TYPE Original Research
PUBLISHED 19 September 2024
DOI 10.3389/fgene.2024.1451024

Check for updates

OPEN ACCESS

EDITED BY
Andrey Grigoriev,
The State University of New Jersey,
United States

REVIEWED BY
Qingnan Liang,
University of Texas MD Anderson Cancer Center, United States
Lingyu Guan,
Children's Hospital of Philadelphia,
United States

*CORRESPONDENCE
Peter A. C. 't Hoen,
✉ Peter-Bram.tHoen@radboudumc.nl

RECEIVED 18 June 2024
ACCEPTED 30 August 2024
PUBLISHED 19 September 2024

CITATION
Riepe TV, Stemerdink M, Salz R, Rey AD, de Bruijn SE, Boonen E, Tomkiewicz TZ, Kwint M, Gloerich J, Wessels HJCT, Delanote E, De Baere E, van Nieuwerburgh F, De Keulenaer S, Ferrari B, Ferrari S, Coppieters F, Cremers FPM, van Wyk E, Roosing S, de Vrieze E and 't Hoen PAC (2024) A proteogenomic atlas of the human neural retina.
Front. Genet. 15:1451024.
doi: 10.3389/fgene.2024.1451024

A proteogenomic atlas of the human neural retina

Tabea V. Riepe[1,2,3,4], Merel Stemerdink[5], Renee Salz[1], Alfredo Dueñas Rey[6,7], Suzanne E. de Bruijn[2,3], Erica Boonen[2,3,4], Tomasz Z. Tomkiewicz[2,3,4], Michael Kwint[2], Jolein Gloerich[8], Hans J. C. T. Wessels[8], Emma Delanote[6,7], Elfride De Baere[6,7], Filip van Nieuwerburgh[9], Sarah De Keulenaer[9], Barbara Ferrari[10], Stefano Ferrari[10], Frauke Coppieters[6,7,11], Frans P. M. Cremers[2,3,4], Erwin van Wyk[5], Susanne Roosing[2,3,4], Erik de Vrieze[5] and Peter A. C. 't Hoen[1]*

[1]Department of Medical BioSciences, Radboud University Medical Center, Nijmegen, Netherlands, [2]Department of Human Genetics, Radboud University Medical Center, Nijmegen, Netherlands, [3]Academic Alliance Genetics, Radboud University Medical Center, Nijmegen, Gelderland, Netherlands, [4]Maastricht University Medical Center+, Maastricht, Netherlands, [5]Department of Otorhinolaryngology, Radboud University Medical Center, Nijmegen, Gelderland, Netherlands, [6]Center for Medical Genetics, Ghent University Hospital, Ghent, Belgium, [7]Department of Biomolecular Medicine, Ghent University, Ghent, Belgium, [8]Department of Human Genetics, Translational Metabolic Laboratory, Radboud University Medical Center, Nijmegen, Gelderland, Netherlands, [9]NXTGNT, Faculty of Pharmaceutical Sciences, Ghent University, Ghent, Belgium, [10]Fondazione Banca degli Occhi del Veneto, Venice, Italy, [11]Department of Pharmaceutics, Ghent University, Ghent, Belgium

The human neural retina is a complex tissue with abundant alternative splicing and more than 10% of genetic variants linked to inherited retinal diseases (IRDs) alter splicing. Traditional short-read RNA-sequencing methods have been used for understanding retina-specific splicing but have limitations in detailing transcript isoforms. To address this, we generated a proteogenomic atlas that combines PacBio long-read RNA-sequencing data with mass spectrometry and whole genome sequencing data of three healthy human neural retina samples. We identified nearly 60,000 transcript isoforms, of which approximately one-third are novel. Additionally, ten novel peptides confirmed novel transcript isoforms. For instance, we identified a novel *IMPDH1* isoform with a novel combination of known exons that is supported by peptide evidence. Our research underscores the potential of in-depth tissue-specific transcriptomic analysis to enhance our grasp of tissue-specific alternative splicing. The data underlying the proteogenomic atlas are available via EGA with identifier EGAD50000000101, via ProteomeXchange with identifier PXD045187, and accessible through the UCSC genome browser.

KEYWORDS

neural retina, isoform, alternative splicing, multi-omics, long-read sequencing, proteogenomics, mass spectrometry, inherited retinal disease (IRD)

1 Introduction

The neural human retina, located at the back of the eye, is a light sensitive tissue composed of six distinct neuronal cell types (rod and cone photoreceptor cells, bipolar cells, horizontal cells, ganglion cells, and amacrine cells) and Müller glial cells. Its primary function is to convert light into electric stimuli that can be interpreted by the brain. Variants in genes responsible for retina function can disrupt this light conversion process resulting in inherited retinal diseases (IRDs). Approximately 1 in 1,500 individuals worldwide are affected by IRD-associated vision loss (Ben-Yosef, 2022). Similar to other neuronal tissues, the retina is enriched for tissue-specific splicing (Cao et al., 2011; Liu and Zack, 2013). Previous research has revealed retina-specific exons, transcript isoforms, and splicing regulators (Murphy et al., 2016; Jayasinghe et al., 2018; Ciampi et al., 2022). For instance, recent discoveries revealed that the dominant *CRB1* isoform in photoreceptors is not the canonical *CRB1-A* isoform that is expressed in Müller glial cells but the novel *CRB1-B* isoform (Ray et al., 2020). Moreover, it was discovered that splicing factors MSI1 (Murphy et al., 2016) and PCBP2 (Ling et al., 2020) play a role in rod-specific splicing events. It is estimated that at least 11% of causative variants in IRD-associated genes interfere with pre-mRNA splicing (Bacchi et al., 2014; Khan et al., 2020).

Understanding of the expressed transcript and protein isoforms, as well as knowledge of retina-specific splicing events, is required to correctly classify genetic variants identified in IRD patients (Farkas et al., 2013; Swamy et al., 2020; Aísa-Marín et al., 2021). Several mechanisms by which retina-specific splicing influences variant classification have been revealed. Firstly, variants in *ABCA4* and *BBS8* can lead to photoreceptor-specific pseudo-exon inclusions by activating cryptic splice sites or exonic splice enhancers that are only recognized by the photoreceptor splicing machinery (Riazuddin et al., 2010; Murphy et al., 2015; Albert et al., 2018). Other variants, like the frequent deep intronic Leber congenital amaurosis-associated variant c.2991 + 1655A>G in *CEP290,* demonstrated a higher pseudo-exon inclusion in retinal organoids compared to lymphocytes and fibroblasts (Den Hollander et al., 2006; Parfitt et al., 2016). Another mechanism through which the retina transcriptome affects variant calling involves retina-specific isoforms. Notably, about 80% of variants in *RPGR* are located in a retina-specific exon in a non-canonical isoform called ORF15 (Liu and Zack, 2013). These examples emphasize the importance of studying tissue-specific splicing when interpreting variants in IRD patients.

Several retinal transcriptome studies have been published (Farkas et al., 2013; Pinelli et al., 2016; Ratnapriya et al., 2019; Schumacker et al., 2020; Ruiz-Ceja et al., 2023). However, the existing literature provides limited evidence on the transcript isoforms expressed in the retina for two main reasons. Firstly, the human neural retina can only be accessed post-mortem, and rapid extraction and sample preservation are crucial to maintaining RNA integrity. Therefore, collecting high-quality neural retina samples for transcriptome studies is challenging. Secondly, current datasets are derived from short-read RNA-sequencing, and inference of full-length transcript isoforms comes with uncertainties (Sarantopoulou et al., 2021). Recent advances in long-read RNA-sequencing allow the sequencing of entire transcripts, providing insight into the expressed transcript isoforms and encoded open reading frames (ORFs). Moreover, earlier studies primarily focused on transcript isoforms, but it is also important to investigate the influence of transcriptome diversity on proteome diversity (Smith et al., 2013).

To address the lack of transcript and protein isoform data of the retina, we generated a comprehensive retina proteogenomic atlas by combining PacBio long-read mRNA sequencing data of three high-quality human post-mortem neural retina samples with mass spectrometry (MS)-based proteomic data and short-read whole genome sequencing (WGS) data.

2 Materials and methods

2.1 Tissue collection for WGS, PacBio mRNA sequencing and mass-spectrometry

Human neural retinal samples (n = 3) from non-visually impaired individuals were obtained from Fondazione Banca degli Occhi del Veneto (Venice, Italy), with written consent from the donor's next of kin to be used for research purposes in accordance with the tenets of the Declaration of Helsinki. Detailed information about the donors is shown in Supplementary Table S1. The eyes were enucleated within 2–12 h after donor's death and the neural retina extraction was performed as described previously (Niyadurupola et al., 2011; Osborne et al., 2016). Briefly, after retrieval of the donor cornea, the eyeball was cut at the ora serrata; iris, lens, and the vitreous body were removed and the retina carefully detached from the sclera and the retinal pigment epithelium, cutting at the optic nerve head. The retinal samples were transferred into cryovials and snap-frozen in liquid nitrogen. Samples were shipped in dry ice.

2.2 DNA isolation and WGS library preparation

DNA was isolated from frozen retinal tissue with the QIAamp DNA mini kit (QIAGEN, Aarhus, Denmark) following the standard protocol. The samples were sequenced using a 2 × 150 base pair (bp) paired-end module on a BGISeq500. The minimal median coverage per genome was 30-fold.

