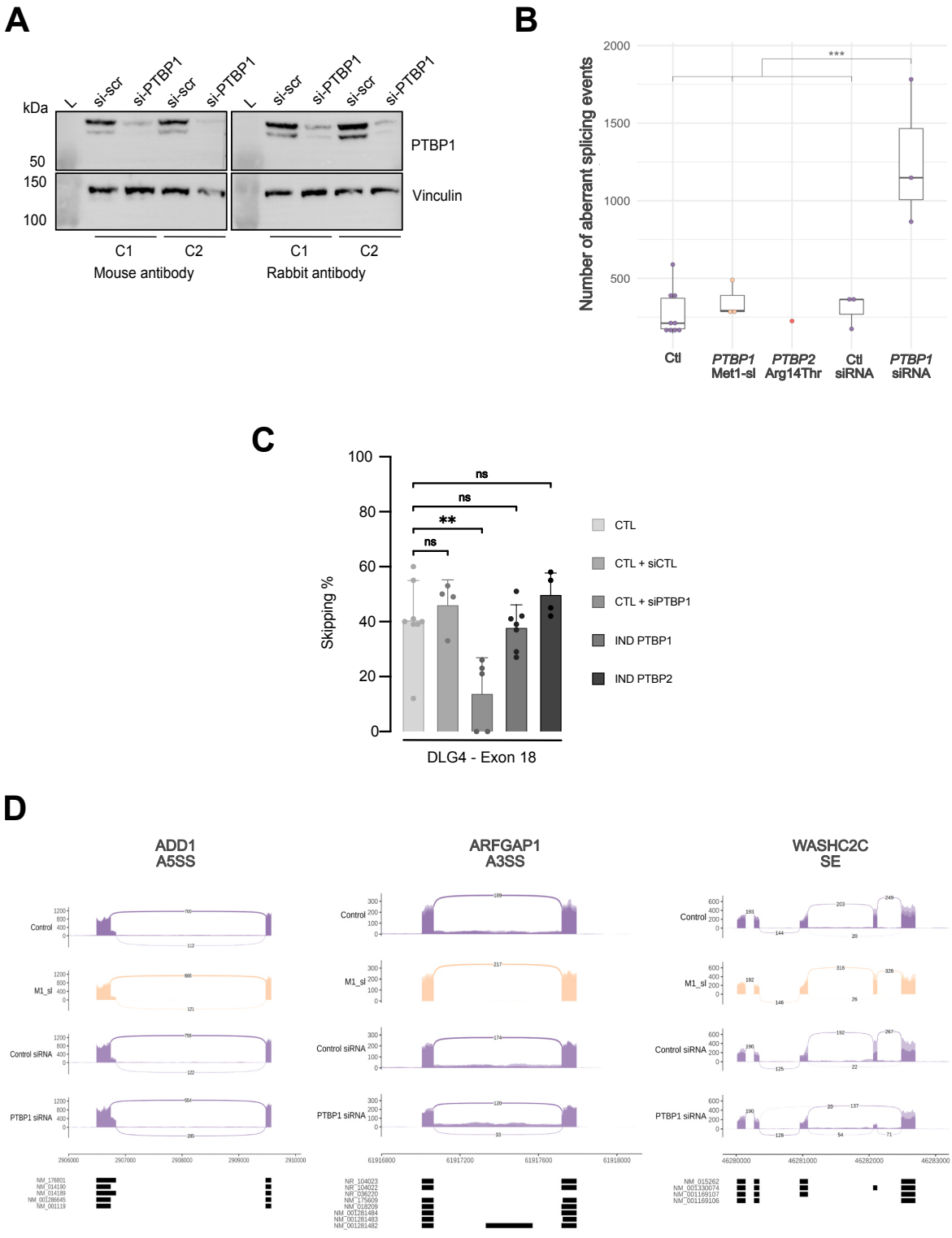


Figure S7. PTBP1 splicing activity in knockdown conditions. (A) Western blot showing PTBP1 expression extinction by more than 80% in two control fibroblast cell lines treated with siRNA targeting PTBP1 transcript (si-PTBP1) compared to the scrambled siRNA (si-scr) treatment and using two different specific antibodies. **(B)** Quantification of the splicing defects detected in primary fibroblasts obtained from

PTBP1_Start-loss individuals and the Arg14Thr-PTBP2 individual compared to control samples untreated or treated with scrambled siRNA or siRNA specific for PTBP1. **(C)** Targeted splicing activity assessment on *DLG4* exon 18, analyzed in fibroblasts derived from healthy (control), treated or untreated with scramble or PTBP1 siRNA, and PTBP1 start-loss or Arg14Thr-PTBP2 individuals. Control untreated or treated cells with scramble siRNA, and fibroblasts from individuals carrying variants in PTBP1 or PTBP2 showed comparable exclusion rates. Conversely, siPTBP1-treated control lines showed aberrant exclusion rates. The statistical analysis was performed using the Mann–Whitney U test. **(D)** Sashimi plots illustrating aberrant splice events detected in RNA-seq data. The abnormal alternative 5' splice site in *ADD1* (left), alternative 3' splice site in *ARFGAP1* (middle), or skipped exon in *WASHC2C* (right) present in each biological replicate of the PTBP1 siRNA group (knockdown of PTBP1 splicing function) and absent or less frequent in all other groups (n=3 per group) are shown as examples. Average numbers of split-reads per group normalized to the lowest expression across all samples are indicated on splice junctions. WT, wild-type.

