BMC Medical Genomics

RESEARCH

Open Access

CLN3 transcript complexity revealed by long-read RNA sequencing analysis

Hao-Yu Zhang[1], Christopher Minnis[1], Emil Gustavsson[1], Mina Ryten[1] and Sara E. Mole[1*]

Abstract

Background Batten disease is a group of rare inherited neurodegenerative diseases. Juvenile CLN3 disease is the most prevalent type, and the most common pathogenic variant shared by most patients is the "1-kb" deletion which removes two internal coding exons (7 and 8) in *CLN3*. Previously, we identified two transcripts in patient fibroblasts homozygous for the 1-kb deletion: the 'major' and 'minor' transcripts. To understand the full variety of disease transcripts and their role in disease pathogenesis, it is necessary to first investigate *CLN3* transcription in "healthy" samples without juvenile CLN3 disease.

Methods We leveraged PacBio long-read RNA sequencing datasets from ENCODE to investigate the full range of *CLN3* transcripts across various tissues and cell types in human control samples. Then we sought to validate their existence using data from different sources.

Results We found that a readthrough gene affects the quantification and annotation of *CLN3*. After taking this into account, we detected over 100 novel *CLN3* transcripts, with no dominantly expressed *CLN3* transcript. The most abundant transcript has median usage of 42.9%. Surprisingly, the known disease-associated 'major' transcripts are detected. Together, they have median usage of 1.5% across 22 samples. Furthermore, we identified 48 *CLN3* ORFs, of which 26 are novel. The predominant ORF that encodes the canonical CLN3 protein isoform has median usage of 66.7%, meaning around one-third of *CLN3* transcripts encode protein isoforms with different stretches of amino acids. The same ORFs could be found with alternative UTRs. Moreover, we were able to validate the translational potential of certain transcripts using public mass spectrometry data.

Conclusion Overall, these findings provide valuable insights into the complexity of *CLN3* transcription, highlighting the importance of studying both canonical and non-canonical CLN3 protein isoforms as well as the regulatory role of UTRs to fully comprehend the regulation and function(s) of *CLN3*. This knowledge is essential for investigating the impact of the 1-kb deletion and rare pathogenic variants on *CLN3* transcription and disease pathogenesis.

Keywords Juvenile CLN3 disease, Batten disease, Neuronal ceroid lipofuscinoses, *CLN3*, Transcription, Readthrough gene, Alternative splicing, Long-read RNA sequencing

Background

The neuronal ceroid lipofuscinoses (NCLs, also known as Batten disease) are a group of rare inherited neurodegenerative lysosomal storage diseases characterised by the accumulation of autofluorescent lipofuscin and/or ceroid in lysosomes, with many causative genes. Juvenile CLN3 disease (Juvenile NCL, JNCL), is the most common, accounting for more than 50% of all NCLs cases [1–3]. Classic juvenile CLN3 disease is an autosomal recessive

Mina Ryten is a joint last author.

*Correspondence:
Sara E. Mole
s.mole@ucl.ac.uk
[1] Great Ormond Street Institute of Child Health, University College London, London WC1E 1EH, UK

disorder caused by pathogenic variants in *CLN3* gene [4]. The first symptom of juvenile CLN3 disease is usually visual loss (retinitis pigmentosa), followed by seizures, cognitive and behavioural decline, motor impairment, and premature death in early adulthood [5]. However, differences in clinical manifestations have been reported for patients who are homozygous for the intragenic 1-kb deletion [6] and in those with different pathogenic variants [7]. *CLN3* encodes a protein of 438 amino acids, most of which (amino acids 23–435) share similarities with members of the major facilitator superfamily (MFS) according to InterPro [8] (https://www.ebi.ac.uk/interpro/). The function of CLN3 is not fully understood, however accumulation of glycerophosphodiesters (GPDs) in CLN3-deficient lysosomes suggests it is involved in their clearance [9], and these GPDs inhibit lysosomal phospholipases activity, resulting in the buildup of toxic lysophospholipids [10].

Over 70% of patients with juvenile CLN3 disease are homozygous for a 1-kb intragenic deletion [11], which is a deletion of 966-bp from intron 6 to intron 8 (rs1555468634, g.28485965_28486930del), causing the loss of two exons (coding exons 7 and 8) of the *CLN3* canonical transcript [3, 11]. Previous research has identified at least two transcripts arising from the 1-kb deletion: a 'major' transcript, in which coding exon 6 is spliced to coding exon 9 out of frame so generating a premature stop codon in coding exon 9 with 28 novel amino acids; and a 'minor' transcript, in which coding exon 6 is spliced to coding exon 10 that brings the transcript back into the amino acid reading frame [12]. These transcripts may exert different functions. Modelling the 1-kb deletion 'minor' transcript in fission yeast suggests that it has some residual functionality as well as gaining novel function [13].

To date, there are six *CLN3* transcripts in Refseq [14] and 62 *CLN3* transcripts in Ensembl 110/GENCODE 44 [15, 16]. Existing genome-wide PacBio long-read RNA sequencing data in human and mouse cerebral cortex also suggests diversity in *CLN3* transcripts, with four novel *CLN3* transcripts identified in human data [17]. The high variation in the numbers of different *CLN3* transcripts in current annotations raises the possibility of additional *CLN3* transcripts that have remained unannotated, as well as questions about which transcripts are functional.

To fully decipher the mechanism of juvenile CLN3 disease pathogenesis caused by the 1-kb deletion, it is important to first understand *CLN3* transcription in "healthy" tissues. Given the existing complexity of *CLN3* transcription, several questions arise that need to be addressed: (a) How many *CLN3* transcripts remain unannotated? (b) How do *CLN3* transcripts vary, in terms of their open reading frames (ORFs) and untranslated regions (UTRs); (c) What is the expression of *CLN3* at both the transcript and gene level across tissues? As long-read RNA sequencing enables both the capture of full-length transcripts and their accurate quantification, these questions can now be addressed [18, 19]. Long-read RNA sequencing studies have successfully detected novel transcripts using both untargeted sequencing [17] and targeted sequencing for specific genes, including *GBA1 and GBAP1* [20], *SNCA* [21], and *MYBPC3* [22], as well as single-cell long-read RNA sequencing [23, 24]. In this study, we addressed these questions by utilising the untargeted PacBio long-read RNA sequencing datasets across 23 tissues and 10 cell lines available on the ENCODE portal (https://www.encodeproject.org) to study *CLN3* transcription (Fig. 1).

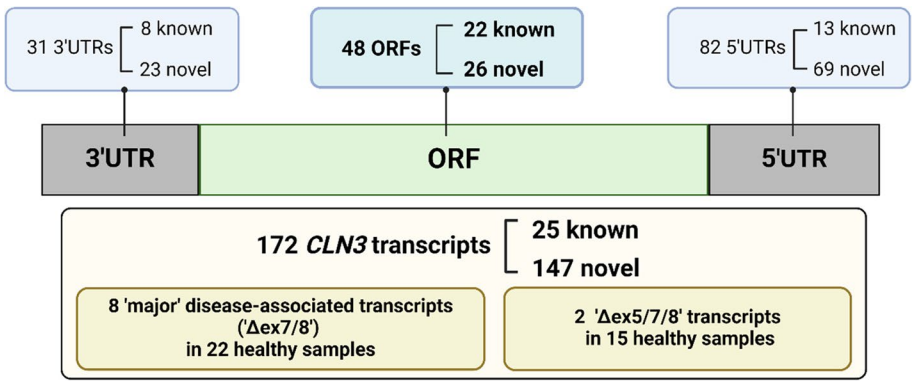

Fig. 1 Summary of key findings. From 99 ENCODE long-read RNA sequencing samples, 172 *CLN3* transcripts are detected, of which 147 of them are novel. We are also able to detect the 'major' disease-associated transcripts ('Δex7/8') across 22 samples and transcripts matching a reported ASO-induced exon skipping ('Δex5/7/8') across 15 samples. Amongst all *CLN3* transcripts detected, 26 novel open reading frames (ORFs), 69 novel 5' untranslated regions (5'UTRs) and 23 novel 3' untranslated regions (3'UTRs) are identified

Methods

Sequence similarity examination

Comprehensive gene annotation based on GENCODE release 29 (GENCODE 29) was downloaded from https://www.gencodegenes.org/human/release_29.html. All unique exons were studied, including 164 *CLN3* exons, and an additional 702,165 other exons. Exon sequences were extracted using the function getseq() of the R package BSgenome. Then BLASTN version 2.9.0 [25] was used with a threshold for filtering exons with a minimum percentage identity of 95% and a minimum bitscore of 100, compared to any given *CLN3* exons.

GTEx V8 short-read RNA sequencing data

Gene-level quantifications in transcripts per million (TPMs), exon-exon junction read counts, and sample attributes of GTEx V8 were downloaded from GTEx portal [26] (https://gtexportal.org/home/datasets) (GTEx_Analysis_2017-06-05_v8_RNASeQCv1.1.9_gene_reads.gct.gz, GTEx_Analysis_2017-06-05_v8_STARv2.5.3a_junctions.gct.gz, GTEx_Analysis_v8_Annotations_SampleAttributesDS.txt).

Five exon-exon junctions which belong to the readthrough transcripts, i.e., splice donors in *CLN3* and splice acceptors in *NPIPB7,* were selected and investigated: 1) chr16:28,466,903–28476250:- as shown in transcript ENST00000635887; 2) chr16:28,466,903–28482104:- as shown in transcript ENST00000637378; 3) chr16:28,466,903–28477463:- as shown in transcripts ENST00000637376, ENST00000636078, ENST00000637745, ENST00000568224 and ENST00000636503; 4) chr16:28,471,175–28,476,250:- as shown in transcripts ENST00000636017, ENST00000636866, ENST00000637299, and ENST00000638036; and 5) chr16:28,466,903–28477011:- as shown in transcript ENST00000636766.

The detection rates of these junctions were determined by proportion of tissue donors in which a specific exon-exon junction was detected. For tissues with 100 or more donors, tissues with the maximum detection rates for specific junctions were checked. For each tissue, the average read counts for specific junctions across all donors were calculated. Then the average read count for each junction was investigated for minimum, mean, and maximum values across all tissue types.