2.3 WGS data analysis

Burrows-Wheeler Aligner (v.0.7814) was used to map the reads to the Human Reference Genome build GRCh38/hg38. Single-nucleotide variants (SNVs), structural variants (SVs), and copy number variants (CNVs) were called with Genome Analysis Toolkit HaplotypeCaller (Broad Institute) (Li and Durbin, 2009), Manta structural variant caller (Chen et al., 2016), and Canvas Copy Number Variant Caller (Roller et al., 2016), respectively. After sequencing, the data was processed with an in-house pipeline for pathogenic variant detection using the Human Reference Genome build GRCh38/hg38 according to the procedure of de Bruijn et al. (2023). To ensure that no IRD-associated pathogenic variants were observed in our non-IRD samples, we assessed the WGS data for putative pathogenic variants in currently known IRD-associated genes. In short, all SNVs with allele frequencies higher than 0.5% in gnom AD v.3.1.2 were discarded. We prioritized variants with the predicted effects: stop-loss or gain, start-loss or gain, frameshift, in-frame deletion or insertion. Moreover, we included canonical splice variants, missense variants near splice sites, non-canonical splice variants, and deep intronic and exonic splice variants with a SpliceAI score ≥0.2. Next, missense variants with a phyloP score ≥2.7, CADD_PHRED score ≥15, or Grantham score ≥80, and silent variants with a phyloP score ≥2.7 or a CADD_PHRED score ≥15 were prioritized. For remaining variants that met the criteria, we checked the ACMG/AMP classification using Franklin Genoox Platform (https://franklin.genoox.com). Additionally, ClinVar (Landrum et al., 2018) and LOVD (Fokkema et al., 2011) provided information about previous occurrences of the variants in individuals with IRD. For CNVs and SVs, we prioritized variants with an allele frequency in the inhouse control SV and CNV database smaller or equal to 0.5%. Variants spanning an exon in an IRD-associated gene were evaluated. Inversion events were only considered when at least one of the breakpoints was located within an IRD-associated gene.

2.4 RNA isolation and PacBio library preparation

To obtain the total RNA from human retina samples, 500 µL of Trizol was added and each sample was homogenized in two rounds using a Tissuelyser kit (QIAGEN, Aarhus, Denmark) for 30 s at 30 Hz. After a 5-minute incubation at room temperature (RT), 100 µL chloroform was added, samples were mixed, incubated for 3 min at RT, and centrifuged at 12,000 g for 15 min. Afterwards, the aqueous phase was mixed with glycogen (5 µg/µL) and 1 volume of isopropanol, and the resulting mixture was incubated at 20°C for 75 min and subsequently centrifuged at 12,000 g for 30 min at 4°C. The supernatant was discarded, and the resulting RNA pellet was further purified and DNAse treated using the Nucleospin RNA Clean-up Kit (Macherey-Nagel, Duren, Germany) according to the manufacturer's protocol. The total isolated RNA was quantified using a Qubit fluorometer (Thermo Fisher Scientific, Waltham, MA, United States) and RNA integrity number (RIN) values were assessed using a 2,100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, United States). The RIN values of the three samples were above 7.0 and 300 ng of RNA input was used to generate the Iso-Seq SMRTbell libraries using the Iso-Seq-Express-Template-Preparation protocol version 2.0 (Pacific Biosciences, California, United States). Libraries were prepared following the standard workflow (for samples composed primarily of transcripts centered ~2 kb) and using the SMRTbell® library binding kit 2.1 (Pacific Biosciences, California, United States). The on-plate loading concentration of the final Iso-Seq SMRTbell libraries was 80 pM, and a 24-h movie time was used for sequencing on a Sequel II system (Pacific Biosciences, California, United States).

2.5 IsoSeq data analysis

The Iso-Seq data was processed as recommended by PacBio. Circular consensus sequencing (CCS) reads with a minimum read accuracy of 0.99, maximum length of 25,000, and minimum number of three passes were generated using CCS (v6.2.0). Lima (v2.4.0) was applied to remove primers and SMRT adapters to generate full-length (FL) reads. Full-length non concatemer (FLNC) reads were generated with Isoseq3 (v3.4.0) refine and converted from BAM to FASTQ. Minimap2 (v2.24) with parameters *-ax splice -uf--secondary = no -C5 -O6,24 -B4 --MD* was used to align the reads to the Human Reference Genome build GRCh38/hg38. The mapped isoforms were classified using IsoQuant (v.3.1.2) (Prjibelski et al., 2023) with the GENCODE v39 primary assembly annotation (Frankish et al., 2019). As parameters we applied*--complete_gened, --datatype pacbio_ccs, --fl_data, --sqanti_output, --count_exons, --model_construction_strategy fl_pacbio, --check_canonical, --transcript_quantification unique_only, gene_quantification unique_only*, and*--splice_correction_strategy default_pacbio*. A list of 294 IRD-associated genes was downloaded from RetNet (Daiger et al., 1998) (19 May 2022) to filter for IRD-associated genes.

2.6 Protein digestions

The neural retinal tissue was homogenated with the PlusOne Sample grinding kit following the standard protocol. Retina protein homogenates were subjected to in-solution digestion using trypsin (Promega), chymotrypsin (Merck), or a combination of LysC (Wako chemicals) and AspN (Merck). For each digest, 10 µg of total protein was reduced by incubation at RT for 20 min with 1 µL 10 mM dithiotrytol (Sigma-Aldrich) followed by incubation for 20 min at RT in the dark with 1 µL 50 mM chloroacetamide (Sigma-Aldrich). For tryptic digestion, alkylated proteins were pre-digested by addition of 0.2 µg LysC enzyme and incubated for 3 h at RT. Samples were diluted with 3 volumes of 50 mM ammoniumbicarbonate (Sigma-Aldrich) prior to the addition of 0.2 µg trypsin and subsequent overnight incubation at 37°C. Chymotrypsin digestion was performed by diluting alkylated proteins with 3 volumes of 50 mM ammoniumbicarbonate prior to addition of 0.2 µg chymotrypsin and subsequent overnight incubation at 25°C. For the combined LysC + AspN digestion, alkylated proteins were pre-digested for 3 h at RT using 0.2 µg LysC after which the sample was diluted by adding 3 volumes of 50 mM ammoniumbicarbonate prior to overnight digestion with AspN at 37°C in 1 mM methylamine (Sigma-Aldrich). Digests were diluted 1:1 with 2% trifluoroacetic acid and stored at −20°C prior to shotgun proteomics analysis.

2.7 Shotgun proteomics

All samples were measured using an Evosep One nanoflow liquid chromatography system (Evosep) connected online to a timsTOF Pro2 mass spectrometer (Bruker Daltonics) via a CaptiveSprayer nanoflow electrospray ionization source (Bruker Daltonics). From each digest, 200 ng was loaded onto Evosep tips according to manufacturer's instructions and separated using the pre-defined 30 samples per day protocol in combination with an Evosep 150 mm × 0.15 mm C18 reversed phase column (EV1106 endurance column packed with 1.9 µm C18AQ particles from Dr Maisch) using 0.1% formic acid (Merck) in ultra-pure water (Biosolve) as solvent A and 0.1% formic acid in acetonitrile (Biosolve) as solvent B. The mass spectrometer was operated in positive ionization mode using the instrument default 1.1 s duty cycle time data dependent acquisition parallel accumulation serial fragmentation (dda-PASEF) method. Settings: 100 ms trapped ion mobility spectrometry (TIMS) accumulation and ramp times, 0.6–1.6 1/K0 mobility range, 100–1,700 m/z range, 10 PASEF frames, 20eV collision energy at 0.6 1/K0 up to 59eV at 1.6 1/K0, dynamic exclusion enabled for 0.4 min. The timsTOF Pro2 instrument was calibrated prior to measurements using ESI-Low tune mix (Agilent technologies) infusion.

2.8 Proteomics analysis

To create a retina-specific peptide search database, we modified the long-read proteogenomic pipeline by Miller et al. (2022) to the IsoQuant output. We generated Python scripts to convert the IsoQuant output Gene transfer format (GTF) file into a SQANTI-like classification file. The ORFs of novel transcripts were predicted with Coding-Potential Assessment Tool (CPAT) (v3.0.4), and ORFs were called and combined. A GTF file with coding regions was created and the coding sequences were renamed to exons. After using SQANTI3 protein to classify the ORFs, the SQANTI3 classification of the transcripts was replaced by the IsoQuant classification. The untranslated regions (UTRs) were added, proteins were classified, renamed, and filtered for nonsense mediated decay, protein truncation or unlikely protein classification. Philosopher (v4.8.1) was used to add decoys to the PacBio-GENCODE hybrid database. MSFragger (v3.7) was run with open search parameters and digestion enzymes stricttrypsin, chymotrypsin, or aspn and lysc-p with a maximum of two missed cleavages. The fragment mass tolerance was set to 20 ppm. We included carbamidomethyl of cysteine as fixed modification and oxidation of methionine and acetylation of the N-terminus as variable modifications. The false discovery rate (FDR) was 1%, and protein quantification was performed using IonQuant (v1.8.10). We adapted the scripts by Miller *et al.* for the MSFragger output to create files for visualization in the genome browser and to detect novel peptides. Pyteomics (v4.6) (Goloborodko et al., 2013; Levitsky et al., 2018) was used to perform *in silico* digestion of the hybrid database into peptides with a minimum length of seven amino acids and a maximum of two missed cleavages. Spectra of novel peptides were visualized with PDV (v1.7.4) for manual validation (Li et al., 2019). The criteria for manual validation included detecting at least eight ions, including both b- and y-ions, a probability higher than 0.95, and the majority of peaks being accounted for by fragment ions.