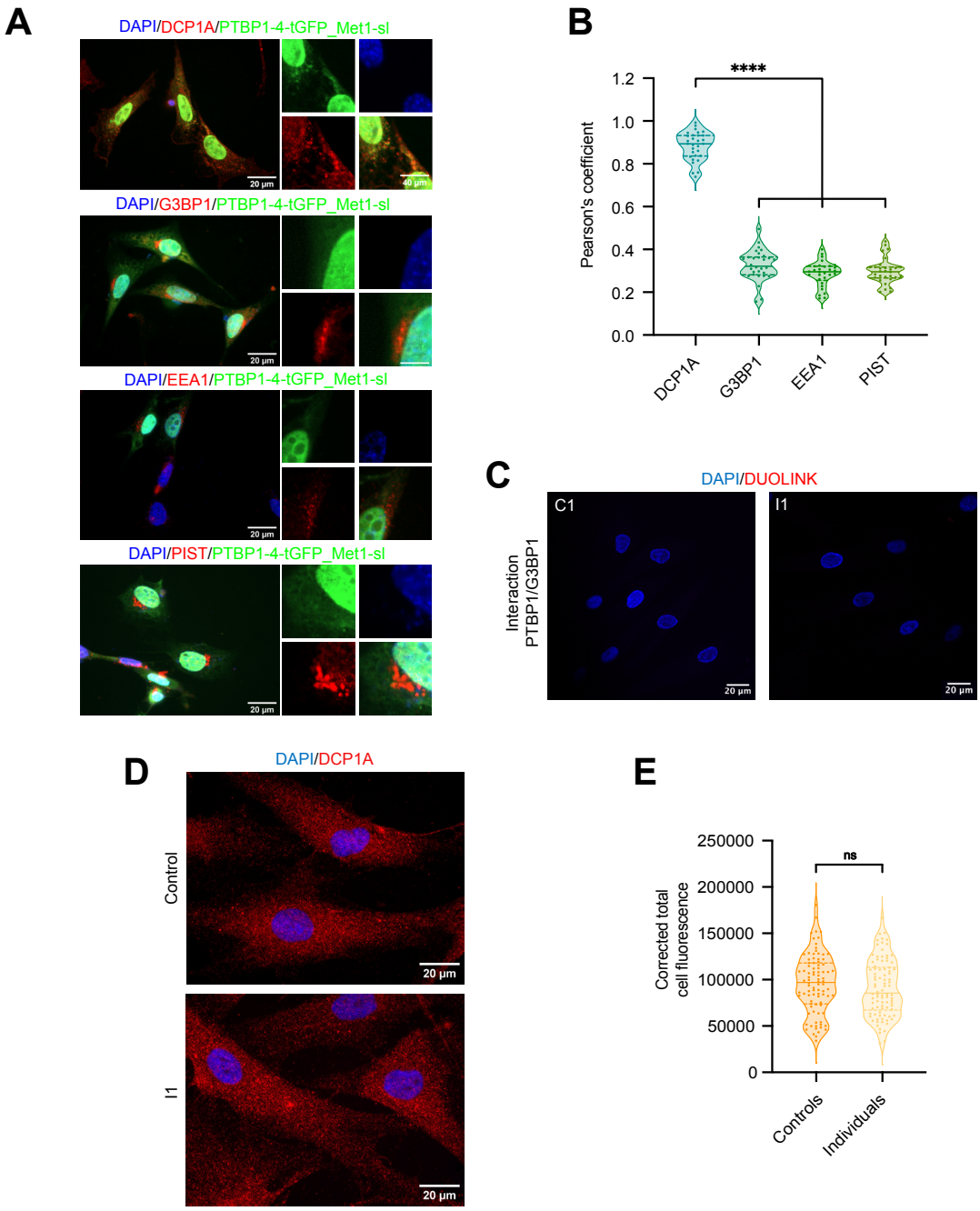


Figure S8. Colocalization assays and DCP1A expression. (A) Fluorescence imaging of transfected NIH-3T3 cells expressing recombinant start-loss PTBP1-4-tGFP plasmids (green), immunostained for DCP1A, G3BP1, EEA1, or PIST (red) and stained with DAPI (blue). Colocalization signals between one of the four markers and PTBP1 are shown on zoomed images at the bottom right of each condition. Only DCP1A colocalizes with PTBP1. **(B)** Measurement of the colocalization between PTBP1_Start-loss and DCP1A, G3BP1, EEA1 or PIST with the Pearson's coefficient obtained via the ImageJ plugin JACoP. Only DCP1A shows a strong colocalization signal with PTBP1_Start-loss. **(C)** Confocal imaging of the proximity ligation assay between PTBP1 and G3BP1 performed on control and PTBP1_Start-loss fibroblasts

(negative control). **(D)** Immunostaining of DCP1A in control and PTBP1_Start-loss fibroblasts. **(E)** Quantification of the DCP1A fluorescence signal measured in D. Analysis performed on ninety cells, from three controls and three affected individual fibroblasts, using Students *t*-test. WT, wild-type. Start-loss, Methionine 1 start-loss variation.