Long-read RNA sequencing data processing

ENCODE untargeted long-read RNA sequencing data

ENCODE untargeted long-read RNA sequencing data was downloaded from the ENCODE portal [27] (https://www.encodeproject.org/., accessed on Aug 26, 2022). We selected 99 samples from nine ENCODE-defined organs, including blood vessel, brain, connective tissue, embryo, endocrine gland, heart, blood, lung, and skin. Data from cancer cell lines was not included. The processed quantification files in.tsv format and annotation files in.gtf format were downloaded for further analysis. All the downloaded quantification and annotation files were generated by ENCODE long-read RNA sequencing pipeline (GitHub—ENCODE-DCC/long-read-rna-pipeline: ENCODE long read RNA-seq pipeline). Corrected transcripts were annotated and quantified by the TALON package [28] (https://github.com/mortazavilab/TALON). GENCODE 29 was used as a reference when generating these datasets.

Transcript separation

For all downloaded ENCODE annotation files, transcripts overlapping with the *CLN3* locus (chr16:28,474,111–28,495,575:-) were selected using GffRead [29]. To separate *CLN3* transcripts and transcripts of the readthrough gene ENSG00000261832, transcripts that also overlapped with the *NPIPB7* locus (chr16:28,456,329–28,472,336:-) were identified. Transcripts that overlapped with both *CLN3* and *NPIPB7* loci were assigned to ENSG00000261832. Transcripts that overlapped with the *CLN3* locus but not with the *NPIPB7* locus were assigned to *CLN3*.

Gene-level quantification

The quantification of transcripts in read counts within each sample was obtained from downloaded quantification files. All transcripts marked as "Genomic", and transcripts of genes marked "Antisense" or "Intergenic" were removed. The read counts for each gene were calculated by summing the read counts of all transcripts identified at each locus. Then for each sample, TPM for each gene was calculated by: 1) calculating the reads per kilobase (RPK) for each gene by dividing the read count by the length of gene in kilobases; 2) calculating the scaling factor by dividing the sum of RPK values within the sample by 1,000,000; 3) calculating the TPM for each gene by dividing the RPK value for each gene by the scaling factor. Then the TPMs of *CLN3* and ENSG00000261832 were checked within each ENCODE-defined organ.

Open reading frame (ORF) prediction

The ORFs for *CLN3* transcripts were predicted using the findMapORFs() function from the R package ORFik [30] (https://github.com/Roleren/ORFik). Exon coordinates were obtained from annotation files downloaded from ENCODE portal. GRCh38 was used as the reference genome. "ATG" and "TAA|TAG|TGA" were used as the start codon and stop codons, respectively. When predicting the ORFs, only the longest ORF per stop codon was kept. Then, the longest ORF per transcript was selected and the length of the product was calculated. ORF

identifiers (ID) were assigned to each ORF, for example, "CLN3_24_438aa", based on the length rank (e.g., '24') and length of the product (e.g., 438 amino acids).

Transcript merging

Transcripts were called separately in each sample by the ENCODE pipeline so to study all transcripts and their specific features using a common framework and naming convention, transcripts were merged based on the following criteria: (a) the same ORF; (b) the same proximal 5' and 3'UTR internal boundaries; (c) distal 5' and 3'UTR ends located within 20 bps.

UTRs were extracted using the fiveUTRsByTranscript() and threeUTRsByTranscript() functions from R package GenomicFeatures [31]. The same ID was assigned for a given UTR with the same internal boundaries and with 5' and 3' ends within 20 bps (±). IDs for UTRs (e.g., "5UTR_136") consisted of type (e.g., '5' or '3') and arbitrary numbers (e.g., '136').

For the merged transcripts, IDs were assigned using a combination of ORF IDs, 5'UTR IDs, and 3'UTR IDs, for example, "CLN3_125_316aa_5UTR_132_3UTR_79". This allows an immediate appreciation of the structure which would not be possible using a completely arbitrarily assigned ID. Transcripts detected in three or more samples were considered valid.

Transcript-level quantification

For merged transcripts, transcript-level expression was assessed on the basis of transcript occurrence and usage. Transcript occurrence refers to the number of samples in which a specific transcript was detected. Transcript usage refers to all transcription that was assigned to a specific *CLN3* transcript divided by the total transcription from the locus (in read counts). ORF usage was defined as proportion of transcripts containing specific ORFs and was calculated by summing up the usage of transcripts containing the same ORFs in each sample. Tissue-specific transcripts were defined as transcripts that had usage in a given tissue that was at least two times higher than in any other tissue.

Nonsense-mediated decay prediction

We predicted whether a transcript was subject to nonsense-mediated decay (NMD) using the function predictNMD() from R package factR [32] (https://github.com/fursham-h/factR). This function is based on the commonly used rule that if the stop codon of a transcript is more than 50 nucleotides upstream of the most downstream exon-exon junction, the transcript will be NMD-sensitive.

Determining the novelty of ORFs, UTRs and transcripts

Known *CLN3* transcripts were extracted from GENCODE version 29. Then, ORFs were predicted by R package ORFik [30] using the same arguments as above. UTRs from GENCODE 29 and ENCODE datasets were extracted using fiveUTRsByTranscript() and threeUTRsByTranscript() functions from R package GenomicFeatures. ORFs/UTRs detected in GENCODE 29 but not in the selected ENCODE datasets were considered novel. Merged transcripts were classified as novel if all the ENCODE transcripts they were derived from were novel.

Transcripts classification

Transcripts were categorised into different types based on NMD prediction, coding potential and novelty. Transcripts that were predicted to undergo NMD were categorised based on their novelty into NMD_Known and NMD_Novel. Non-NMD transcripts with ORFs encoding products smaller than 150aa were considered non-coding. These non-coding transcripts were categorised into Non_coding_Known and Non_coding_Novel based on their novelty; and then coding transcripts were categorised into Coding_Known and novel coding transcripts. Novel coding transcripts were further categorised based on the novelty of the UTRs and ORFs: transcripts with novel 3' or 5' UTRs but known ORFs (Novel_3'/5'UTR_only), coding transcripts with both known ORFs and known UTRs but the combination of UTRs and ORFs were novel (Novel_combination), coding transcripts with novel ORFs and at least one novel UTR (Novel_ORF_and_UTR), and coding transcripts with only novel ORFs (Novel_ORF_only).

Transcripts visualisation

CLN3 transcripts were visualised using R Package ggtranscript [33] (https://github.com/dzhang32/ggtranscript). The coding sequences (CDSs) coordinates used were generated by the ORF prediction step using the R package ORFik.

Validation of transcripts

Transcripts were validated in terms of their 5' transcription start sites (TSSs), splice junctions, and 3' polyadenylation sites (PASs). The reference TSS (refTSS) data v3.3 in BED format was downloaded from http://reftss.clst.riken.jp/ (accessed on Nov 09, 2022). This included 5' sequencing data from FANTOM5, EPDnew, ENCODE RAMPAGE, ENCODE CAGE, stem cell CAGE (DDBJ accession number DRA000914), and dbTSS. The integrated data was reprocessed and mapped to the latest version of the reference genome [34]. RJunBase (www.RJunBase.org,

accessed on Dec 13, 2022) is a web database of splicing junctions from 18,084 normal samples and 11,540 cancer samples [35]. A total number of 64 non-tumour-specific linear splice junctions of *CLN3* were identified and downloaded. SpliceVault 300K-RNA database was downloaded using https://storage.googleapis.com/misspl-db-data/misspl_events_300k_hg38.sql.gz (accessed on Jul 31, 2024). The 300K-RNA database contains splice junctions 335,663 public RNA-seq datasets from GTEx and SRA [36]. PolyAdenylation annotation for human GRCh38.96, Homo sapiens v2.0, was downloaded from PolyAsite 2.0 (https://polyasite.unibas.ch/atlas#2, accessed on Dec 5, 2022). The polyA human atlas included 221 different 3'end sequencing libraries prepared by different protocols, including 3'-Seq (Mayr), 3'READS, DRS, QuantSeq_REV, SAPAS, PAPERCLIP, PolyA-seq, PAS-seq, A-seq, and 3P-Seq [37].

Public mass spectrometry data

Public mass spectrometry datasets PXD026370 and PXD028605 were downloaded from ProteomeXchange (https://www.proteomexchange.org/). PXD026370 contains data derived from post-mortem human brain tissue from patients with multiple system atrophy (MSA) ($N=45$) and controls ($N=30$) [38]. PXD028605 contains data from non-small-cell lung cancer (NSCLC) patients ($N=5$) and healthy individuals ($N=5$) whole blood cell pellets, and whole blood collected from healthy volunteers [39].

Transcripts predicted to have ORFs with unique peptide sequences were selected. In each case, the public mass spectrometry datasets were searched using MetaMorpheus [40] 1.0.1 for evidence of the unique peptide sequence with default settings to determine their presence. Alignments of identified peptides and corresponding CLN3 protein sequences were visualised using R package ggmsa [41].

Protein structures

Translated amino acid sequences of transcripts of interest were used as input for AlphaFold 2.3.2 Colab to enable prediction of the protein structure [42] (https://colab.research.google.com/github/deepmind/alphafold/blob/main/notebooks/AlphaFold.ipynb). Pairwise Structure Alignment was performed using jFATCAT (rigid) on RCSB Protein Data Bank [43] (https://www.rcsb.org/).

Results

CLN3 transcription shows complexity in current annotation

Of the 62 *CLN3* transcripts in Ensembl 110, 29 of them are classified as protein-coding, and 10 of them have the highest transcript support level with all splice junctions supported by mRNAs from other sources (e.g., SPTrEMBL and RefSeq). Currently, only 172 of the 23,217 protein-coding genes in Ensembl 110 have more than 50 transcripts, and *CLN3* is ranked 81st among all protein-coding genes in terms of number of transcripts (Figure S1).

CLN3 overlaps with readthrough gene ENSG00000261832

To assess whether multimapping could be limiting the annotation of *CLN3* transcripts we first determined if there was any other gene with high sequence similarity to the locus. This analysis demonstrated that 42 different *CLN3* exons are identical in sequence to exons of ENSG00000261832, a readthrough gene containing exons from both *CLN3* and *NPIPB7* (Fig. 2A). All other *CLN3* exons were identically assigned to the *CLN3* locus alone. In GTEx V8, ENSG00000261832 was not included in the gene-level quantification data. Five unique exon-exon junctions spanning *CLN3* and *NPIPB7* were found, with average read counts ranging from 1.12 to 2.12 (Figure S2, Table S1).