2.9 Tissue collection for oxford nanopore technology sequencing

For the independent ONT validation dataset, rest material from human donors (n = 3) without a history or clinical evidence of retinal disease was collected from either Ghent or Antwerp University Hospital tissue banks in accordance with the ethical principles of the Declaration of Helsinki and under ethical approval of the Ethics Committee of the Ghent University Hospital (IRB approval B670201837286). More information about the donors is shown in Supplementary Table S2. Eyes were transported in CO2 Independent Medium (Gibco) until dissection. Protocols for retina dissection were optimized in-house. To limit the possible effects of autolysis time on RNA integrity, retinas were isolated only from eyes with a total post-mortem interval lower than 20 h. After visual inspection to exclude any cross-contamination with retinal pigment epithelium, neural retinas were immediately processed (total RNA isolation) or snap-frozen and stored at −80°C until used (total RNA isolation).

2.10 RNA isolation and oxford nanopore technology sequencing

Total RNA from post-mortem adult human neural retina was extracted following manufacturer's guidelines (RNeasy Mini kit®, Qiagen) followed by DNase treatment (ArcticZymes, Tromsø, Norway) and poly-A capture. Poly-A mRNA samples of sufficient quality (RNA Integrity Value, RIN>8.0) were subjected to direct-cDNA (SQK-DCS109, ONT) library preparation following the supplier protocol with minor adaptations. Each library was then loaded (FLO-PRO002, ONT) and sequenced on a PromethION device (ONT) for 72 h. Information about the number of reads can be found in Supplementary Table S2.

2.11 ONT data analysis (selected genes)

MinKNOW (v5.1.0) was applied to generate fast5 files, which were then base called with Guppy (v6.1.5). Reads were aligned to the Human Reference Genome build GRCh38/hg38 using minimap2 (v2.24) with the -ax splice flags to allow spliced alignments. Alignment files were converted to BAM format, sorted, and indexed using SAMtools (v1.15). The mapped isoforms were classified using IsoQuant (v.3.1.2) (Prjibelski et al., 2023) with the GENCODE v39 primary assembly annotation (Frankish et al., 2019). As parameters we applied*--complete_gened, --datatype nanopore, --fl_data, --model_construction_strategy default_ont, --check_canonical, --transcript_quantification unique_only, gene_quantification unique_only*, and*--splice_correction_strategy default_ont*. Additionally, StringTie2 (v2.1.1) was run for transcriptome assembly using the -L parameter and the reference transcriptome annotations (Ensembl human release 103) as guide. The three individual GTF files for selected genes were merged with gffcompare (v.0.12.6). For both transcriptomes, aFASTA file was created using TransDecoder (v.5.5.0), ORFs were predicted using CPAT (v.3.0.4), and UTRs were added to the CPAT output.

2.12 Gene ontology analysis

Gene ontology (GO) enrichment analysis of the 300 most expressed genes was performed using goatools (v1.3.1) and Benjamini-Hochberg false discovery rate correction. We included biological process, molecular function, and cellular component GO terms and used an adjusted p-value cutoff of smaller than 0.05.

2.13 Comparison of PacBio and reference transcripts

To compare the length of PacBio transcripts to reference transcripts, a list of protein coding genes was downloaded from BioMart (14 March

2023) to select protein coding genes from the IsoQuant output. Additionally, a list of protein-coding reference transcripts and their corresponding length was downloaded (12 July 2022).

2.14 Correlation with number of isoforms

The length of the different transcripts for all genes was downloaded from BioMart (16 May 2023) and for each gene, the longest transcript was selected. The FL count for each gene was calculated using the IsoQuant output and the correlation coefficient was calculated using the scipy (v1.10.1) Spearman correlation.

2.15 Enrichment analysis of novel isoforms in RetNet genes

The distribution of known and novel transcripts between all genes and RetNet genes was compared using a chi squared test.

2.16 Length and TPM comparison between known and novel transcripts

Comparison of length between novel and known isoforms was tested using a Mann-Whiney-U test.

2.17 CAGE peak analysis

To test if the TSS of the IsoQuant transcripts were supported by CAGE peak data, we made use of data and scripts from SQANTI3 (Tardaguila et al., 2018). We therefore used refTSS CAGE peaks (v3.1) (Abugessaisa et al., 2019) and the search window was set to 50 base pairs upstream of the TSS.

2.18 Comparison to short-read sequencing studies

First, the data of the retina-specific exons from Murphy et al. (2016) and Ciampi et al. (2022) was converted into a BED file and exons of IsoQuant transcripts were extracted from the IsoQuant GTF file. Bedtools (v2.27.1) intersect with parameters -f, -u and 1 was used to find the previously reported retina-specific exons that were also included in the IsoQuant transcriptome.

2.19 Analysis of novel transcript and protein isoforms

Protein domain predictions were performed with HmmerWeb (v2.41.2) against the Pfam database. AlphaFold (Jumper et al., 2021) predictions were obtained using the AlphaFold3 server with standard settings. Protein structures were aligned in PyMOL (v2.5.5). The *EPB41L2* brain data and *IMPDH1* exon expression data used for comparison were obtained from the GTEx Portal on 12 June 2023.

3 Results

In this study, we created a human neural retina proteogenomic atlas by combining WGS, PacBio long-read RNA-sequencing, and MS-based proteomics data of three healthy human neural retinal samples. DNA-sequencing was used to rule out that the donors were carriers of known pathogenic variants in IRD genes. PacBio long-read RNA-sequencing was applied to discover full-length (FL) transcript isoforms in the retina. After ORF prediction on the novel retina transcripts, the MS-based proteomics data was analyzed using a custom peptide search database to identify novel peptides validating novel transcript isoforms. Additionally, selected transcripts were validated using an independent ONT retina dataset.

3.1 Whole-genome sequencing confirms that the donors do not have an IRD-associated genotype

The isolated DNA of the three samples was sequenced with WGS to rule out the presence of known IRD-associated variants among our study participants. We identified a heterozygous frameshift variant c.7027del in *CEP290*, which is classified as ACMG/AMP pathogenic, in sample 2. Since about 20%–40% of the population carries a recessive IRD-associated variant, this finding was not unexpected (Hanany et al., 2018). Sample 2 was still included in the dataset because further analysis of *CEP290* did not reveal a second causative allele and we did not observe any *CEP290* transcripts that were only detected in sample 2.

3.2 Discovering the transcript landscape in the human neural retina with long-read sequencing

From the three retina samples, we obtained a total of 11.6 M (million) circular consensus sequencing (CCS) reads (Supplementary Table S3) that resulted in 58,541 unique FL transcripts from 15,807 known and 470 novel genes (Supplementary Data Sheet S1). The individual samples had between 3.4 and 4.2 M CCS reads, which is in line with previous PacBio Iso-Seq studies (Leung et al., 2021; Mehlferber et al., 2022). Sample 1 had the lowest number of reads and in this sample the average CCS read length was 3.3 kb compared to around 2.2 kb for the other two samples (Supplementary Figure S1A). Detected transcripts had an average transcript length of 3.2 kb, which is longer than the average of 2.4 kb for GENCODE reference transcripts (Figure 1A). The average transcripts per million (TPM) count for the three different samples is comparable (Supplementary Figure S1B). For 11,225 genes (68.79%), more than one isoform was identified, and for 4,664 genes (28.64%) five or more isoforms were sequenced (Figure 1B). The number of isoforms demonstrated a weak correlation with the longest reference transcript length (Spearman correlation coefficient = 0.19, *p*-value <0.05, Supplementary Figure S1C). Novel transcripts were not lower expressed than known transcripts (Figure 1C). Furthermore, 213 out of 294 IRD-associated genes extracted from RetNet (72.45%) had more than one isoform, and 99 genes (33.33%) had five or more isoforms (Supplementary Figure S1D), which is comparable to the results for the whole dataset. The genes with most isoforms were *ATE1* and