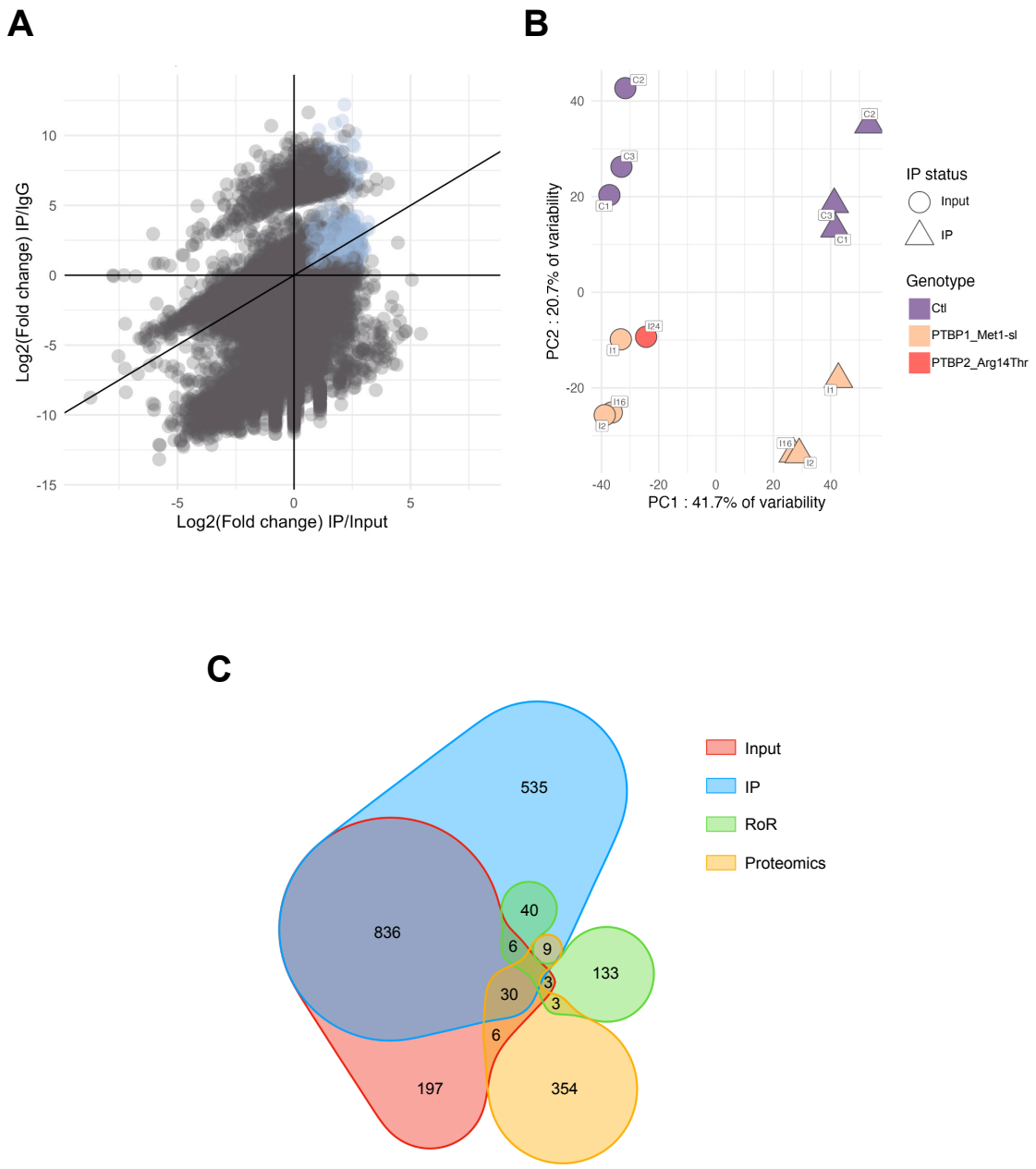


Figure S9. Integrative analysis of RIP-seq, RNA-seq, and proteomics data from fibroblasts derived from affected individuals. (A). Scatter plots showing the $\log_2$ fold change of each gene for the IP over input control samples (x-axis) and for the IP over IgG control samples (y-axis). Blue dots represent PTBP1-associated RNAs with a fold change superior to $\log_2$ (1.5) and an adjusted *p*-value ≤ 0.05 in both comparisons. **(B)** Principal component analysis performed on bulk (input) and PTBP1-immunoprecipitated (PTBP1-IP) mRNA-seq. The first principal component (PC1) explains 41.7% of the entire variance associated with the dataset and distinguishes input samples from IP samples. The second principal component (PC2) supports

20.7% of the variability and separates the control from the start-loss samples. **(C)** Venn diagram showing the overlap between differentially expressed genes from the input, IP, ratio of ratios analysis, and genes corresponding to differentially expressed proteins. Protein isoforms were grouped by gene.