CLN3 annotation is complicated by the *CLN3-NPIPB7* readthrough gene

To examine the accuracy of *CLN3* annotation and quantification, we analysed ENCODE long-read RNA-sequencing data. Transcripts that overlap with the *CLN3* locus (chr16:28,474,111–28,495,575:-) were selected within 99 chosen samples, which includes data generated from 23 different human tissue types and 10 different human cell lines (Table S2). The selected transcripts were annotated to either *CLN3* or the *CLN3-NPIPB7* readthrough gene, ENSG00000261832 (Fig. 2A and B). To avoid misannotation, we categorised selected transcripts that overlap with the *CLN3* locus based on whether they also overlap with the *NPIPB7 locus*. Those transcripts that overlap with both *CLN3* and *NPIPB7* loci were assigned to ENSG00000261832. We found that ~ 10–25% of all *CLN3* transcripts had originally been assigned to the readthrough gene, ENSG00000261832 by the ENCODE Long Read RNA-Seq Analysis Protocol for Human Samples (Fig. 2C).

CLN3 is expressed in all 99 samples. We observed that gene-level expression of *CLN3* is variable across different organs (Table 1). Among these selected organs, *CLN3* has the highest expression in blood (median TPM 63.10), and the lowest expression in heart (median TPM 11.59). Compared with GTEx V8 bulk gene expression for *CLN3,* our data shows similar variability across different organs. Both datasets show relatively low expressions in the brain and heart, and higher expressions in blood and lung (Table 1). In contrast, the readthrough gene, ENSG00000261832, is not detected in all samples of a given tissue, for example, it is only detected in 35.3% (6 out of 17) of heart samples. It exhibits low expression

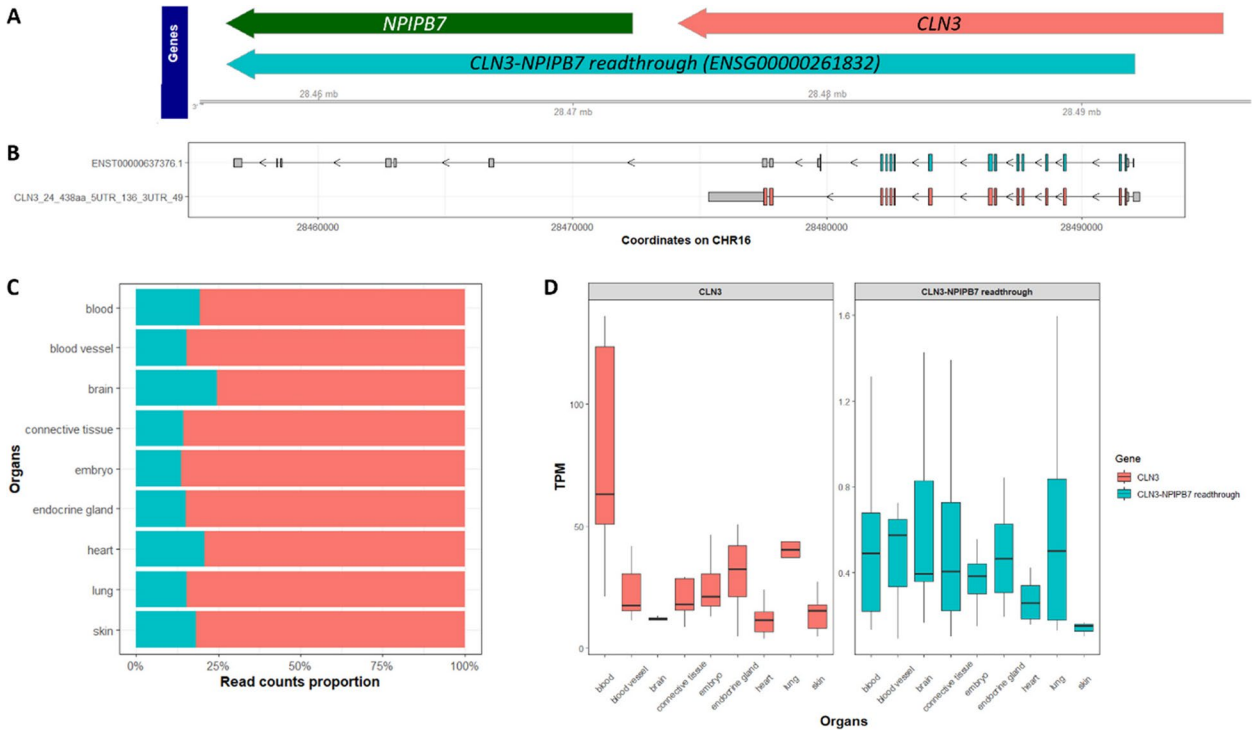

Fig. 2 Assigning *CLN3* transcripts to the *CLN3-NPIPB7* readthrough gene affects the accurate quantification of *CLN3* in ENCODE long-read RNA sequencing data. **A** The *CLN3, NPIPB7,* and *CLN3-NPIPB7* readthrough gene (ENSG00000261832) loci are shown according to their genomic coordinates from Ensembl 110. **B** The readthrough gene, ENSG00000261832, contains exons of *CLN3* and *NPIPB7*. This plot shows representative transcripts of *CLN3* and ENSG00000261832. Coding sequences are coloured by red for *CLN3* and cyan for ENSG00000261832, grey boxes show the UTRs. Note, *CLN3* transcript is present on the antisense strand, so it reads right to left. **C** Over 10% of *CLN3* transcripts were assigned to the readthrough gene in selected ENCODE data, affecting the accurate annotation and gene-level quantification of *CLN3*. **D** The gene-level quantification of *CLN3* and the readthrough gene ENSG00000261832 across nine ENCODE-defined organs are plotted. The readthrough gene has rather lower transcripts per million (TPMs) across all samples compared with *CLN3*

Table 1 Gene-level quantification of *CLN3* across nine ENCODE-defined organs, expressed as transcripts per million (TPMs)

Organs	Number of Samples	Tissues	Min TPM	Median TPM	Max TPM	GTEx V8 median TPMs within specific tissues
Blood	16	HL-60	21.20	63.10	136.10	19.07
Blood vessel	6	posterior vena cava, aorta, endothelial cell of umbilical vein	11.44	17.56	41.87	11.52 – 13.36
Brain	9	middle frontal area 46	5.54	11.97	13.25	3.31—10.47
Connective tissue	16	IMR-90, osteocyte, WTC11, chondrocyte, mesenteric fat pad, HFFc6	8.61	18.07	62.77	NA
Embryo	18	H9, neural crest cell, endodermal cell, H1	12.93	21.09	46.38	NA
Endocrine gland	9	adrenal gland, type B pancreatic cell, right lobe of liver, progenitor cell of endocrine pancreas	4.93	32.51	50.77	6.10 – 16.76
Heart	17	heart left ventricle, heart right ventricle, left cardiac atrium, right cardiac atrium, cardiac septum, Right ventricle myocardium inferior, left ventricle myocardium superior, left ventricle myocardium inferior, Right ventricle myocardium superior	4.02	11.59	24.08	3.20 – 4.55
Lung	6	upper lobe of right lung, lower lobe of left lung, lower lobe of right lung, left lung	23.25	40.25	107.08	22.18
Skin	8	WTC11, GM23338	4.79	15.50	27.08	16.61—17.85

levels across all nine organs, with median TPMs ranging from 0.15 (skin) to 0.57 (blood vessel) (Fig. 2D, Table S3).

172 different *CLN3* transcripts are detected with no dominantly expressed transcript

We identified 172 *CLN3* transcripts across 99 samples. They were categorised based on their novelty, likelihood of undergoing nonsense-mediated decay (NMD) and coding potential (Fig. 3A). Among these 172 transcripts, 147 are absent from GENCODE 29, including 75 transcripts with coding potential. Within these 75 novel coding transcripts, 13 have novel ORFs and known UTRs; 34 have novel ORFs and novel 5' or 3'UTRs; 25 have only novel 5'UTRs; and 3 have known ORFs and UTRs but novel combinations (Fig. 3B).

The expression of transcripts was analysed, both in terms of the number of samples in which a transcript is detected and the usage of that transcript across tissues. Interestingly, unlike previous research showing that dominant transcripts of protein-coding genes account for around 80% of transcription of the locus [44], there is no dominant transcript detected for *CLN3*. The most abundant transcript is the canonical *CLN3* transcript "CLN3_24_438aa_5UTR_136_3UTR_49" as in GENCODE 29, which is detected in all 99 samples and has median usage of 42.9%. The second most abundant transcript "CLN3_24_438aa_5UTR_65_3UTR_52" has the same ORF of 438aa but different UTRs and is detected in almost all samples ($N=98$), with median usage of 18.7%. The third most abundant transcript, "CLN3_57_384aa_5UTR_17_3UTR_52", has a smaller ORF due to a non-canonical start codon; this is detected in 63 samples with median usage of 3.7% (Fig. 3C, Table S4).

Transcripts can show tissue-specific usage [45]. In this study, tissue-specific transcripts are defined as those exhibiting twofold higher expression in one tissue as compared to that in any other tissue. Amongst the 172 *CLN3* transcripts, 75 tissue-specific transcripts are detected (Table S8, Figure S3). Across nine organs, blood contains the largest number of tissue-specific transcripts ($N=21$), followed by heart ($N=15$) and embryo ($N=10$) (Table S8).

"Disease-associated" transcripts are detected in control samples

We identified transcripts matching the splicing patterns of those previously reported in patient-derived fibroblasts homozygous for the 1-kb deletion namely the 'major' and 'minor' disease-associated transcripts [7]. In this dataset, we identified eight transcripts showing the same ORF pattern as the 'major' disease-associated transcript, that is coding exon 6 spliced to exon 9 (Δex7/8) introducing a non-canonical coding sequence and premature stop codon; these transcripts have the same ORF but different 5'/3' UTRs (Fig. 3D). Together, they have median usage of 1.5% and are detected in 22 samples from 13 tissue donors (Table S5). We also identified two transcripts showing coding exon 6 spliced to coding exon 10 and the canonical stop codon as the 'minor' disease-associated transcript (Δex7/8/9), but these transcripts did not pass the filtration of detection in three or more samples. These two 'minor' transcripts are detected once in male newborn endothelial of umbilical vein primary cell and female embryo (5 days) chondrocyte in vitro differentiated cells, respectively. Furthermore, two transcripts (Δex5/7/8) matching a recent antisense oligonucleotides (ASO)-induced exon skipping [46], which skips canonical coding exon 5 and reads through an alternate frame in exon 6 and restores the ORF from the 'major' disease-associated transcripts (Δex7/8) from exon 9, are detected across 15 samples, with median usage of 1.1% (Figure S3, Table S6).