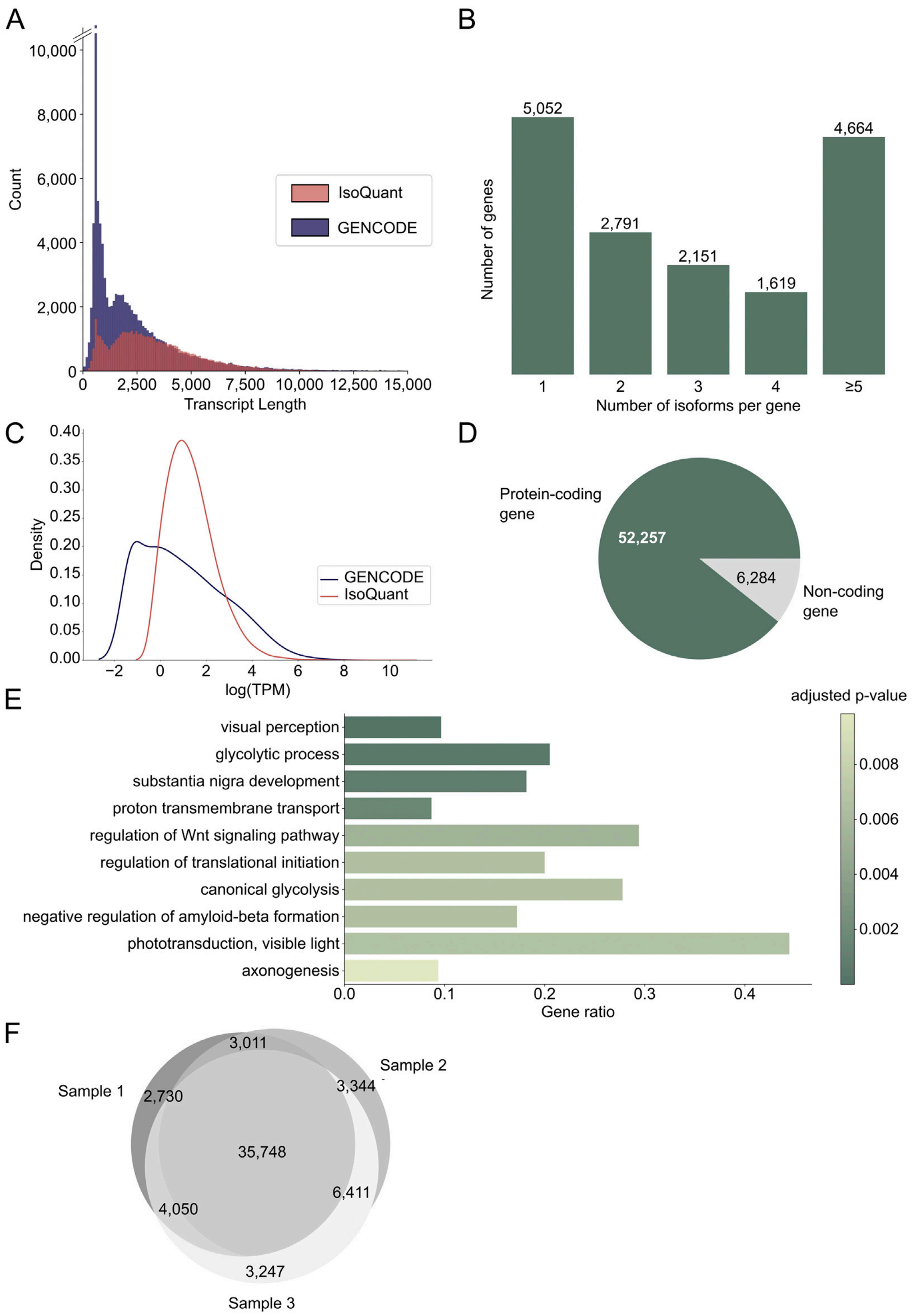

FIGURE 1
Discovering the transcript landscape in the human neural retina with long-read sequencing. **(A)** Transcript length distribution of protein coding PacBio IsoQuant (red) and GENCODE reference (blue) isoforms. **(B)** Number of isoforms detected per gene across the three samples. The number of isoforms for RetNet genes can be found in Supplementary Figure S1D. **(C)** Transcript per million (TPM) count for known GENCODE transcripts (blue) and novel IsoQuant transcripts (red). **(D)** Number of isoforms from protein coding genes (green) and non-coding genes (grey). **(E)** Visualization of the ten most significant hits of the gene ontology (GO) term enrichment analysis of the 300 most expressed genes. The bars are colored by the *p*-value and the gene ratio shows the percentage of genes associated with the GO-term among the 300 genes. **(F)** Overlap of transcripts identified in the three individual samples.

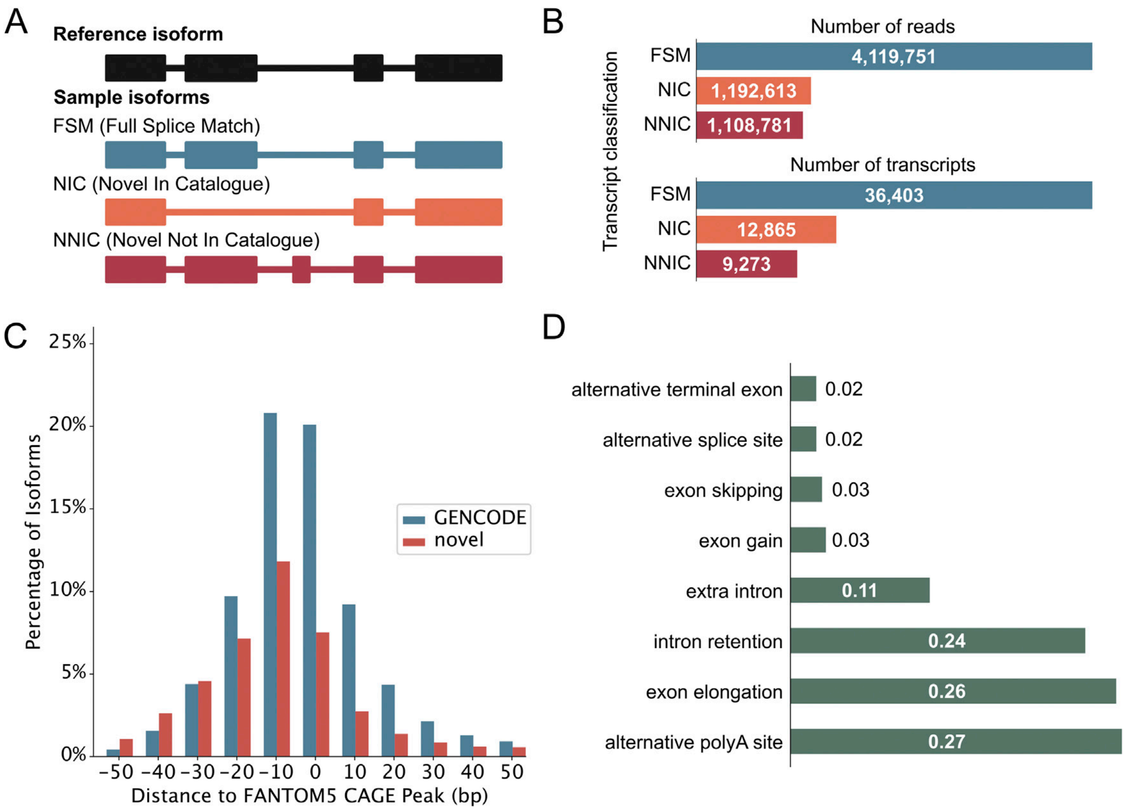

FIGURE 2
Iso-Seq reveals novel neural retinal transcripts. **(A)** IsoQuant classification schematic that compares retina transcripts with GENCODE transcripts. Full Splice Matches (FSMs) match the reference completely. Novel transcripts either match the reference splice junctions and are then called Novel In Catalog (NIC), or they contain novel splice junctions and are called Novel Not In Catalog (NNIC). **(B)** Number of reads and transcripts from the three retina samples associated with each transcript class. The classification of RetNet transcripts is shown in Supplementary Figure S1D. **(C)** Distance of transcription start sites (TSS) of known and novel transcripts to annotated refTSS CAGE peaks. **(D)** Most common novel elements of NIC and NNIC transcripts as classified by IsoQuant. The *x*-axis represents the fraction of transcripts with that specific event.

TMEM161B-DT, both with 30 isoforms. Most transcripts (89%) belonged to protein coding genes (Figure 1D) and gene ontology (GO) enrichment analysis of the 300 most expressed genes revealed that these genes are linked to vision (visual perception, phototransduction, visible light) and the nervous system (substantia nigra development, axonogenesis, negative regulation of amyloid-beta formation) (Figure 1E). Moreover, 84% of the identified transcripts were present in at least two samples, indicating that most transcripts are robustly expressed across individuals (Figure 1F). For the transcripts detected only in one sample, 9,313 map to known transcripts and eight to novel transcripts. 141 of these transcripts, all known, belong to IRD genes. Moreover, 73% of single sample transcripts have an FL count of one, and with an average length of 2.0 kb they are shorter than the average transcript.

3.3 Iso-Seq reveals novel neural retinal transcripts

Following quality control, PacBio transcripts were classified using the IsoQuant classification (Prjibelski et al., 2023). Figure 2A illustrates the different structural categories: Full splice match (FSM), novel in catalog (NIC), and novel not in catalog (NNIC). FSMs match the reference transcript completely, whereas NIC and NNIC are considered novel transcripts. NIC transcripts only contain known splice junctions, however, in new combinations (e.g., through the process of exon skipping). NNIC transcripts contain at least one novel splice junction. We identified 36,403 FSMs, 12,865 NICs, and 9,273 NNICs (Figure 2B). The RetNet genes accounted for 522 novel isoforms from 176 genes, and they are enriched in NIC isoforms compared to the complete dataset (X^2 (2, 59,741) = 24.99, $p = 3.74 \times 10^{-6}$) (Supplementary Figure S1E). Novel transcripts (NIC and NNIC) were longer than known transcripts (Mann-Whitney-U test: W = 5.78×10^8, *p*-value <0.05 (Supplementary Figure S1F). To obtain further evidence for the integrity of our novel transcripts, we aligned the 5'-ends of our transcripts with Cap Analysis Gene Expression (CAGE) peak data marking transcription start sites (TSSs). This revealed that 24,382 (66.98%) TSSs of known transcripts and 7,974 (36.02%) TSSs of novel transcripts were overlapping or within 100 base pairs of a CAGE peak (Figure 2C). The most common novel elements in novel transcripts were alternative polyadenylation [poly(A)], followed by exon elongation and intron retention (Figure 2D). Our results are consistent with previous short-read RNA-sequencing studies (Supplementary Data Sheet S2). However, we can call intron retention events with more confidence than in short-read RNA-sequencing studies, because short-read sequencing