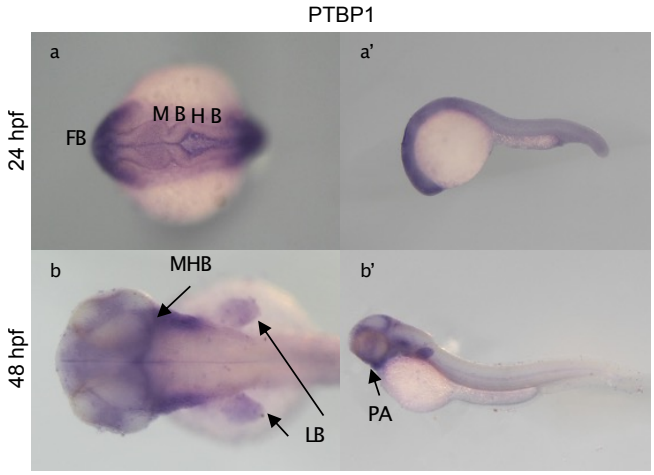


Figure S10. Ptbp1a and ptbp1b mRNA expression in zebrafish. Expression pattern of zebrafish *ptbp1* in zebrafish embryos obtained by whole mount *in situ* hybridization using a *ptbp1* RNA antisense probe. Dorsal (a) and lateral (a') view of WT embryos at 24 hpf. *Ptbp1* expression is evident in the three primary brain vesicles. Dorsal (b) and lateral (b') view of WT embryos at 48 hpf. *ptbp1* expression shows a spatial restriction in the embryonic primordia of the craniofacial skeleton (pharyngeal arches, PA) and pectoral fins (limb bud, LB), and in the midbrain/hindbrain border (MHB). WT, wild-type. dpf, days post-fertilization. FB, forebrain. HB, hindbrain. MB, midbrain.