48 different ORFs are detected across all *CLN3* transcripts

Different transcripts can contain the same open reading frames (ORFs) and produce the same protein products. Therefore, to infer the proportion of different CLN3 protein isoforms generated, we investigated the usage of ORFs. ORF usage was calculated by summing all transcripts containing the same ORFs in each individual sample. From the 99 selected samples, we detected 48 different ORFs, of which 26 are novel. The most abundant ORF is that encoding the canonical CLN3 protein isoform of 438aa (UniProt ID: Q13286-1), "CLN3_24_438aa". Summing all transcripts containing this ORF gave median usage of 66.7% (Fig. 4A, Table S7), suggesting that transcripts encoding non-canonical protein isoforms account for around one-third of *CLN3* transcription. The most abundant non-canonical ORF is "CLN3_57_384aa" which encodes a shorter protein of 384aa (UniProt ID: B4DFF3) generated through a different start codon position that misses the first two coding exons present in the canonical transcript. Collectively, transcripts containing the "CLN3_57_384aa" ORF have median usage of 7.4% (Fig. 4A, Table S7) and are detected in 87 samples. The most abundant novel ORF is "CLN3_207_223aa" which has a start codon in the coding exon 8 of the canonical transcript and the canonical stop codon. It has median usage of 3.1% (Fig. 4A, Table S7) and is detected in 72 samples. Additionally, the ORF identified in the 'major' disease-associated transcript, "CLN3_235_181aa" (UniProt ID: Q9UBD8), has median usage of 1.5% across 22 samples (Table S7). Moreover, the same ORFs could have different usage across different organs. For example, the ORF "CLN3_125_316aa" which is generated through a retained intron event and a different stop codon (Fig. 6A)

is highly expressed in brain while "CLN3_207_223aa" is highly expressed in blood (Fig. 4B). These particular transcripts encode protein isoforms with different stretches of amino acid sequences and so may have different functional properties.

CLN3 transcripts with the same ORFs have alternative UTRs

Among the 172 *CLN3* transcripts, 82 different 5'UTRs were identified, 69 of which are novel. Twenty-nine of the 48 ORFs were found with multiple 5'UTRs. The canonical ORF "CLN3_24_438aa" has the most 5'UTRs ($N=23$) (Fig. 5A). Out of the 23 5'UTRs found with "CLN3_24_438aa," 16 of them show distal transcription start sites (TSSs) and additional exons at the 5' ends compared to the 5'UTR of the canonical transcript, i.e., "5UTR_161". By examining the usage of transcripts with specific ORFs and specific 5' or 3' UTRs, we found that for ORF "CLN3_24_438aa", there are 11 5'UTRs showing tissue-specific patterns, with blood harbouring the most tissue-specific 5'UTRs ($N=5$). Out of these tissue-specific 5'UTRs, "5UTR_194" has the highest usage of 8.3% in blood vessel.

There are fewer 3'UTRs than 5'UTRs. Thirty-one different 3'UTRs were identified, of which 23 are novel. Only 12 ORFs were found with multiple 3'UTRs. There are three 3'UTRs found with ORF "CLN3_24_438aa" (Fig. 5B). The distal 3'UTR, "3UTR_49", is the only distal 3'UTR identified within 172 *CLN3* transcripts. It is specifically detected with the canonical ORF.

The variety of *CLN3* transcripts can be validated

Given the large variety of *CLN3* transcripts identified, we sought to validate these transcripts and their translation potential using data from different independent technologies. The transcription start sites (TSSs) of *CLN3* transcripts, which represent 5' ends of these transcripts, were validated using the reference TSS dataset from refTSS [24]. Among the 172 *CLN3* transcripts, TSSs of 65.7% of them ($N=113$) are located within 50bps of the TSS peaks, providing confidence in the existence of these transcripts. There are TSSs signals upstream of the canonical *CLN3* transcript which indicate that *CLN3* might be controlled by upstream promoters (Fig. 3C).

We found that 41 out of 58 splice junctions found in *CLN3* transcripts are detected in RJunBase (www.RJunBase.org). These same 41 junctions can also be detected in SpliceVault 300K-RNA data, where 15 out of 17 absent splice junctions are found. Overall, 19 splice junctions are associated with cryptic splice sites within SpliceVault 300K-RNA data (Table S9).

The polyadenylation (polyA) sites, which represent the 3' ends of transcripts, of 91.3% of *CLN3* transcripts ($N=157$) match with the polyA sites from polyAsite 2.0 [26]. PolyA signals of the distal 3'UTR, "3UTR_49", and shorter 3'UTRs, for example, "3UTR_52" are detected in polyAsite 2.0 data (Fig. 3C). Strong polyA signals are also found in the region of coding exons 7 and 8, and exon 3 of the canonical *CLN3* transcript (Fig. 3C). These polyA signals could indicate premature termination of transcription. Transcripts with premature transcription termination sites near the canonical coding exons 8 and 4 are detected at low levels in a small proportion of our selected samples (Figure S3).

The translation of *CLN3* transcripts can be validated

Finally, we investigated the translational potential of transcripts predicted to have an ORF (including NMD transcripts) using public mass spectrometry data from human post-mortem brain and whole blood. Unique sequence stretches from seven specific ORFs were tested: CLN3_4_489aa (retained intron), CLN3_80_363aa (retained intron), CLN3_125_316aa (retained intron), CLN3_46_414aa (exon skipping), CLN3_101_338aa (alternative translation start site, exon skipping), CLN3_235_181aa ('major' disease-associated transcript, exon skipping), CLN3_115_328aa ('minor' disease-associated transcript, exon skipping) (Fig. 6A). Peptides "ADSAPGGHARSGRAPESR" for

(See figure on next page.)

Fig. 3 Summary of identified *CLN3* transcripts. **A** Data processing pipeline for ENCODE long-read RNA sequencing data is shown. **B** Valid transcripts are categorised based on transcript novelty, open reading frame (ORF) and untranslated region (UTR) novelty, coding potential and nonsense-mediated decay (NMD) prediction. Numbers of transcripts within different categories are shown in the bar plot. Bars showing the number of transcripts with novel ORFs are highlighted. **C** Top 15 *CLN3* transcripts based on the number of samples in which they are detected (occurrence, left panel) are selected. The occurrence numbers are shown in the left panel, coloured by categories of these transcripts. The middle panel shows the structures of transcripts, the coloured taller boxes show the ORFs while the grey shorter boxes show the UTRs. The transcript usage is shown in the right panel. Note, *CLN3* transcripts are present on the antisense strand, so read right to left. Transcription start sites (TSSs) and polyadenylation sites (PASs) signals are shown under the transcript structures, aligned by genomic coordinates. **D** Eight transcripts with the same ORF as the 'major' disease-associated transcript but different UTRs are detected. This plot shows the structural differences between these 'major' transcripts and the canonical *CLN3* transcript "CLN3_24_438aa_5UTR_136_3UTR_49". All grey boxes show structures of the 'major' transcripts, with taller grey boxes showing the ORF and shorter grey boxes showing the UTRs. The pink boxes show sequences in the canonical transcript but not in the 'major' transcripts, blue boxes show sequences in the 'major' transcripts but not in the canonical transcript

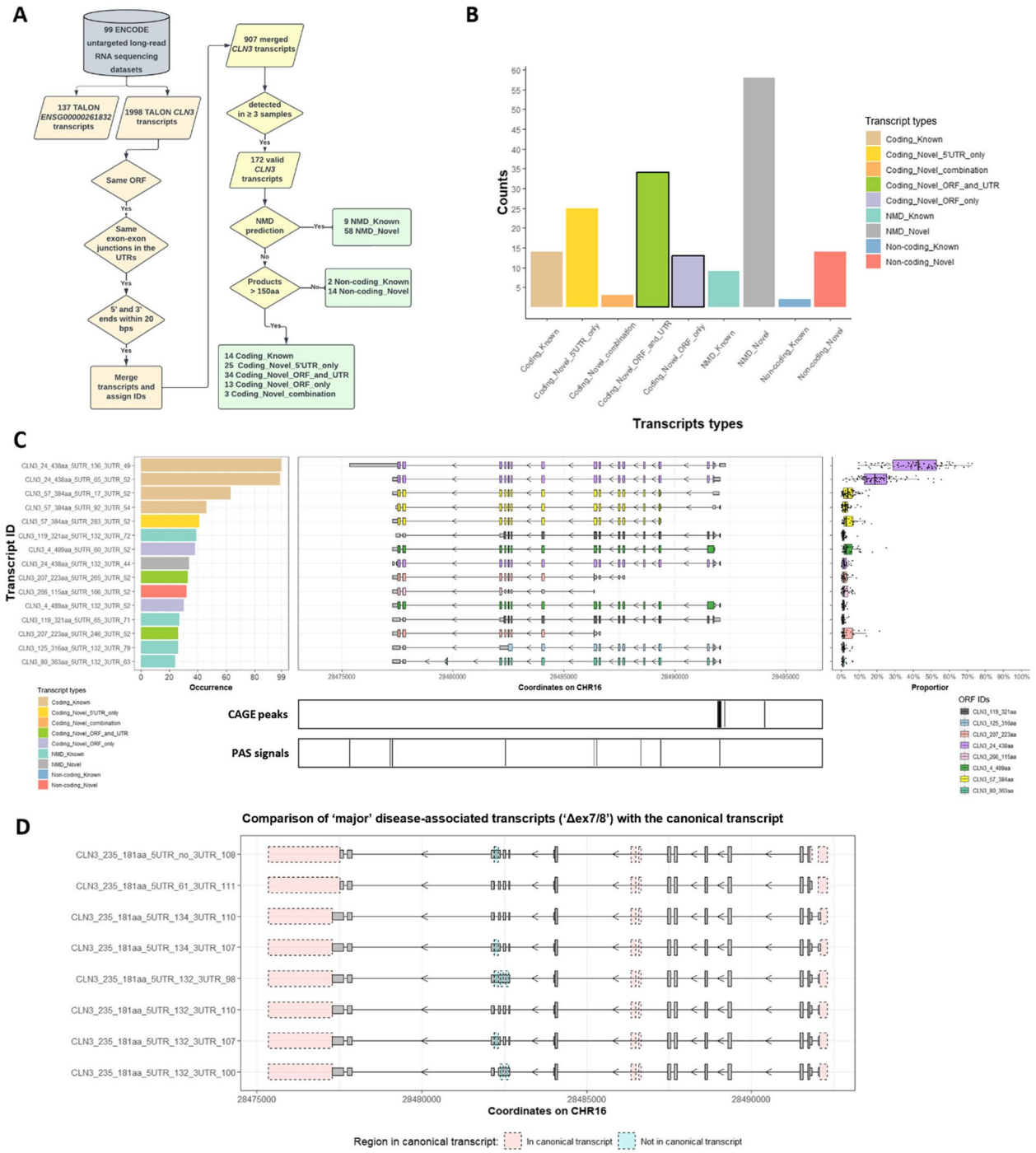

Fig. 3 (See legend on previous page.)