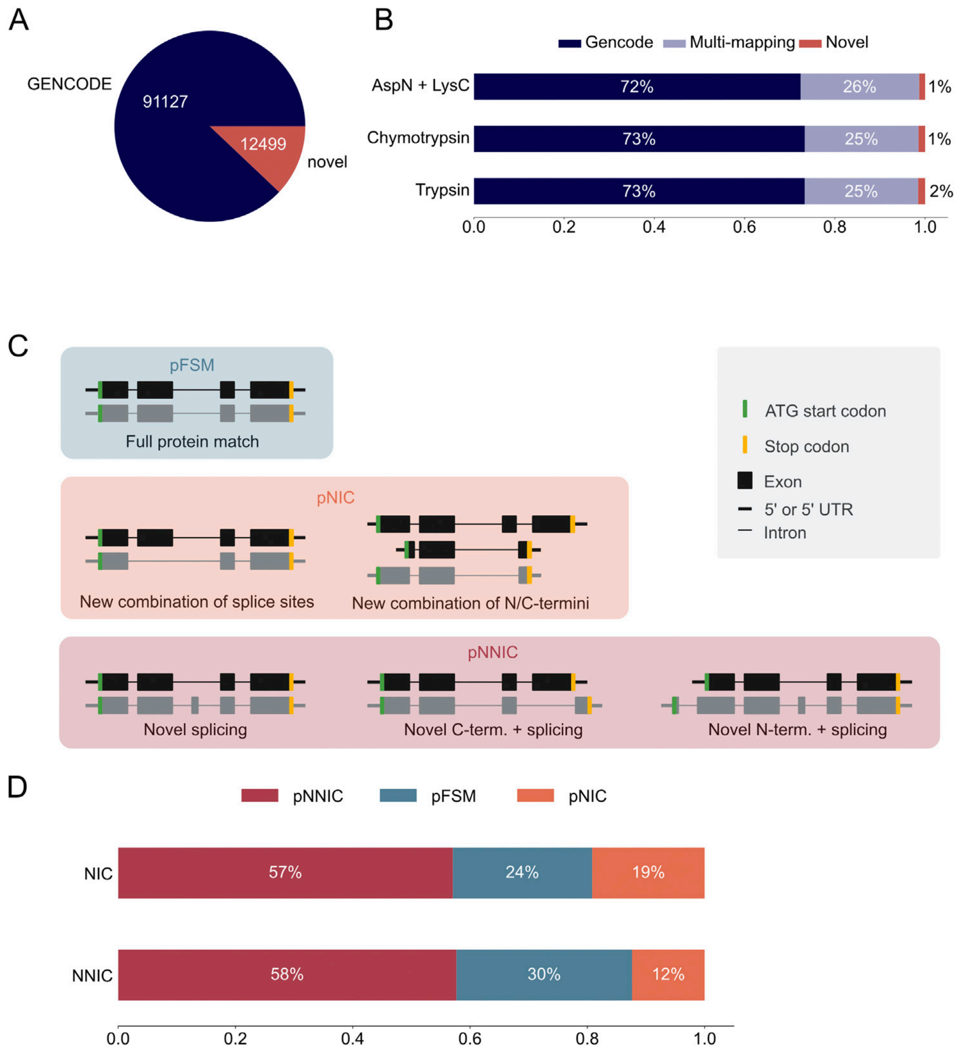

FIGURE 3
Most novel ORFs in the IsoQuant-GENCODE hybrid database contain novel protein sequence elements. **(A)** Number of known GENCODE open reading frames (ORFs) (blue) and novel ORFs in the hybrid database (red). **(B)** *In-silico* digestion of the ORFs in the hybrid database with trypsin, chymotrypsin, or AspN and LysC shows the number of multi-mapping peptides (purple) and uniquely mapping peptides. GENCODE peptides are shown in blue and novel peptides in red. **(C)** The SQANTI Protein classification scheme compares the long-read ORFs to the reference ORFs. ORFs are classified using the N-terminus, C-terminus, and the splice junctions. A protein full splice match (pFSM) isoform matches a reference isoform. In a protein novel in catalog (pNIC) isoform the N-terminus, the C-terminus, and the splice junctions are known but the combination is new. A protein novel not in catalog (pNNIC) isoform contains at least one novel protein element. **(D)** Comparison of the transcript (two different bars) and protein (subdivision in the bars) classification of novel ORFs.

protocols also capture a fraction of nascent mRNAs that are not yet fully spliced (David et al., 2022). In further support of the quality of our data, we identified eight human homologues out of the ten mice photoreceptor-specific exons identified by Murphy et al. (2016) and 45/75 retina-specific short exons and 84/116 retina-specific long exons found by Ciampi et al. (2022).

3.4 Most novel ORFs in the IsoQuant-GENCODE hybrid database contain novel protein sequence elements

The long-read proteogenomic pipeline by Miller et al. (2022) was used to create a custom peptide sequence database for MS-based proteomic data analysis based on the novel transcripts identified with PacBio sequencing. Our custom database contained 12,499 unique novel ORFs from 22,138 novel transcripts (Supplementary Data Sheet S3). We combined these novel ORFs with 91,124 GENCODE ORFs to form a complete peptide search database (Figure 3A). *In silico* digestion of the ORFs in the hybrid database demonstrated that most theoretical peptides map uniquely to a GENCODE ORF, about a quarter map to both a GENCODE and an IsoQuant ORF, and 1,500 peptides map uniquely to a novel IsoQuant ORF (Figure 3B).

Next, we compared the classification of the transcripts with the classification of the ORFs to analyze which transcripts resulted in novel ORFs. Transcripts were classified using IsoQuant as described earlier and ORFs were classified using SQANTI protein (Figure 3C)

TABLE 1 Number of peptides and proteins identified with MSFragger and the custom PacBio-GENCODE hybrid database.

	Trypsin	Chymotrypsin	AspN + LysC	Total (unique)
Number of peptides	18,671	8,649	10,537	33,503
Number of proteins	2,828	1,434	1,855	3,122

(Miller et al., 2022). Both classifications are based on the GENCODE reference annotation. A protein full splice match (pFSM) has a known N-terminus, known splice junctions, and a known C-terminus, and the combination of the three protein sequence elements is known. A protein novel in catalog (pNIC) isoform also only contains known protein sequence elements, but the combination of the elements is novel. If at least one of the protein sequence elements, such as the N- or C-terminus or a splice junction, is novel, the protein isoform is classified as protein novel not in catalog (pNNIC). It should be noted that a novel transcript can still encode for a pFSM because the variation can be located in the untranslated region (UTR). We observed that most novel ORFs are classified as pNNIC, followed by pFSM and pNNC (Figure 3D; Supplementary Table S4). This shows that most novel ORFs are truly novel because of novel protein sequence elements.

3.5 Novel peptides confirm novel IsoQuant isoforms

Digestions with three different enzymes were performed to maximize the sequence coverage of the proteome and the chances of detecting novel peptides. Analysis of the data using the PacBio-GENCODE hybrid database resulted in 33,503 unique peptides from 3,122 unique genes (Table 1). Digestion with trypsin contributed the most peptides (18,671), followed by the combination of AspN and LysC (10,537) and chymotrypsin (8,649). We identified peptides for 15,589 of the GENCODE transcripts and 2,335 of the novel PacBio transcripts. 12,713 peptides were mapped to both a GENCODE and novel transcript while 12 peptides were unique for novel isoforms and 20,723 peptides unique for known isoforms. Ten novel peptides mapping to AMPH, ARHGDIA, EPB41L2, TPM3, and VTI1A passed manual validation (Table 2). The novel peptide in VIT1A initially did not pass manual validation but it was included in the analysis because it was previously detected by Leung et al., (2021) in the cortex. None of the sequences of the novel peptides were present in UniProt. Alternative splicing events confirmed by novel peptides are exon elongation, novel intron, novel exon, and novel combination of known splice sites. The two peptides that were mapped to intron retention events did not pass manual validation of the spectra. An overview of all novel peptides and their spectra including manual validation is presented in Supplementary Data Sheet S4.

Figure 4 shows examples of the novel peptides and their spectra. Five of the novel peptides map to an exon elongation event in *AMPH* (Figure 4A) that is present in three PacBio transcripts (TRANSCRIPT6109.CHR7.NIC, TRANSCRIPT6086.CHR7.NIC, TRANSCRIPT6138.CHR7.NIC), and partly in TRANSCRIPT6110.CHR7.NIC. The transcripts code for proteins with a 102, 406 and 415 amino acid insertion, respectively. The 406 amino acid exon elongation could be confirmed by ONT long-read mRNA sequencing on independent retina samples 1–3 (TRANSCRIPT14.CHR7.NIC).

A second example of a novel peptide is AISEELDHALNDMTSIASLQPT in TPM3 that is derived from TRANSCRIPT38170.CHR1.NNIC and/or TRANSCRIPT38168.CHR1.NNIC (Figure 4B). The novel isoforms contain a 79 nt penultimate exon that is also observed in reference isoforms ENST00000328159.9, ENST00000651731.1, ENST00000368530.7, and ENST00000651641.1. Nevertheless, our newly identified isoforms exhibit a distinct final exon compared to the reference isoforms. While they share the last exon with reference isoforms ENST00000368531.6 ENST00000323144.12, and ENST00000341485.10, the incorporation of the additional exon leads to a reading frame shift and an alternative upstream stop codon. This splice junction is also present in TRANSCRIPT617.CHR1.NNIC, TRANSCRIPT600.CHR1.NNIC, and TRANSCRIPT620.CHR1.NNIC of the ONT retina data.

A third example is the novel splice peptide SHLLESSHETL identified in EPB41L2 (Figure 4C). This peptide maps to a novel junction in TRANSCRIPT22463. CHR6. NNIC. The novel *EPB41L2* isoform contains a novel combination of known splice junctions, resulting in an exon coding for 56 amino acids that is also present in two shorter isoforms, ENST00000527017.6 and ENST00000525198.1. In comparison to the two shorter isoforms, the novel isoform does not contain the next small 54 nucleotide exon, like the shorter reference isoforms, but the exon used by most other reference isoforms. The novel junction is not detected in the ONT data but was found with a FL count of 183 in all three PacBio samples.

3.6 RetNet genes demonstrating highest expression of a novel isoform

For 35 out of 294 RetNet genes, the transcript with the novel ORF demonstrated higher expression than all other transcripts from the same gene containing a known ORF. Seven of the 35 genes, *DYNC2H1*, *RPGRIP1*, *KIAA1549*, *LAMA1*, *CEP290*, *DMD*, *USH2A*, and *EYS*, were not completely covered with the PacBio sequencing due to the length of the transcripts (>8 kb) and were excluded from this analysis. Of the remaining 28 genes, 14 genes had a novel isoform that also contained novel coding sequences (Table 3). Figure 5 shows three examples of these genes.