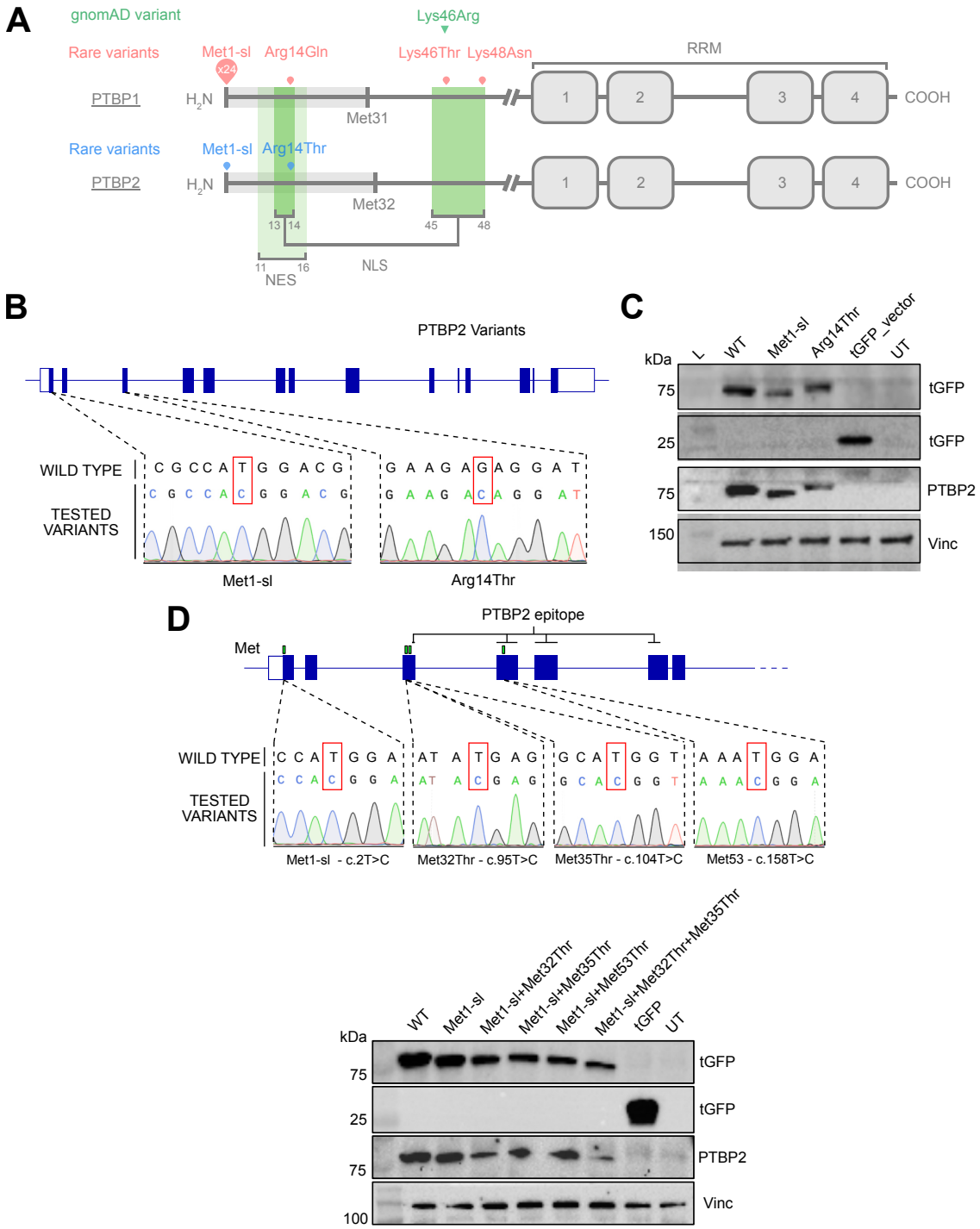


Figure S11. PTBP2 variant generation and characterization of translation initiation. (A) Schematic representation of the PTBP1 (red) or PTBP2 (blue) start-loss and missense variants identified in our cohort and the benign variant (green). Positions of the NES, bipartite NLS, and RRM are depicted. Light gray rectangles represent the part of protein not translated in start-loss individuals. Protein schematics are not to scale. **(B)** Representation of variants generated in the PTBP2 expression plasmid and verified by Sanger sequencing. **(C)** Western blot performed on proteins extracted from NIH-3T3 cells transfected with PTBP2-tGFP fusion constructs. PTBP2-tGFP fusion proteins appear at 80 kDa (anti-PTBP2 and anti-tGFP blots). Note that the lower band

represents start-loss proteins without the first thirty aa. The band at 28 kDa corresponds to the empty tGFP vector. Vinculin is used as a loading control. UT, Untransfected. **(D)** Representation of the different mutated PTBP2 methionines, mutation verification by Sanger sequencing, and protein expression assessment by western blot. Notice the shifted bands according to the specific methionine translation initiation. PTBP2 epitopes recognized by PTBP2 antibody are indicated.