"CLN3_4_489aa" and "TLEGKKK" for "CLN3_235_181aa" are detected with protein Q-value < 0.05 in multiple samples (Fig. 6B). The latter is particularly surprising given that "CLN3_235_181aa" belongs to the 'major' disease-associated transcripts and they are predicted to undergo NMD. The identification of relevant peptides supports the translation of at least two transcripts in the human brain and whole blood. We further analysed the structures of these two protein isoforms and found that they show different orientations when compared with the canonical CLN3 protein based on AlphaFold 2 predictions: an inverted N-terminus in "CLN3_4_489aa" (Fig. 6D) and an inverted C-terminus in

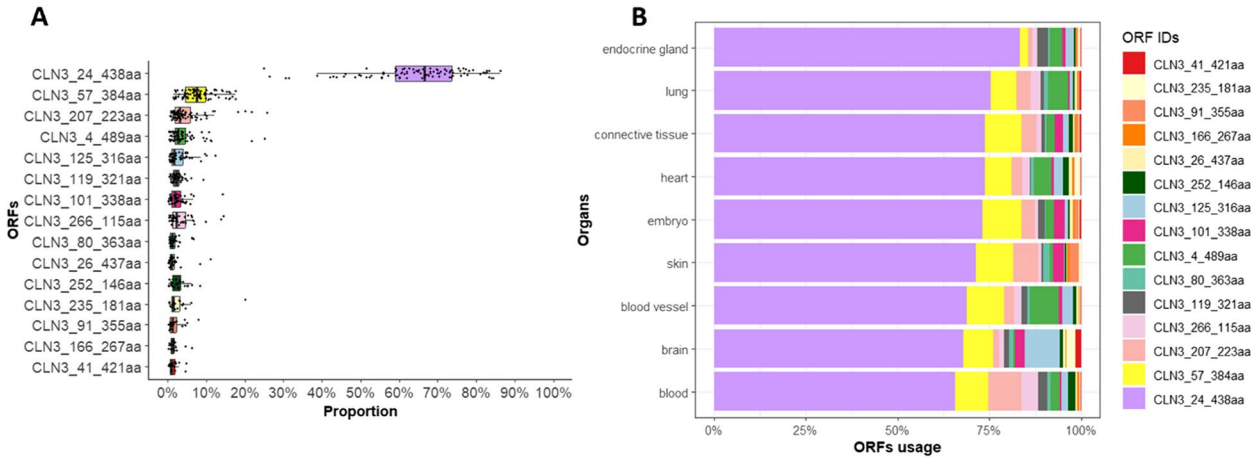

Fig. 4 Summary of the top 15 *CLN3* open reading frames (ORFs). The top 15 ORFs calculated by summing usage of all *CLN3* transcripts containing specific ORFs are plotted in **A**; the canonical *CLN3* ORF "CLN3_24_438aa" has a median usage of 66.7%, therefore around one-third of *CLN3* transcripts encode different protein isoforms. The top 15 ORFs' usage in different organs is plotted in **B**. The usage of the transcripts containing non-canonical ORFs is variable across organs

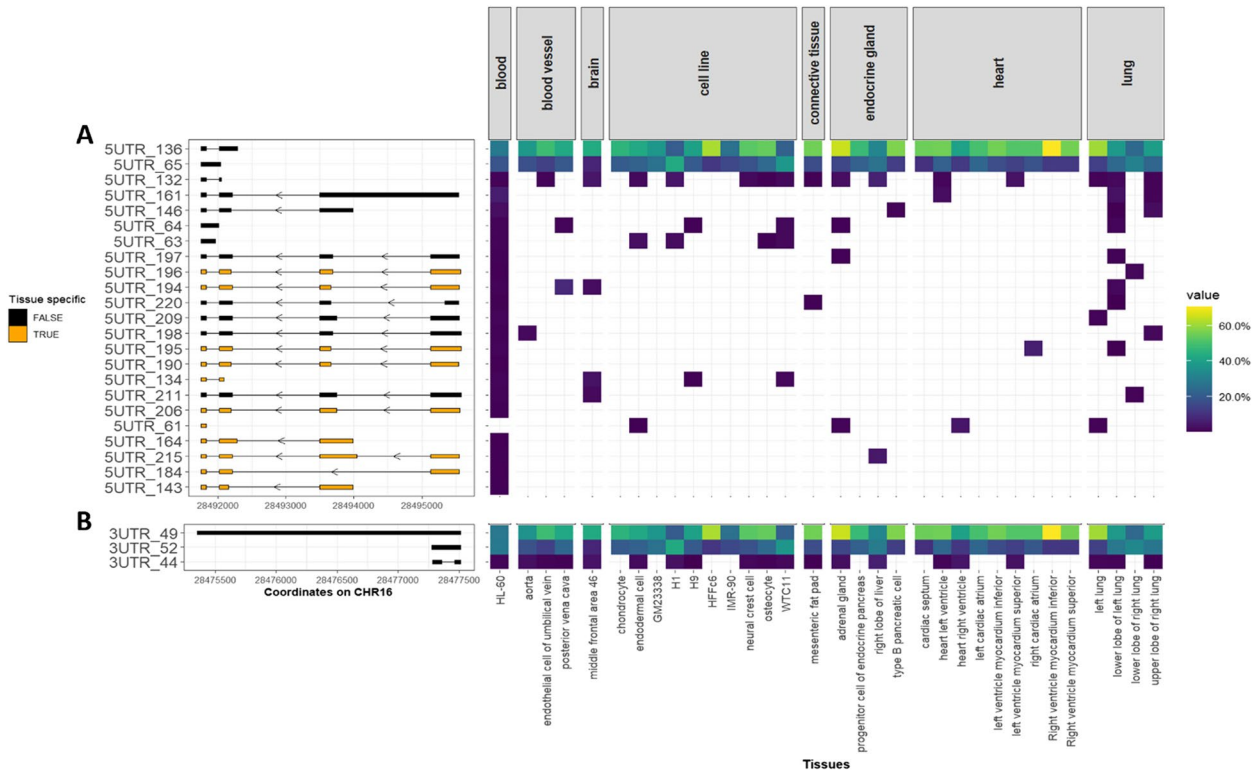

Fig. 5 The same *CLN3* open reading frames (ORFs) are associated with different 5' untranslated regions (5'UTRs) and 3' untranslated regions (3'UTRs). The presence of untranslated regions (UTRs) for three selected *CLN3* ORFs is plotted; UTRs with tissue-specific expression are shown in orange, and non-tissue-specific UTRs are shown in black. The usage of these UTRs (with specific ORFs) in different tissues is plotted using heat maps with yellow showing high usage and dark blue showing low usage. Cell lines are grouped together and shown separately. **A** Twenty-three different 5'UTRs were found with the canonical *CLN3* ORF "CLN3_24_438aa", 16 of them show additional upstream exons, 11 of them show tissue-specificity; **B** Three different 3'UTRs were identified with the canonical ORF "CLN3_24_438aa", none of them shows tissue-specificity. The distal 3'UTR, "3UTR_49", is specific to the canonical ORF

"CLN3_235_181aa" (Fig. 6C). These findings suggest potential differences in the functionality of these two protein isoforms as compared to canonical CLN3 protein.

Discussion

The *CLN3* transcription shows greater complexity than other genes. The number of transcripts for *CLN3* is 62 on Ensembl 110, which is much higher than the average of 3.42 transcripts per gene for GTEx V8 short-read RNA sequencing data [47]. We detected 172 *CLN3* transcripts with a large number of unannotated *CLN3* transcripts ($N=147$) in this study. Interestingly, only 25 out of the 64 annotated transcripts (GENCODE 29) are detected, plus five annotated transcripts not detected in three or more samples. There are 34 transcripts without support from ENCODE long-read RNA-seq data, in addition to the MANE select *CLN3* transcript ENST00000636147.2 in the current annotation (Ensembl 110, GENCODE 44). This MANE transcript does not match the most commonly used transcript (CLN3_24_438aa_5UTR_136_3 UTR_49) or any other *CLN3* transcripts detected in our study. In fact, within our *CLN3* data, there is no dominantly expressed *CLN3* transcript. The most abundant *CLN3* transcript has median usage of 42.9% across 99 samples, well below the estimated usage of dominant transcripts of around 85% in human tissues and 75% in cell lines [44]. In addition, the median usage of transcripts containing the canonical ORF is 66.7%, leaving around one-third of *CLN3* transcripts encoding non-canonical CLN3 protein isoforms. These findings highlight the importance of investigating the full variety of *CLN3* transcripts and studying *CLN3* in terms of highly used non-canonical transcripts and their products in both healthy individuals and juvenile CLN3 disease patients to better understand their role in disease pathogenesis. Also, given the important regulatory role of alternative splicing in tissue-/cell- and development-specificity [48, 49], the high transcript variability necessitates studying the spatial and temporal localisation of different *CLN3* transcripts in relation to biological effects.