We identified a novel major isoform in *SAMD11* (TRANSCRIPT529.CHR1.NNIC) (Figure 5A). This isoform contains a new first exon that is also found in TRANSCRIPT556.CHR1.NNIC and TTRANSCRIPT531.CHR1.NNIC. The resulting ORF of TRANSCRIPT529.CHR1.NNIC is missing the N-terminal SAND domain but it still contains the C-terminal SAM domain. The

TABLE 2 Overview of the novel peptides detected with the three different digestion enzymes.

Gene	Digestion enzyme	Transcript	Peptide sequence	Probability[a]	Passes manual validation	Novel event
AMPH	Trypsin	TRANSCRIPT6086.CHR7.NIC	AESSLIEGSER	0.87	Yes	Exon elongation
AMPH	Trypsin	TRANSCRIPT6086.CHR7.NIC	DQDINNSDLSEDEIANQR	1.00	Yes	Exon elongation
AMPH	Trypsin	TRANSCRIPT6086.CHR7.NIC	DTEGLDNSWTHSDVVEHK	1.00	Yes	Exon elongation
AMPH	Trypsin	TRANSCRIPT6086.CHR7.NIC	TLEGTEEFEEK	1.00	Yes	Exon elongation
AMPH	AspN + LysC	TRANSCRIPT6086.CHR7.NIC	DEIANQRYGLLYQEIEA	0.95	Yes	Exon elongation
ARHGDIA	AspN + LysC	TRANSCRIPT33743.CHR17.NNIC	DYMVGSYSIKSRFT	0.99	Yes	Novel intron
ARHGDIA	AspN + LysC	TRANSCRIPT33743.CHR17.NNIC	TDYMVGSYSIK	1.00	Yes	Novel intron
ATP5PD	Chymotrypsin	TRANSCRIPT29328.CHR17.NIC	TAQVDAEEKEDVSS	1.00	No	Intron retention
CAPG	Trypsin	TRANSCRIPT16133.CHR2.NIC	MQYAPNTQVRR	0.81	No	Intron retention
EPB41L2	Chymotrypsin	TRANSCRIPT22463.CHR6.NNIC	SHLLESSHETL	0.99	Yes	Novel exon
TPM3	Trypsin	TRANSCRIPT38168.CHR1.NNIC	AISEELDHALNDMTSIASLQPT	1.0	Yes	Novel combination of splice sites
VTI1A	Trypsin	TRANSCRIPT20975.CHR10.NIC	NELLGDDGNSSENQLIK	0.99	No[b]	Novel exon

[a]PeptideProphet confidence score.

[b]The peptide is still included in the analysis because it was previously identified in an IsoSeq study by Leung et al., (2021).

difference between TRANSCRIPT529.CHR1.NNIC and TRANSCRIPT531.CHR1.NNIC is that TRANSCRIPT531.CHR1.NNIC has a longer exon 3. TRANSCRIPT556.CHR1.NNIC contains partial exon skipping of exon 3 and intron retention that results in a frameshift and a preliminary stop codon. TRANSCRIPT556.CHR1.NNIC and TRANSCRIPT529.CHR1.NNIC were also identified in the ONT data.

Figure 5B shows two novel *SLC24A1* transcripts, TRANSCRIPT10713.CHR15.NIC and TRANSCRIPT10694.CHR15.NIC, which both account for 50 percent of the *SLC24A1* transcripts. TRANSCRIPT10713.CHR15.NIC results in a novel ORF while TRANSCRIPT10694.CHR15.NIC corresponds to a known ORF. The novel TRANSCRIPT10713.CHR15.NIC isoform contains an alternative C-terminus with a penultimate coding exon of (33 aa) that is also present in the short reference transcript ENST00000505666.2. The novel ORF was not identified in the IsoQuant ONT dataset, but could be confirmed after processing the ONT datasets with StringTie2 (Supplementary Figure S2A).

The third example is a novel *IMPDH1* isoform (TRANSCRIPT23663.CHR7.NIC) that accounts for about 80% of all *IMPDH1* transcripts present in the retina samples (Figure 5C). This isoform is a NIC transcript that combines known splice junctions. The novel part is the inclusion of a small 17-nt exon before the last exon, which results in a reading frameshift and ORF elongation with an alternative stop codon. The same 17-nt exon was also identified in the mouse retina (Wang et al., 2024) and the StringTie2 ONT dataset (Supplementary Figure S2B). This additional exon and alternative stop codon are present in the short reference isoform ENST00000648462.1, but this isoform has an alternative 5’ start, while the novel retinal isoform uses the reference start codon. There are also two peptides, DAPQCPLLGTASLHN and GGGGDAPQCPLLGTASLHN, that map to this alternative last exon (Figure 5D; Supplementary Figure S3). These overlapping peptides were not called with our novelty analysis because they also map to a reference isoform. However, our transcript analysis shows that the novel isoform is likely to be the most common isoform for *IMPDH1* in the human neural retina, and that this transcript is translated into protein. Therefore, the additional coding sequence should be considered when looking for variants in patients with *IMPDH1*-associated vision loss.

Overall, our study created an important reference dataset for future retina and IRD research. All datasets are available in the corresponding repositories and are also available as genome browser tracks.

4 Discussion

The comprehensive proteogenomic atlas generated in our study, combining PacBio long-read RNA-sequencing, MS, and whole-genome sequencing, provides a rich dataset for advancing our understanding of retina-specific transcript- and protein isoforms. We provided several examples of novel protein isoforms supported by MS data, and for many more novel transcripts that are potentially coding for novel protein isoforms. These novel protein isoforms may demonstrate (novel) functionalities different from the reference protein. Moreover, also novel transcripts not validated with mass-spectrometry can be relevant for IRD-research and genetic diagnosis; for example, identified (rare) splice junctions can be predictive of alternative splicing induced by a pathogenic variant. Furthermore, PacBio long-read RNA sequencing adds a unique layer

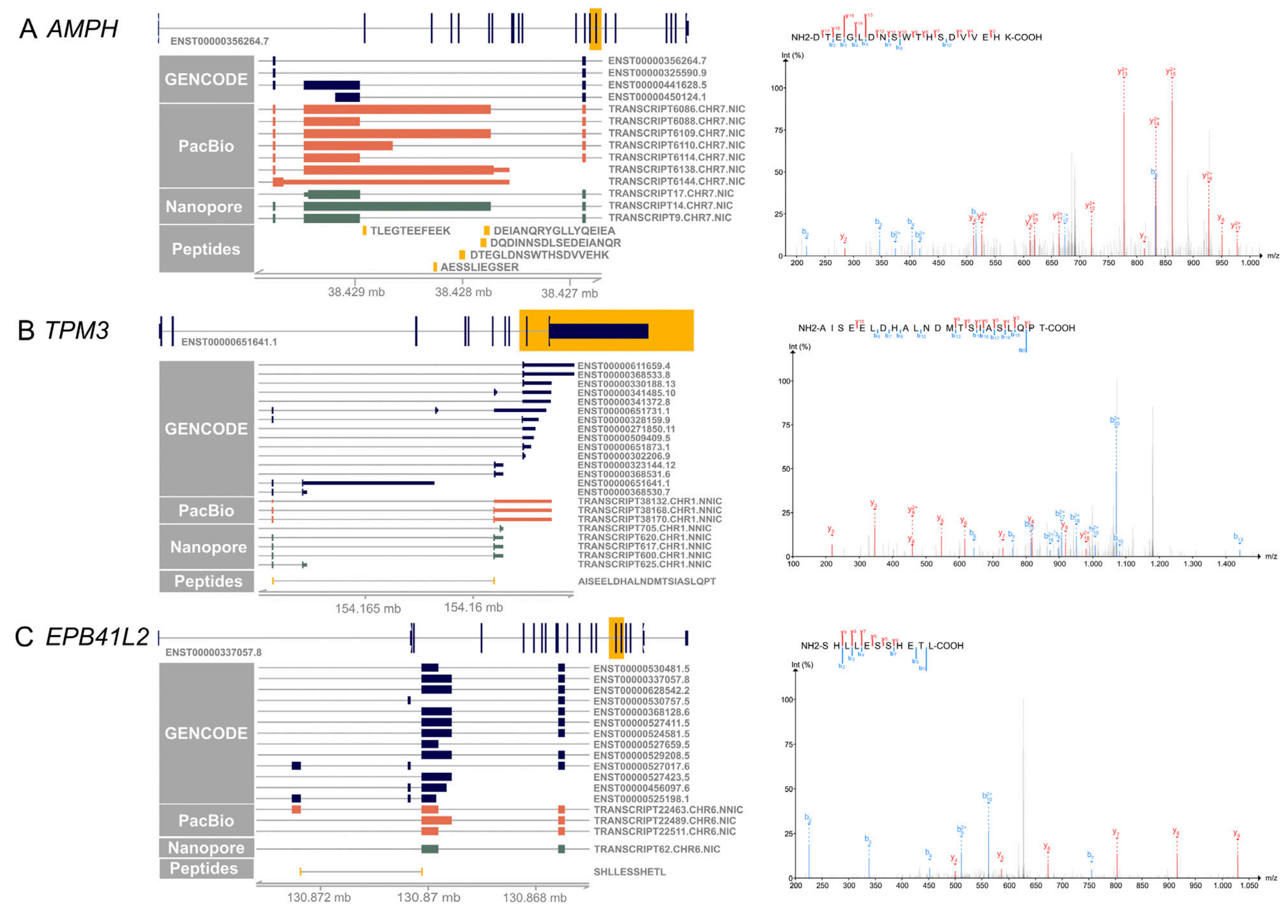

FIGURE 4
Novel peptides confirm novel IsoQuant isoforms. On the left side, the canonical transcript of each gene is shown with the yellow highlighted region shown in more detail. GENCODE transcripts are shown in blue, novel IsoQuant transcripts in orange, novel Oxford Nanopore Technology (ONT) sequencing transcripts in green, and novel peptides in yellow. Transcripts were filtered for transcripts with at least one exon in the displayed region. On the right, the spectrum of the novel peptide is shown with b-ions in blue and y-ions in red. **(A)** Five novel peptides confirm an intron retention event in *AMPH* that is also supported by the ONT data. The spectrum of one novel peptide is shown, the other four spectra can be found in Supplementary Data Sheet 4 **(B)** The novel splice peptide in *TPM3* maps to two novel transcripts that contain a novel combination of known splice junctions. **(C)** A novel splice junction peptide supports TRANSCRIPT22463. CHR6. NNIC derived from the *EPB41L2* gene.

of refinement on retina-specific transcript isoforms, providing information on *in cis* or *trans* occurrence of events. Our dataset does not only identify novel isoforms, but it also confirms known retina isoforms that previously have been assembled from short read RNA-sequencing data. Together, these insights can also provide a valuable resource for therapeutic developments in the context of IRDs.