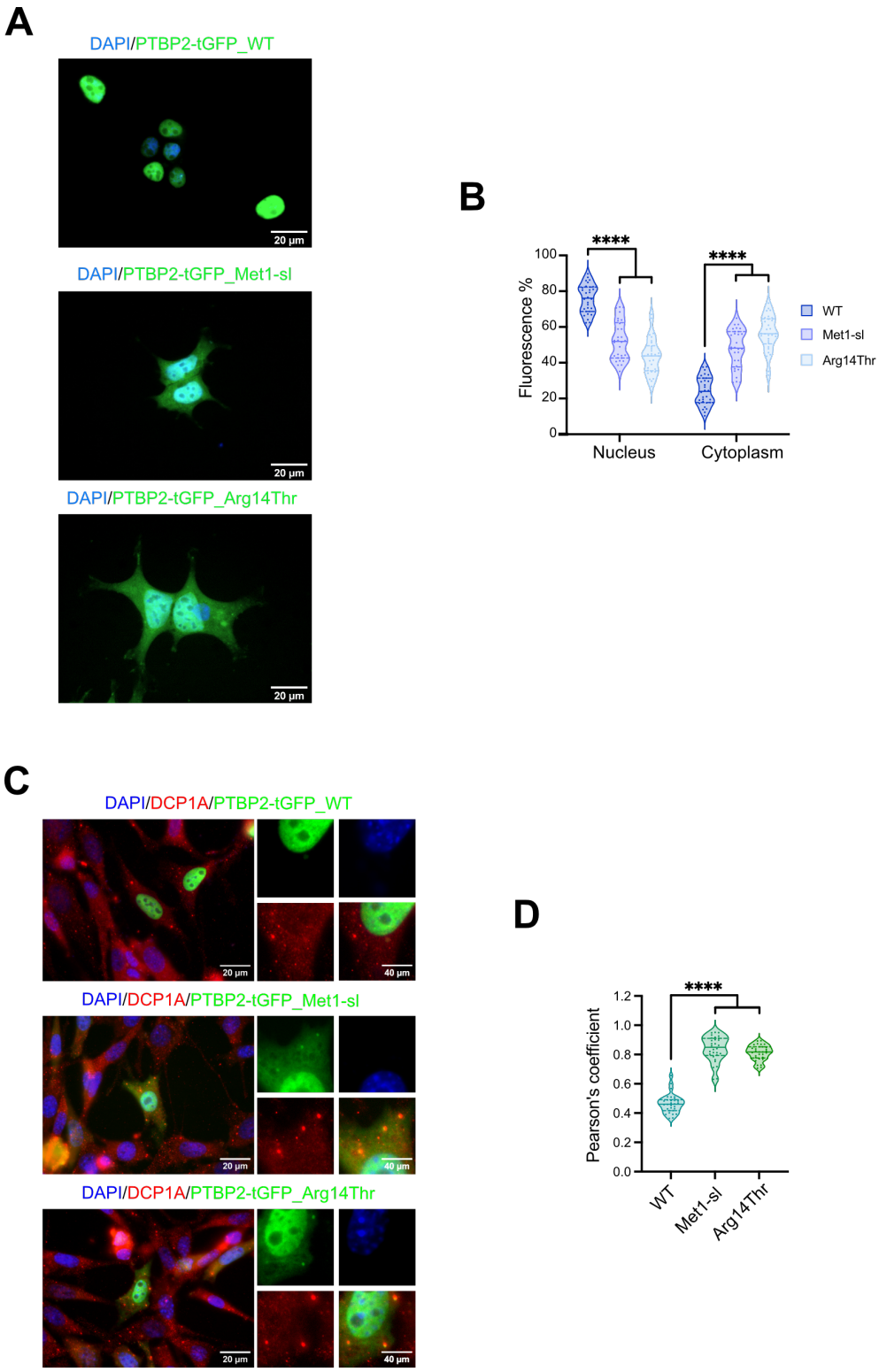


Figure S12. Characterization of PTBP2 nucleocytoplasmic localization. (A) Fluorescence imaging of transfected NIH-3T3 cells expressing recombinant WT, start-loss, or p.Arg14Thr PTBP2-tGFP plasmids (green) and counterstained with DAPI (blue). WT shows nuclear localization. Start-loss and p.Arg14Thr variants show cytoplasmic accumulation. **(B)** Quantification of the fluorescence of the nuclear or

cytoplasmic compartment observed in PTBP2-tGFP transfected cells shown in A. The statistical analysis was performed on thirty cells per condition using the Mann–Whitney U test. **(C)** Imaging of transfected NIH-3T3 cells expressing recombinant WT, start-loss, or missense PTBP2-tGFP plasmids (green), immunostained for DCP1A (red) and counterstained with DAPI (blue). Right: zoomed-in images. **(D)** Pearson's coefficient of PTBP2 WT, start-loss, or missense variants, and DCP1A colocalization.

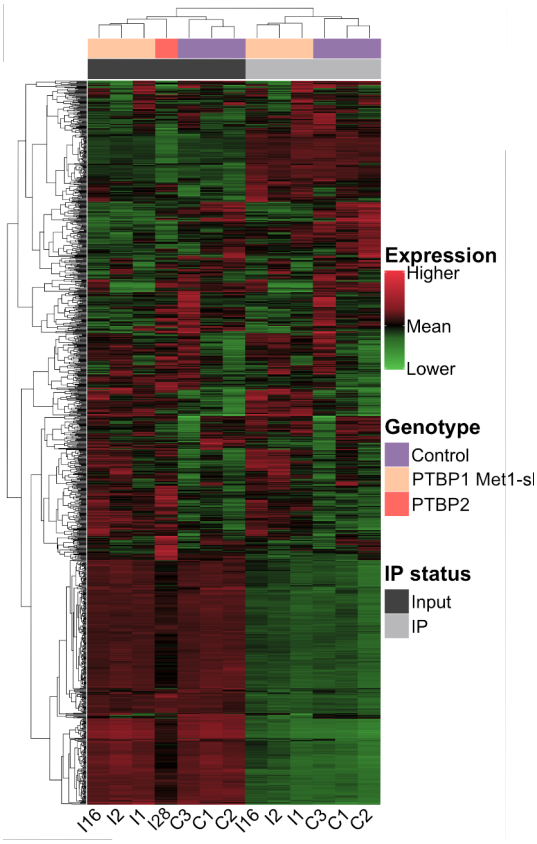


Figure S13. Differential gene expression between control, PTBP1 start-loss and PTBP2 p.Arg14Thr individuals. Hierarchical clustering of gene expression heatmaps showing molecular signatures of fractions of either control, PTBP1, or PTBP2 start-loss samples. The top 5% variable expressed genes (1184) are shown. M1-sl refers to samples harboring PTBP1 start-loss variants. M2 refers to the individual with the p.Arg14Thr variant (I28).

Supplemental References

1. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinforma Oxf Engl. 2009 Jul 15;25(14):1754–60.
2. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010 Sep;20(9):1297–303.
3. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011 May;43(5):491–8.
4. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang 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 Jun;6(2):80–92.
5. Chen S, Francioli LC, Goodrich JK, Collins RL, Kanai M, Wang Q, et al. A genome-wide mutational constraint map quantified from variation in 76,156 human genomes [Internet]. Genetics; 2022 Mar [cited 2023 Jul 11]. Available from: http://biorxiv.org/lookup/doi/10.1101/2022.03.20.485034
6. Fromer M, Moran JL, Chambert K, Banks E, Bergen SE, Ruderfer DM, et al. Discovery and Statistical Genotyping of Copy-Number Variation from Whole-Exome Sequencing Depth. Am J Hum Genet. 2012 Oct;91(4):597–607.