The large number of novel transcripts identified in this study can be partly explained by the fact that previous discoveries based on short-read RNA sequencing data would have been limited by multi-mapping by the *CLN3-NPIPB7* readthrough gene, ENSG00000261832. ENSG00000261832 is annotated due to the existence of transcripts generated through the splicing together of *CLN3* and *NPIPB7* exons. This gene contains exons of both *CLN3* and *NPIPB7* and so will result in duplicated sequences in the reference, preventing the unique mapping of *CLN3* short-read sequencing data [50]. These multi-mapped reads are typically removed in multiple sources including GTEx [26], IntroVerse [51], and recount3 [52]. As a result, *CLN3* may have been both inaccurately quantified and annotated. Future long-read RNA sequencing studies for *CLN3* will need to consider this readthrough gene to ensure that the analysis pipeline does not duplicate or remove reads inappropriately.

The existence of *CLN3* transcripts with the same ORFs but different UTRs brings further complexity to *CLN3* annotation. The different transcription start sites detected in this study suggest the presence and use of alternative promoters for *CLN3*. Work is needed to investigate whether the usage of alternative promoters to generate transcripts with the same ORF but different 5'UTRs is cell- or tissue-specific [53], and whether regulatory elements, such as binding sites of RNA-binding proteins (RBPs), or secondary structures [54] are regulating translation efficiency [55] or mRNA stability [56]. Most *CLN3* transcripts have closely located polyadenylation sites (PASs), however the 3'UTR of the most abundant *CLN3* transcript in our dataset has a distinct structure spanning 2164 nucleotides and is detected in all selected samples. Work is needed to investigate whether these different 3'UTRs serve to regulate gene expression levels by altering the mRNA localization, regulatory elements including micro-RNA and RBPs binding sites, or NMD status [57–62].

Relevant to understanding CLN3 disease is our observation of two specific *CLN3* transcripts previously thought to be generated only in disease states are

(See figure on next page.)

Fig. 6 Validation of *CLN3* open reading frames (ORFs) with unique sequences. Selected ORFs with unique sequences (CLN3_4_489aa, CLN3_125_316aa, CLN3_101_338aa, CLN3_80_363aa, CLN3_46_414aa, CLN3_235_181aa, and CLN3_115_328aa) are compared with the canonical *CLN3* ORF "CLN3_24_438aa". Grey boxes show exons of these ORFs, with pink boxes showing regions included in the canonical ORF but not in these ORFs and blue boxes showing regions included in these ORFs but not in the canonical ORF in **A**. In **B**, detected peptides "ADSAPGGHARSGRAPESR" for "CLN3_4_489aa" and "TLEGKKK" for "CLN3_235_181aa" are aligned to the corresponding ORFs, with amino acids that are not present in the canonical CLN3 protein shown as "Novel". Protein products structures of "CLN3_24_438aa", "CLN3_4_489aa" and "CLN3_235_181aa" (from the 'major' transcripts) are predicted by AlphaFold 2.0 in **C** and **D**. The structure of the canonical CLN3 protein (blue and grey in **C** and **D**) is aligned against products of "CLN3_235_181aa" (orange and cream in **C**) and "CLN3_4_489aa" (orange and cream in **D**). Both protein isoforms show different orientations compared with the canonical 438aa CLN3 protein, "CLN3_235_181aa" has an inverted C-terminus and "CLN3_4_489aa" has an inverted N-terminus

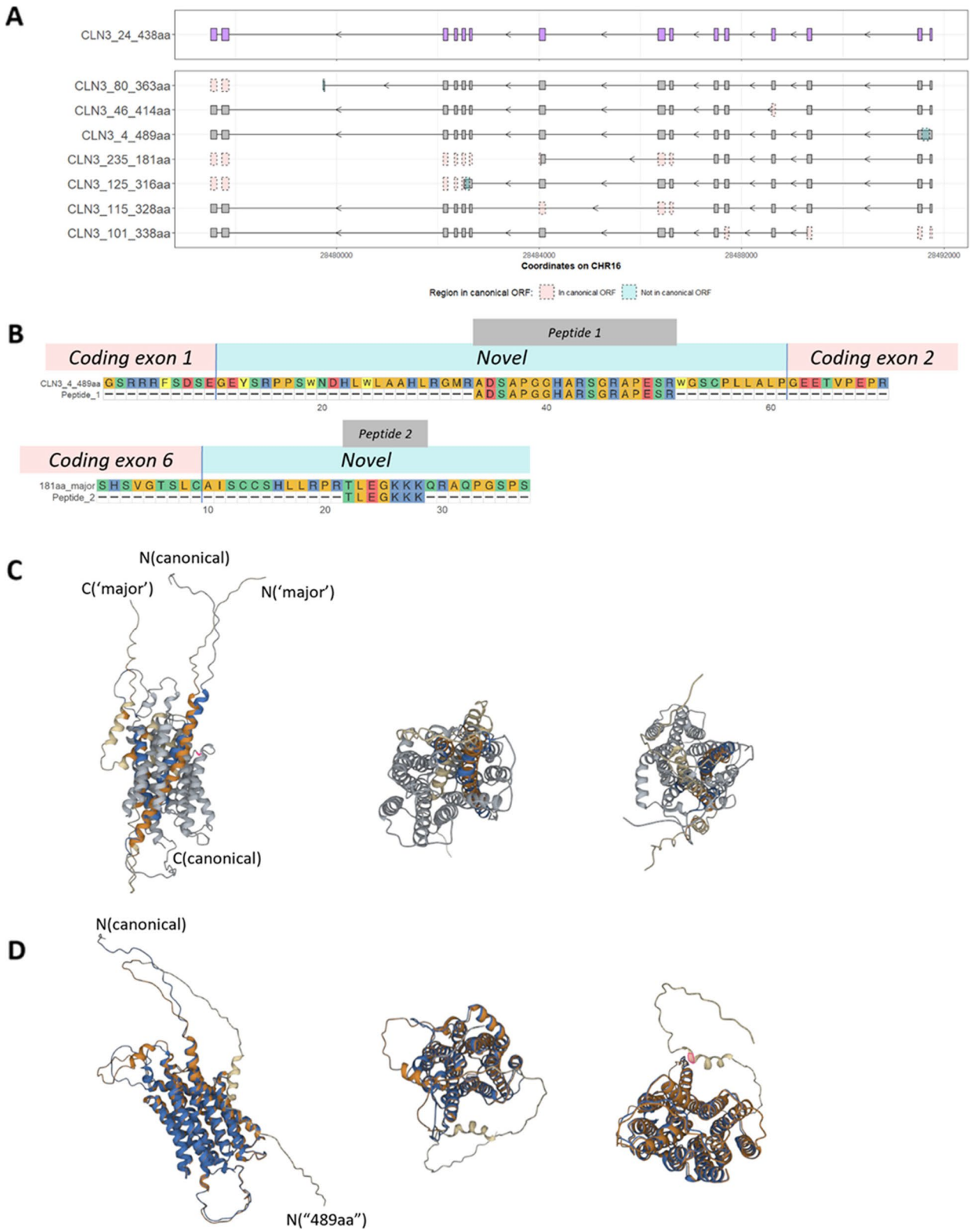

Fig. 6 (See legend on previous page.)

detected in "healthy" control samples. Due to the rarity of juvenile CLN3 disease [63], it is extremely unlikely that all ENCODE tissue donors containing these disease-associated transcripts are carriers or patients. Instead, these transcripts are likely to be naturally occurring, but significantly increase in expression in patients with classic juvenile CLN3 disease. Detection of these disease-associated transcripts in healthy individuals suggests that their mere presence is not sufficient to cause juvenile CLN3 disease. Instead, it indicates the importance of expression levels of disease-associated transcripts in disease pathogenesis. Previous research has indicated these translated disease-associated transcripts of *CLN3* retain some functions and are not deleterious [12]. This suggests that these translated protein isoforms might be involved in normal cellular processes, and that disease might result from dysregulation in the balance of CLN3 protein isoforms, i.e., decreased expression of canonical protein isoform and increased expression of the translated disease-associated isoforms. Furthermore, according to in-house and public RNA sequencing data from GTEx and ENCODE, the predominant variant-associated mis-spliced transcripts detected in patients with other diseases can also be found as rare splice junctions in control data [36]. This has been reported for *BRCA1* and *BRCA2* variant-associated transcripts using targeted RNA sequencing [64]. These findings align with our data, suggesting a potentially common mechanism in which disease-associated transcripts are products of normal alternative splicing but contribute to pathogenesis when their regulation is disrupted. The results of long-read RNA sequencing of patients with CLN3 disease are not yet available to allow comparisons of the relative usage of transcripts from healthy tissues.

We also found that 'Δex5/7/8' transcripts (CLN3_99_339aa_5UTR_63_3UTR_52 and CLN3_99_339aa_5UTR_132_3UTR_52) are naturally occurring. The production of 'Δex5/7/8' transcripts by ASO-induced coding exon 5 skipping in $Cln3^{\Delta ex7/8}$ mice is used to ameliorate the phenotypes of the model [46]. However, it is not yet known if the function of the 'Δex5/7/8' transcripts can fully replace function of the canonical transcript, and if the disease can be prevented or slowed in humans by a similar approach, as the $Cln3^{\Delta ex5/7/8}$ mice still showed elevated subunit c of mitochondrial ATP synthase accumulation, the main accumulated storage material found in Batten disease patients, at later ages [65]. Thus, further studies comparing the function of the 'Δex5/7/8' protein in comparison with the canonical CLN3 protein, and long-term studies for the therapy, are needed. Furthermore, the induced coding exon 5 skipping has to be considered within the full context of the complexity of *CLN3* transcription, including the naturally occurring alternative splicing, alternate transcription start sites (TSSs) and polyadenylation sites (PASs), to make sure that no toxic peptides are produced and that the desired product is highly expressed.

The existence of different CLN3 protein isoforms is supported by mass spectrometry data. Notably, all 'major' disease-associated transcripts containing ORF "CLN3_235_181aa" are predicted to undergo NMD based on the distance between the most downstream exon-exon junction and the stop codons. It has been reported that transcripts with premature termination codons (PTCs) could escape NMD and produce truncated proteins [66, 67]. A previous study of *DMD*, which encodes dystrophin, has shown truncated proteins could rescue the Duchenne muscular dystrophy disease phenotypes if they escape NMD [68]; this could be an interesting avenue to explore in CLN3 disease. Thus, protein-level evidence will be crucial to determine if a transcript with a PTC is NMD-sensitive or produces a truncated protein.