For *AMPH*, a protein associated with the cytoplasmic surface of synaptic vesicles with enriched expression in the retina and the brain, we observed different amino acid insertions downstream of the BAR domain. We could not identify protein domains in these additional sequences. AlphaFold (Jumper et al., 2021) predicts that the proteins with the insertions have similar 3D structures as the reference protein H0Y7T8 (Supplementary Figure S4). The inserted amino acids are located in the disordered part of the protein that has a low Local Distance Difference Test score. Therefore, further research is necessary to determine if the new isoforms generate stable proteins with different functionalities than the reference protein. Other novel retina isoforms include *TPM3*, *EPB41L2*, *SAMD11*, *SLC24A1* and *IMPDH1*. The novel *TPM3* isoform contains a 79 nt exon inclusion that results in a frameshift and a preliminary stop codon. The alternative exon and corresponding frameshift do not affect any known protein domains. *EPB41L2* was previously found to interact with cyclic-nucleotide gated channels in the rod outer segments (Cheng and Molday, 2013). The novel isoform accounts for only 4 percent of the *EPB41L2* transcripts, however, it is still detected on protein level. This shows that even low abundance isoforms can be translated into proteins and likely impact the function of the protein. The sequence encoded by the additional exon does not fall into a known protein domain. The canonical transcript ENST00000337057.8 only has a FL count of two in our dataset. The difference between this isoform and all other isoforms detected in the retina samples is that it contains a Spectrin and Actin Binding (SAB) domain. This shows that in the retina mainly EPB41L2 transcripts without SAB domain are expressed, which is also observed in the brain (GTEx). Mutations

TABLE 3 RetNet genes with a novel most abundant isoform.

Transcript	Gene	Length	Structural category	Relative abundance of the isoform	FL count	Subcategory[a]
TRANSCRIPT4695.CHR10.NIC	*ACBD5*	5,261	NIC	0.33	158	Alternative poly(A) site
TRANSCRIPT16166.CHR3.NIC	*ATXN7*	2,535	NIC	0.31	11	-
TRANSCRIPT16650.CHR4.NIC	*BBS7*	2,638	NIC	0.54	87	Extra intron, alternative poly(A) site
TRANSCRIPT12294.CHR19.NIC	*C3*	3,880	NIC	0.88	59	Intron retention, alternative poly(A) site
TRANSCRIPT3755.CHR4.NNIC	*CC2D2A*	5,411	NNIC	0.17	94	Alternative structure, alternative poly(A) site
TRANSCRIPT18031.CHR2.NIC	*CNNM4*	3,830	NIC	0.39	255	Alternative poly(A) site
TRANSCRIPT47181.CHR1.NIC	*CRB1*	5,341	NIC	0.30	187	Intron retention, exon elongation, alternative poly(A) site
TRANSCRIPT8849.CHR1.NNIC	*EMC1*	4,913	NNIC	1.00	267	Extra intron, intron shift, exon elongation
TRANSCRIPT3910.CHR1.NIC	*ESPN*	6,326	NIC	1.00	565	Extra intron, exon elongation, alternative poly(A) site
TRANSCRIPT9878.CHR13.NNIC	*GRK1*	7,091	NNIC	0.65	680	Intron shift, exon elongation, alternative poly(A) site
TRANSCRIPT24253.CHR5.NNIC	*GRM6*	6,192	NNIC	0.77	128	Extra intron, alternative poly(A) site
TRANSCRIPT23663.CHR7.NIC	*IMPDH1*	2,866	NIC	0.79	867	Intron shift
TRANSCRIPT11816.CHR9.NIC	*INVS*	7,086	NIC	0.58	90	Intron retention, exon elongation, alternative poly(A) site
TRANSCRIPT20055.CHR11.NIC	*MYO7A*	10,907	NIC	0.72	76	Alternative poly(A) site
TRANSCRIPT3636.CHR1.NIC	*NPHP4*	2,668	NIC	0.39	14	Intron retention
TRANSCRIPT1308.CHRX.NIC	*OFD1*	2,539	NIC	0.40	21	Intron retention, exon elongation, alternative poly(A) site
TRANSCRIPT8654.CHR10.NIC	*PCDH15*	4,743	NIC	0.18	203	Intron retention, exon elongation, alternative poly(A) site
TRANSCRIPT9279.CHR1.NNIC	*PLA2G5*	4,523	NNIC	0.24	42	Alternative structure, intron retention
TRANSCRIPT18772.CHR12.NIC	*POC1B*	3,460	NIC	0.21	39	Intron retention, exon elongation, alternative poly(A) site
TRANSCRIPT18871.CHR6.NNIC	*PRDM13*	5,142	NNIC	0.52	12	Alternative structure, alternative poly(A) site
TRANSCRIPT26925.CHR17.NIC	*RGS9*	1,667	NIC	0.37	47	Extra intron, exon elongation, alternative poly(A) site
TRANSCRIPT14425.CHR16.NIC	*RPGRIP1L*	2,528	NIC	0.23	5	Intron retention, exon elongation, alternative poly(A) site
TRANSCRIPT529.CHR1.NNIC	*SAMD11*	2,382	NNIC	0.62	42	Alternative structure
TRANSCRIPT10713.CHR15.NIC	*SLC24A1*	5,140	NIC	0.50	72	Extra intron, alternative poly(A) site
TRANSCRIPT4995.CHR3.NIC	*SLC4A7*	7,577	NIC	0.51	632	Exon elongation
TRANSCRIPT9410.CHR8.NNIC	*TTPA*	2,466	NNIC	0.44	11	Alternative structure, exon elongation, alternative poly(A) site

(Continued on following page)

TABLE 3 (*Continued*) RetNet genes with a novel most abundant isoform.

Transcript	Gene	Length	Structural category	Relative abundance of the isoform	FL count	Subcategory[a]
TRANSCRIPT11681.CHR2.NNIC	*WDPCP*	7,044	NNIC	0.51	54	Alternative structure, alternative poly(A) site

NNIC, novel not in catalog; NIC, novel in catalog; FL (full-length) count, Transcript count of the full-length isoform.
[a]Only novel subcategories are shown.

in *SAMD11* are associated with adult-onset retinitis pigmentosa (Corton et al., 2016). Within photoreceptor cells, SAMD11 serves as a component of PRC1, required for establishing the proper identity of rod photoreceptors (Kubo et al., 2021). Except for ENST00000618323.5 all *SAMD11* isoforms detected in the retina samples lack the SAND DNA-binding domain. This could mean that in the retina, the majority of SAMD11 proteins do not bind to DNA but instead mainly interacts with other proteins or RNA via its SAM domain. The *SLC24A1* isoform with the novel ORF contains an alternative terminal exon resulting in a shorter sodium/calcium exchanger protein domain compared to most reference isoforms. However, it is only 10 amino acids shorter than the sodium/calcium exchanger protein domain in transcript ENST00000537259.5. Variants in *IMPDH1* are associated with retinitis pigmentosa type 10. The most common novel isoform detected in the retina contains a 17 nt exon before the last exon resulting in a frameshift with an alternative ORF end of 37 amino acids. This alternative sequence does not affect any known protein domains. According to GTEx, the small exon has no high inclusion in any tissue included in that study, suggesting that this proteoform is retina-specific.

Most novel transcripts do not cause big changes in the encoded protein structure. However, this was also expected because the advantage of long-read sequencing is the sequencing of whole isoforms. Therefore, we expected to mainly find transcripts with novel combinations of known junctions. For *SAMD11*, the novel transcripts miss the part coding for the N-terminus of the protein but contain a novel first non-coding exon resulting in a shortened protein. For *EPB41L2*, *SLC24A1,* and *IMPDH1* we observe a fusion of junctions present in prominent reference isoforms with those identified in less prominent, smaller isoforms. For *IMPDH1*, the new combination of splice junctions results in a frameshift and an alternative last exon or ORF end. There is no frameshift in *SLC24A1,* but the novel transcript also contains an alternative terminal exon. Moreover, many novel transcripts contained an alternative polyA site, exon elongation, or intron retention. While intron retention and exon elongation might affect the ORF, an alternative polyA site by itself does not. However, it is not uncommon to find different 3′UTR lengths for the same gene as 70% of human genes have more than one polyadenylation site (Navarro et al., 2021). It is known that the 3’ UTR has regulatory functions as its length can affect translation efficiency, stability of the mRNA, and even tissue-specific expression (Tanguay and Gallie, 1996; Navarro et al., 2021).