7. Tran Mau-Them F, Duffourd Y, Vitobello A, Bruel A, Denommé-Pichon A, Nambot S, et al. Interest of exome sequencing trio-like strategy based on pooled parental DNA for diagnosis and translational research in rare diseases. Mol Genet Genomic Med. 2021;9:e1836.
8. Neveling K, Feenstra I, Gilissen C, Hoefsloot LH, Kamsteeg EJ, Mensenkamp AR, et al. A Post-Hoc Comparison of the Utility of Sanger Sequencing and Exome Sequencing for the Diagnosis of Heterogeneous Diseases. Hum Mutat. 2013 Dec;34(12):1721–6.
9. Retterer K, Juusola J, Cho MT, Vitazka P, Millan F, Gibellini F, et al. Clinical application of whole-exome sequencing across clinical indications. Genet Med Off J Am Coll Med Genet. 2016 Jul;18(7):696–704.
10. McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinforma Oxf Engl. 2010 Aug 15;26(16):2069–70.
11. Paila U, Chapman BA, Kirchner R, Quinlan AR. GEMINI: Integrative Exploration of Genetic Variation and Genome Annotations. Gardner PP, editor. PLoS Comput Biol. 2013 Jul 18;9(7):e1003153.
12. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nat Biotechnol. 2011 Jan;29(1):24–6.
13. Abyzov A, Urban AE, Snyder M, Gerstein M. CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res. 2011 Jun;21(6):974–84.

14. Eisfeldt J, Vezzi F, Olason P, Nilsson D, Lindstrand A. TIDDIT, an efficient and comprehensive structural variant caller for massive parallel sequencing data. F1000Research. 2017 Jun 30;6:664.
15. Miller CR, Lee K, Pfau RB, Reshmi SC, Corsmeier DJ, Hashimoto S, et al. Disease-associated mosaic variation in clinical exome sequencing: a two-year pediatric tertiary care experience. Cold Spring Harb Mol Case Stud. 2020 Jun;6(3):a005231.
16. Kelly BJ, Fitch JR, Hu Y, Corsmeier DJ, Zhong H, Wetzel AN, et al. Churchill: an ultra-fast, deterministic, highly scalable and balanced parallelization strategy for the discovery of human genetic variation in clinical and population-scale genomics. Genome Biol. 2015 Jan 20;16(1):6.
17. Köhler S, Vasilevsky NA, Engelstad M, Foster E, McMurry J, Aymé S, et al. The Human Phenotype Ontology in 2017. Nucleic Acids Res. 2017 Jan 4;45(D1):D865–76.
18. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinforma Oxf Engl. 2013 Jan 1;29(1):15–21.
19. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016 Jan 4;44(D1):D733-745.
20. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc. 2012 Mar 1;7(3):562–78.
21. Gustavsson EK, Zhang D, Reynolds RH, Garcia-Ruiz S, Ryten M. ggtranscript: an R package for the visualization and interpretation of transcript isoforms using ggplot2. Bioinforma Oxf Engl. 2022 Aug 2;38(15):3844–6.

22. Shen S, Park JW, Lu Z xiang, Lin L, Henry MD, Wu YN, et al. rMATS: Robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. Proc Natl Acad Sci [Internet]. 2014 Dec 23 [cited 2022 Aug 4];111(51). Available from: https://pnas.org/doi/full/10.1073/pnas.1419161111
23. Garrido-Martín D, Palumbo E, Guigó R, Breschi A. ggsashimi: Sashimi plot revised for browser- and annotation-independent splicing visualization. Pertea M, editor. PLOS Comput Biol. 2018 Aug 17;14(8):e1006360.
24. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014 Dec;15(12):550.
25. Pérez-Silva JG, Araujo-Voces M, Quesada V. nVenn: generalized, quasi-proportional Venn and Euler diagrams. Wren J, editor. Bioinformatics. 2018 Jul 1;34(13):2322–4.
26. Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, et al. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. The Innovation. 2021 Aug;2(3):100141.
27. Martens M, Ammar A, Riutta A, Waagmeester A, Slenter DN, Hanspers K, et al. WikiPathways: connecting communities. Nucleic Acids Res. 2021 Jan 8;49(D1):D613–21.
28. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nat Genet. 2000 May;25(1):25–9.
29. Gene Ontology Consortium, Aleksander SA, Balhoff J, Carbon S, Cherry JM, Drabkin HJ, et al. The Gene Ontology knowledgebase in 2023. Genetics. 2023 May 4;224(1):iyad031.
30. Wang JZ, Du Z, Payattakool R, Yu PS, Chen CF. A new method to measure the semantic similarity of GO terms. Bioinformatics. 2007 May 15;23(10):1274–81.
31. Yu G. Gene Ontology Semantic Similarity Analysis Using GOSemSim. In: Kidder BL, editor. Stem Cell Transcriptional Networks [Internet]. New York, NY: Springer US; 2020 [cited