Currently, cryo-electron microscopy or X-ray crystallography data for the canonical CLN3 protein is not available. Therefore, we are relying on AlphaFold topology predictions which show the canonical CLN3 protein has the N-terminus and C-terminus at different sides of the membrane [69]. However, the topology predictions for "CLN3_235_181aa" and "CLN3_4_489aa" show the N-terminus and C-terminus of each protein isoform are on the same sides of the membrane. With the relative orientation of domains/motifs changing, the function of these isoforms could be different to the canonical CLN3 protein [70]. "CLN3_235_181aa", which has the first 153 amino acids of the canonical CLN3 protein and 28 non-canonical out-of-frame amino acids, is missing two lysosomal targeting motifs, ^{242}EEE(X)8LI^{254}, and ^{409}M(X)9G^{419}, of the canonical protein [71, 72]. For "CLN3_4_489aa", which has the N-terminus presenting at the opposite side as the N-terminus of the canonical CLN3 protein, the interaction with protein Rab7 which controls vesicular transport to late endosomes and lysosomes might be affected [73]. Consequently, loss of the terminal portion of the protein (e.g., CLN3_235_181aa) or having an inverted terminus (e.g., CLN3_4_489aa) could lead to different trafficking and thus different functions compared with the canonical CLN3 protein. This identification of potentially structurally and functionally diverse protein isoforms provides us with future new perspectives from which to examine the biology of CLN3 and disease pathogenesis.

Conclusion

Overall, this study expands our knowledge of *CLN3* transcription. We found that the overlapping *CLN3-NPIPB7* readthrough gene could affect accurate quantification and annotation of *CLN3*. With this information, we identified novel *CLN3* transcripts, as well as novel ORFs and UTRs, and characterised their usage. Our analysis reveals there is no dominantly expressed *CLN3* transcript and around one-third of transcripts encoding non-canonical CLN3 protein isoforms have not been studied before. We also identified transcripts that were thought to be expressed only in juvenile CLN3 disease patients in healthy controls. Additionally, alternative UTRs suggest potential regulatory roles which could affect CLN3 protein abundance. Utilising mass spectrometry data to validate the detection of transcripts/ORFs is a powerful tool. The detection of peptides belonging to non-canonical CLN3 protein isoforms demonstrates the translational potential of corresponding transcripts/ORFs and may have functional biological relevance. This highlights the importance of investigating non-canonical protein isoforms in detail to further our understanding of their functions and their role in disease pathogenesis. The insights gained from this study have implications for studying patient samples in the future, as they provide a valuable reference when examining how the 1-kb deletion affects the *CLN3* transcription and contributes to disease pathogenesis.

Abbreviations

NCLs	Neuronal ceroid lipofuscinoses
MFS	Major facilitator superfamily
ORFs	Open-reading frames
UTRs	Untranslated regions
TPM	Transcripts per million
RPK	Reads per kilobase
ID	Identifier
CDS	Coding sequence
NMD	Nonsense-mediated decay
TSS	Transcription start site
PAS	Polyadenylation site
polyA	Polyadenylation
MSA	Multiple system atrophy
NSCLC	Non-small-cell lung cancer
RBP	RNA-binding protein
PTC	Premature termination codon

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12920-024-02017-z.

Supplementary Material 1.

Supplementary Material 2.

Acknowledgements

Not applicable.

Authors' contributions

Conceptualisation: C.M., M.R. and S.E.M; Methodology: H.Y.Z., C.M., E.G., and M.R.; Data acquisition and analysis: H.Y.Z.; Data interpretation: H.Y.Z., C.M., E.G., M.R., and S.E.M.; Writing – original draft: H.Y.Z.; Writing – review & editing: C.M., E.G., M.R. and S.E.M.; Visualization: H.Y.Z.; Supervision: M.R. and S.E.M.; Project administration: C.M. and S.E.M.; Funding acquisition: C.M. M.R and S.E.M. All authors read and approved the final manuscript.

Funding

This work was supported by awards from the UK Medical Research Council (MR/V033956 to S.E.M. (ORCID: 0000–0003-4385–4957), C.M. (ORCID: 0000–0003-4115–8763), M.R. (ORCID: 0000–0001-9520–6957)), and the USA Children's Brain Disease Foundation (to S.E.M., C.M.). E.G. (ORCID: 0000–0003-0541–7537) was supported by the Postdoctoral Fellowship Program in Alzheimer's Disease Research from the BrightFocus Foundation (Award Number: A2021009F). M.R. was supported through the award of a Tenure Track Clinician Scientist Fellowship (MR/N008324/1). All research at Great Ormond Street Hospital NHS Foundation Trust and UCL Great Ormond Street Institute of Child Health is made possible by the NIHR Great Ormond Street Hospital Biomedical Research Centre. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health.

Availability of data and materials

Comprehensive gene annotation of GENCODE release 29 (GENCODE 29) was downloaded from https://www.gencodegenes.org/human/release_29.html. Gene TPMs GTEx_Analysis_2017-06-05_v8_RNASeQCv1.1.9_gene_reads.gct.gz and exon-exon junction read counts GTEx_Analysis_2017-06-05_v8_STARv2.5.3a_junctions.gct.gz of GTEx V8 can be accessed from GTEx portal (https://gtexportal.org/home/datasets). The ENCODE long-read RNA sequencing data used in this paper was downloaded from ENCODE portal (https://www.encodeproject.org/). Public mass spectrometry datasets PXD026370 and PXD028605 can be downloaded from ProteomeXchange (https://www.proteomexchange.org/). Codes used to analyse the data and produce figures are accessible on GitHub (https://github.com/HYzhang800/CLN3_public_long_read).

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 16 January 2024 Accepted: 23 September 2024
Published online: 04 October 2024

References

1. Williams RE, Mole SE. New nomenclature and classification scheme for the neuronal ceroid lipofuscinoses. Neurology. 2012;79(2):183–91.
2. Mitchison HM, Thompson AD, Mulley JC, Kozman HM, Richards RI, Callen DF, et al. Fine genetic mapping of the Batten disease locus (CLN3) by haplotype analysis and demonstration of allelic association with chromosome 16p microsatellite loci. Genomics. 1993;16(2):455–60.
3. Mitchison HM, O'Rawe AM, Taschner PE, Sandkuijl LA, Santavuori P, de Vos N, et al. Batten disease gene, CLN3: linkage disequilibrium mapping in the Finnish population, and analysis of European haplotypes. Am J Hum Genet. 1995;56(3):654–62.
4. The International Batten Disease Consortium. Isolation of a novel gene underlying Batten disease, CLN3. Int Batten Dis Consortium Cell. 1995;82(6):949–57.
5. Schulz A, Kohlschutter A, Mink J, Simonati A, Williams R. NCL diseases - clinical perspectives. Biochim Biophys Acta. 2013;1832(11):1801–6.
6. Lebrun AH, Moll-Khosrawi P, Pohl S, Makrypidi G, Storch S, Kilian D, et al. Analysis of potential biomarkers and modifier genes affecting the clinical course of CLN3 disease. Mol Med. 2011;17(11–12):1253–61.

7. Gardner E, Mole SE. The genetic basis of phenotypic heterogeneity in the neuronal ceroid lipofuscinoses. Front Neurol. 2021;12: 754045.
8. Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, et al. InterPro: the integrative protein signature database. Nucleic Acids Res. 2009;37(Database issue):D211-5.
9. Laqtom NN, Dong W, Medoh UN, Cangelosi AL, Dharamdasani V, Chan SH, et al. CLN3 is required for the clearance of glycerophosphodiesters from lysosomes. Nature. 2022;609(7929):1005–11.
10. Nyame K, Hims A, Aburous A, Laqtom NN, Dong W, Medoh UN, et al. Glycerophosphodiesters inhibit lysosomal phospholipid catabolism in Batten disease. Mol Cell. 2024;84(7):1354-64e9.
11. Munroe PB, Mitchison HM, O'Rawe AM, Anderson JW, Boustany RM, Lerner TJ, et al. Spectrum of mutations in the Batten disease gene, CLN3. Am J Hum Genet. 1997;61(2):310–6.
12. Kitzmuller C, Haines RL, Codlin S, Cutler DF, Mole SE. A function retained by the common mutant CLN3 protein is responsible for the late onset of juvenile neuronal ceroid lipofuscinosis. Hum Mol Genet. 2008;17(2):303–12.
13. Minnis CJ, Townsend S, Petschnigg J, Tinelli E, Bahler J, Russell C, et al. Global network analysis in Schizosaccharomyces pombe reveals three distinct consequences of the common 1-kb deletion causing juvenile CLN3 disease. Sci Rep. 2021;11(1):6332.
14. 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;44(D1):D733–45.
15. Cunningham F, Allen JE, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, et al. Ensembl 2022. Nucleic Acids Res. 2022;50(D1):D988–95.
16. Frankish A, Diekhans M, Jungreis I, Lagarde J, Loveland JE, Mudge JM, et al. Gencode 2021. Nucleic Acids Res. 2021;49(D1):D916–23.
17. Leung SK, Jeffries AR, Castanho I, Jordan BT, Moore K, Davies JP, et al. Full-length transcript sequencing of human and mouse cerebral cortex identifies widespread isoform diversity and alternative splicing. Cell Rep. 2021;37(7): 110022.
18. Wenger AM, Peluso P, Rowell WJ, Chang PC, Hall RJ, Concepcion GT, et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat Biotechnol. 2019;37(10):1155–62.
19. Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat Biotechnol. 2018;36(4):338–45.
20. Gustavsson EK, Sethi S, Gao Y, Brenton JW, Garcia-Ruiz S, Zhang D, et al. The annotation of GBA1 has been concealed by its protein-coding pseudogene GBAP1. Sci Adv. 2024;10(26):eadk1296.
21. Evans JR, Gustavsson EK, Doykov I, Murphy D, Virdi GS, Lachica J, et al. The diversity of SNCA transcripts in neurons, and its impact on antisense oligonucleotide therapeutics. bioRxiv. 2024:2024.05.30.596437.
22. Dainis A, Tseng E, Clark TA, Hon T, Wheeler M, Ashley E. Targeted long-read RNA sequencing demonstrates transcriptional diversity driven by splice-site variation in MYBPC3. Circ Genom Precis Med. 2019;12(5): e002464.
23. Joglekar A, Hu W, Zhang B, Narykov O, Diekhans M, Marrocco J, et al. Single-cell long-read sequencing-based mapping reveals specialized splicing patterns in developing and adult mouse and human brain. Nat Neurosci. 2024;27(6):1051–63.
24. Patowary A, Zhang P, Jops C, Vuong CK, Ge X, Hou K, et al. Developmental isoform diversity in the human neocortex informs neuropsychiatric risk mechanisms. Science. 2024;384(6698):eadh7688.
25. Zhang Z, Schwartz S, Wagner L, Miller W. A greedy algorithm for aligning DNA sequences. J Comput Biol. 2000;7(1–2):203–14.
26. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 2013;45(6):580–5.
27. Luo Y, Hitz BC, Gabdank I, Hilton JA, Kagda MS, Lam B, et al. New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. Nucleic Acids Res. 2020;48(D1):D882–9.
28. Wyman D, Balderrama-Gutierrez G, Reese F, Jiang S, Rahmanian S, Forner S, et al. A technology-agnostic long-read analysis pipeline for transcriptome discovery and quantification. bioRxiv. 2020:672931.
29. Pertea G, Pertea M. GFF utilities: GffRead and GffCompare. F1000Research. 2020;9:9.
30. Tjeldnes H, Labun K, Torres Cleuren Y, Chyzynska K, Swirski M, Valen E. ORFik: a comprehensive R toolkit for the analysis of translation. BMC Bioinformatics. 2021;22(1):336.
31. Lawrence M, Huber W, Pages H, Aboyoun P, Carlson M, Gentleman R, et al. Software for computing and annotating genomic ranges. PLoS Comput Biol. 2013;9(8): e1003118.
32. Hamid F, Alasoo K, Vilo J, Makeyev E. Functional annotation of custom transcriptomes. Methods Mol Biol. 2022;2537:149–72.
33. 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. Bioinformatics. 2022;38(15):3844–6.
34. Abugessaisa I, Noguchi S, Hasegawa A, Kondo A, Kawaji H, Carninci P, et al. refTSS: a reference data set for human and mouse transcription start sites. J Mol Biol. 2019;431(13):2407–22.
35. Li Q, Lai H, Li Y, Chen B, Chen S, Li Y, et al. RJunBase: a database of RNA splice junctions in human normal and cancerous tissues. Nucleic Acids Res. 2021;49(D1):D201–11.
36. Dawes R, Bournazos AM, Bryen SJ, Bommireddipalli S, Marchant RG, Joshi H, et al. SpliceVault predicts the precise nature of variant-associated mis-splicing. Nat Genet. 2023;55(2):324–32.
37. Herrmann CJ, Schmidt R, Kanitz A, Artimo P, Gruber AJ, Zavolan M. PolyASite 2.0: a consolidated atlas of polyadenylation sites from 3' end sequencing. Nucleic Acids Res. 2020;48(D1):D174–9.
38. Rydbirk R, Ostergaard O, Folke J, Hempel C, DellaValle B, Andresen TL, et al. Brain proteome profiling implicates the complement and coagulation cascade in multiple system atrophy brain pathology. Cell Mol Life Sci. 2022;79(6):336.
39. Molloy MP, Hill C, O'Rourke MB, Chandra J, Steffen P, McKay MJ, et al. Proteomic analysis of whole blood using volumetric absorptive microsampling for precision medicine biomarker studies. J Proteome Res. 2022;21(4):1196–203.
40. Solntsev SK, Shortreed MR, Frey BL, Smith LM. Enhanced global post-translational modification discovery with MetaMorpheus. J Proteome Res. 2018;17(5):1844–51.
41. Zhou L, Feng T, Xu S, Gao F, Lam TT, Wang Q, et al. ggmsa: a visual exploration tool for multiple sequence alignment and associated data. Brief Bioinform. 2022;23(4):bbac222.
42. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596(7873):583–9.
43. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The protein data bank. Nucleic Acids Res. 2000;28(1):235–42.
44. Gonzalez-Porta M, Frankish A, Rung J, Harrow J, Brazma A. Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene. Genome Biol. 2013;14(7): R70.
45. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, et al. Alternative isoform regulation in human tissue transcriptomes. Nature. 2008;456(7221):470–6.
46. Centa JL, Jodelka FM, Hinrich AJ, Johnson TB, Ochaba J, Jackson M, et al. Therapeutic efficacy of antisense oligonucleotides in mouse models of CLN3 Batten disease. Nat Med. 2020;26(9):1444–51.
47. Tung KF, Pan CY, Chen CH, Lin WC. Top-ranked expressed gene transcripts of human protein-coding genes investigated with GTEx dataset. Sci Rep. 2020;10(1):16245.
48. Barbosa-Morais NL, Irimia M, Pan Q, Xiong HY, Gueroussov S, Lee LJ, et al. The evolutionary landscape of alternative splicing in vertebrate species. Science. 2012;338(6114):1587–93.
49. Merkin J, Russell C, Chen P, Burge CB. Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. Science. 2012;338(6114):1593–9.
50. Deschamps-Francoeur G, Simoneau J, Scott MS. Handling multi-mapped reads in RNA-seq. Comput Struct Biotechnol J. 2020;18:1569–76.
51. Garcia-Ruiz S, Gustavsson EK, Zhang D, Reynolds RH, Chen Z, Fairbrother-Browne A, et al. IntroVerse: a comprehensive database of introns across human tissues. Nucleic Acids Res. 2023;51(D1):D167–78.
52. Wilks C, Zheng SC, Chen FY, Charles R, Solomon B, Ling JP, et al. recount3: summaries and queries for large-scale RNA-seq expression and splicing. Genome Biol. 2021;22(1):323.
53. Davuluri RV, Suzuki Y, Sugano S, Plass C, Huang TH. The functional consequences of alternative promoter use in mammalian genomes. Trends Genet. 2008;24(4):167–77.
54. Lim Y, Arora S, Schuster SL, Corey L, Fitzgibbon M, Wladyka CL, et al. Multiplexed functional genomic analysis of 5' untranslated region mutations across the spectrum of prostate cancer. Nat Commun. 2021;12(1):4217.

55. Hinnebusch AG, Ivanov IP, Sonenberg N. Translational control by 5′-untranslated regions of eukaryotic mRNAs. Science. 2016;352(6292):1413–6.
56. Yun Y, Adesanya TM, Mitra RD. A systematic study of gene expression variation at single-nucleotide resolution reveals widespread regulatory roles for uAUGs. Genome Res. 2012;22(6):1089–97.
57. Di Giammartino DC, Nishida K, Manley JL. Mechanisms and consequences of alternative polyadenylation. Mol Cell. 2011;43(6):853–66.
58. MacDonald CC. Tissue-specific mechanisms of alternative polyadenylation: testis, brain, and beyond (2018 update). Wiley Interdiscip Rev RNA. 2019;10(4): e1526.
59. Kebaara BW, Atkin AL. Long 3′-UTRs target wild-type mRNAs for nonsense-mediated mRNA decay in Saccharomyces cerevisiae. Nucleic Acids Res. 2009;37(9):2771–8.
60. Hogg JR, Goff SP. Upf1 senses 3′UTR length to potentiate mRNA decay. Cell. 2010;143(3):379–89.
61. Barrett LW, Fletcher S, Wilton SD. Regulation of eukaryotic gene expression by the untranslated gene regions and other non-coding elements. Cell Mol Life Sci. 2012;69(21):3613–34.
62. Kurosaki T, Popp MW, Maquat LE. Quality and quantity control of gene expression by nonsense-mediated mRNA decay. Nat Rev Mol Cell Biol. 2019;20(7):406–20.
63. Haltia M, Goebel HH. The neuronal ceroid-lipofuscinoses: a historical introduction. Biochim Biophys Acta. 2013;1832(11):1795–800.
64. Brandao RD, Mensaert K, Lopez-Perolio I, Tserpelis D, Xenakis M, Lattimore V, et al. Targeted RNA-seq successfully identifies normal and pathogenic splicing events in breast/ovarian cancer susceptibility and Lynch syndrome genes. Int J Cancer. 2019;145(2):401–14.
65. Centa JL, Stratton MP, Pratt MA, Osterlund Oltmanns JR, Wallace DG, Miller SA, et al. Protracted CLN3 Batten disease in mice that genetically model an exon-skipping therapeutic approach. Mol Ther Nucleic Acids. 2023;33:15–27.
66. Neu-Yilik G, Amthor B, Gehring NH, Bahri S, Paidassi H, Hentze MW, et al. Mechanism of escape from nonsense-mediated mRNA decay of human beta-globin transcripts with nonsense mutations in the first exon. RNA. 2011;17(5):843–54.
67. Inoue K, Ohyama T, Sakuragi Y, Yamamoto R, Inoue NA, Yu LH, et al. Translation of SOX10 3′ untranslated region causes a complex severe neurocristopathy by generation of a deleterious functional domain. Hum Mol Genet. 2007;16(24):3037–46.
68. Kerr TP, Sewry CA, Robb SA, Roberts RG. Long mutant dystrophins and variable phenotypes: evasion of nonsense-mediated decay? Hum Genet. 2001;109(4):402–7.
69. Jarvela I, Sainio M, Rantamaki T, Olkkonen VM, Carpen O, Peltonen L, et al. Biosynthesis and intracellular targeting of the CLN3 protein defective in Batten disease. Hum Mol Genet. 1998;7(1):85–90.
70. Chen Q, Denard B, Lee CE, Han S, Ye JS, Ye J. Inverting the topology of a transmembrane protein by regulating the translocation of the first transmembrane helix. Mol Cell. 2016;63(4):567–78.
71. Kyttala A, Yliannala K, Schu P, Jalanko A, Luzio JP. AP-1 and AP-3 facilitate lysosomal targeting of Batten disease protein CLN3 via its dileucine motif. J Biol Chem. 2005;280(11):10277–83.
72. Storch S, Pohl S, Braulke T. A dileucine motif and a cluster of acidic amino acids in the second cytoplasmic domain of the batten disease-related CLN3 protein are required for efficient lysosomal targeting. J Biol Chem. 2004;279(51):53625–34.
73. Bucci C, Thomsen P, Nicoziani P, McCarthy J, van Deurs B. Rab7: a key to lysosome biogenesis. Mol Biol Cell. 2000;11(2):467–80.

Publisher's Note