2024 Mar 13]. p. 207–15. (Methods in Molecular Biology; vol. 2117). Available from: http://link.springer.com/10.1007/978-1-0716-0301-7_11

32. Gu Z, Gu L, Eils R, Schlesner M, Brors B. circlize implements and enhances circular visualization in R. Bioinformatics. 2014 Oct;30(19):2811–2.
33. Abaza MSI, Afzal M, Al-Attiyah RJ, Guleri R. Methylferulate from Tamarix aucheriana inhibits growth and enhances chemosensitivity of human colorectal cancer cells: possible mechanism of action. BMC Complement Altern Med. 2016 Dec;16(1):384.
34. Andrejeva G, Gowan S, Lin G, Wong Te Fong ACL, Shamsaei E, Parkes HG, et al. De novo phosphatidylcholine synthesis is required for autophagosome membrane formation and maintenance during autophagy. Autophagy. 2020 Jun;16(6):1044–60.
35. Manshian BB, Martens TF, Kantner K, Braeckmans K, De Smedt SC, Demeester J, et al. The role of intracellular trafficking of CdSe/ZnS QDs on their consequent toxicity profile. J Nanobiotechnology. 2017 Jun 15;15(1):45.
36. Matsuoka R, Miki M, Mizuno S, Ito Y, Yamada C, Suzuki A. MTCL2 promotes asymmetric microtubule organization by crosslinking microtubules on the Golgi membrane. J Cell Sci. 2022 Jun 1;135(11):jcs259374.
37. Zhang Y, Jiang X, Deng Q, Gao Z, Tang X, Fu R, et al. Downregulation of MYO1C mediated by cepharanthine inhibits autophagosome-lysosome fusion through blockade of the F-actin network. J Exp Clin Cancer Res CR. 2019 Nov 7;38(1):457.
38. Bolte S, Cordelières FP. A guided tour into subcellular colocalization analysis in light microscopy. J Microsc. 2006 Dec;224(Pt 3):213–32.
39. Aref-Eshghi E, Kerkhof J, Pedro VP, Barat-Houari M, Ruiz-Pallares N, Andrau JC, et al. Evaluation of DNA Methylation Episignatures for Diagnosis and Phenotype Correlations in

42 Mendelian Neurodevelopmental Disorders. Am J Hum Genet. 2020 Mar;106(3):356–70.

40. Levy MA, McConkey H, Kerkhof J, Barat-Houari M, Bargiacchi S, Biamino E, et al. Novel diagnostic DNA methylation episignatures expand and refine the epigenetic landscapes of Mendelian disorders. Hum Genet Genomics Adv. 2022 Jan;3(1):100075.
41. Zhou W, Triche TJ, Laird PW, Shen H. SeSAMe: reducing artifactual detection of DNA methylation by Infinium BeadChips in genomic deletions. Nucleic Acids Res [Internet]. 2018 Jul 31 [cited 2024 Apr 25]; Available from: https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gky691/5061974
42. Zhou W, Laird PW, Shen H. Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. Nucleic Acids Res. 2016 Oct 24;gkw967.
43. Ho DE, Imai K, King G, Stuart EA. **MatchIt** : Nonparametric Preprocessing for Parametric Causal Inference. J Stat Softw [Internet]. 2011 [cited 2022 Nov 16];42(8). Available from: http://www.jstatsoft.org/v42/i08/
44. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015 Apr 20;43(7):e47–e47.
45. Kuhn M. Building Predictive Models in *R* Using the **caret** Package. J Stat Softw [Internet]. 2008 [cited 2024 Apr 25];28(5). Available from: http://www.jstatsoft.org/v28/i05/
46. Levy MA, Relator R, McConkey H, Pranckeviciene E, Kerkhof J, Barat-Houari M, et al. Functional correlation of genome-wide DNA methylation profiles in genetic neurodevelopmental disorders. Hum Mutat. 2022 Nov;43(11):1609–28.

47. Peters TJ, Buckley MJ, Statham AL, Pidsley R, Samaras K, V Lord R, et al. De novo identification of differentially methylated regions in the human genome. Epigenetics Chromatin. 2015 Dec;8(1):6.
48. Cavalcante RG, Sartor MA. annotatr: genomic regions in context. Valencia A, editor. Bioinformatics. 2017 Aug 1;33(15):2381–3.
49. Phipson B, Maksimovic J, Oshlack A. missMethyl: an R package for analyzing data from Illumina's HumanMethylation450 platform. Bioinformatics. 2016 Jan 15;32(2):286–8.
50. Franceschi S, Corsinovi D, Lessi F, Tantillo E, Aretini P, Menicagli M, et al. Mitochondrial enzyme GLUD2 plays a critical role in glioblastoma progression. EBioMedicine. 2018 Nov;37:56–67.