Using a conservative approach and manual curation, we detected 12 novel peptides. This number is comparable to the 14 and 30 novel peptides identified by Miller et al. (2022) and Mehlferber et al. (2022) using a similar approach in cultured T- and endothelial cells, respectively. An important difference between our study and the other two is that we used the conservative IsoQuant tool instead of SQANTI3 (Tardaguila et al., 2018).· Our database only contained 12,000 novel ORFs while the other two studies included 35,000 and 27,000 novel ORFs respectively. When we used SQANTI3 or TALON (Wyman et al., 2019) to create the PacBio transcriptome, we also identified more novel transcript isoforms, novel ORFs and novel peptides (Supplementary Table S5). The same was observed for the ONT data where we identified many more isoforms using StringTie2 than IsoQuant. However, we aimed to create an atlas with high-confidence transcript isoforms, and preferred IsoQuant over SQANTI3 and TALON because of the lower level of false positive isoforms (Pardo-Palacios et al., 2023; Prjibelski et al., 2023). Thus, our study used conservative criteria for novel transcript and ORF detection and finds a similar or larger proportion of novel peptides produced from these ORFs compared to the previously mentioned studies.

To increase proteomic coverage, we performed three different digestions, one with trypsin, one with chymotrypsin, and one with a combination of AspN and LysC. As shown before, using multiple enzymes drastically improves proteome coverage (Sinitcyn et al., 2023). A recent study by Sinitcyn et al. (2023) applied deep proteome sequencing to cover a large fraction of the human proteome by MS-based proteomics. Apart from using different digestion enzymes, they applied heavy fractionation and different collision modes, where we used a less accessible biomaterial, only one liquid chromatography fractionation scheme, and one collision mode. Another factor that may have reduced our proteomic coverage in comparison to their study is that many proteins relevant to retina-specific functions and IRDs are membrane proteins that are difficult to detect with MS. Moreover, intron retention was one of the most abundant novel events observed in our study and few intron retention events result in stable proteins due to the presence of premature stop codons. Finally, the number of novel peptides that could be detected was low to start with. A lot of novelty lies in the combination of known splice junctions and in UTR variation.

In conclusion, this study creates a comprehensive overview of the transcript and protein isoforms expressed in healthy human neural retina. Moreover, it highlights the need to study tissue-specific transcriptomes in more detail for better understanding of tissue-specific regulation and for finding disease-causing variants. Choosing a representative transcriptome of the tissue of interest affects variant effect prediction and thereby has an influence on providing a genetic diagnosis to patients with an inherited disease (Taylor et al., 2015; Salz et al., 2023). Therefore, we provide our dataset as reference for future retina and IRD research to contribute to missing heritability in IRD patients.

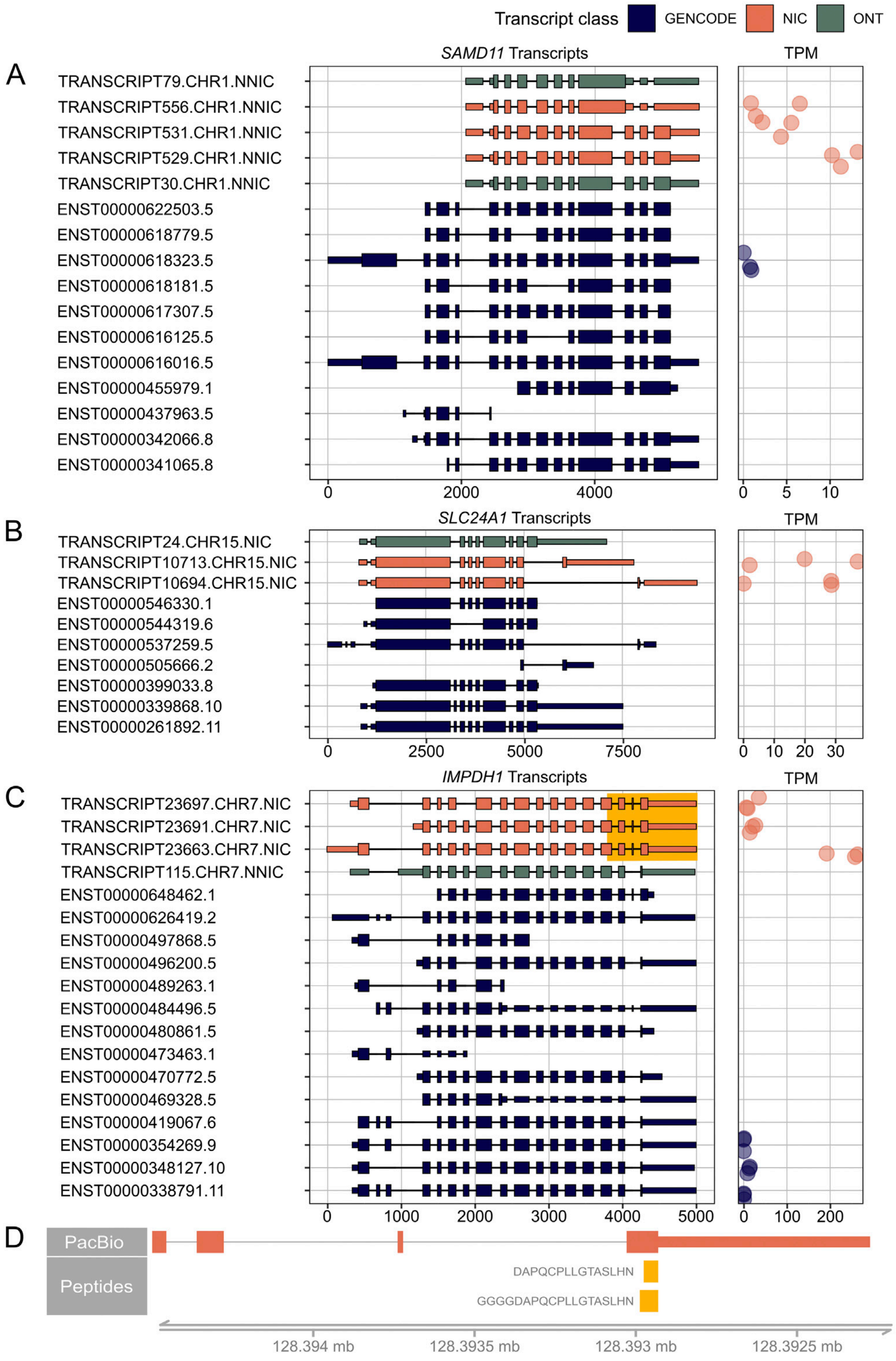

FIGURE 5
RetNet genes demonstrating the highest expression of a novel isoform. On the left, PacBio transcripts are shown in orange, Oxford Nanopore Technology (ONT) sequencing transcripts in green, and reference GENCODE v39 transcripts in blue. We only show PacBio and ONT transcripts that result in a novel open reading frame. For all transcripts, the 5′-end is shown on the left and the 3′-end on the right. On the right, the Transcripts Per Million (TPM) count in the three individual samples is shown. **(A)** *SAMD11* transcripts and their corresponding TPM. **(B)** *SLC24A1* transcripts and their corresponding TPM **(C)** *IMPDH1* transcripts and their corresponding TPM. The highlighted yellow part is shown in **(D)** with peptides that map to the elongated last exon. The spectra of the peptides are shown in Supplementary Figure S3.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: http://www.proteomexchange.org/, PXD045187; https://ega-archive.org, EGAD50000000101. USCS genome browser tracks of the analyzed transcriptomics and proteomics data are available at https://genome-euro.ucsc.edu/s/tabeariepe/retina_atlas. All original code has been deposited at GitHub (https://github.com/cmbi/Neural-Retina-Atlas) and is publicly available as of the date of publication.

Ethics statement

Ethical approval was not required for the studies involving humans because the eyes were donated after a signed consent has been obtained from the donor's next of kin for cornea transplantation. According to the Italian law 91/99 and the guidelines from the Italian National Transplant Service, tissues unsuitable for transplantation can be used for research purposes if the aim of the research project is to improve transplantation or increase the knowledge in the field. The studies were conducted in accordance with the local legislation and institutional requirements. The human samples used in this study were acquired from another research group. Written informed consent to participate in this study was not required from the participants or the participants' legal guardians/next of kin in accordance with the national legislation and the institutional requirements.

Author contributions

TR: Visualization, Writing–original draft, Formal Analysis, Methodology, Project administration. MS: Investigation, Methodology, Writing–review and editing. RS: Formal Analysis, Writing–review and editing. AR: Resources, Writing–review and editing. SB: Formal Analysis, Writing–review and editing. EB: Investigation, Writing–review and editing. TT: Investigation, Writing–review and editing. MK: Investigation, Writing–review and editing. JG: Investigation, Writing–review and editing. HW: Investigation, Writing–review and editing. ED: Resources, Writing–review and editing. EB: Resources, Writing–review and editing. FN: Resources, Writing–review and editing. SK: Resources, Writing–review and editing. BF: Resources, Writing–review and editing. SF: Resources, Writing–review and editing. FC: Resources, Writing–review and editing. FC: Conceptualization, Methodology, Project administration, Supervision, Writing–review and editing. EW: Conceptualization, Writing–review and editing. SR: Conceptualization, Methodology, Project administration, Supervision, Writing–review and editing. EV: Conceptualization, Methodology, Supervision, Writing–review and editing. P't H: Supervision, Writing–review and editing, Conceptualization, Methodology, Project administration.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This work was supported by the Nederlandse Organisatie voor Wetenschappelijk Onderzoek (184.034.019 to the Netherlands X-omics initiative); Ghent University Special Research Fund (BOF20/GOA/023 to EB); H2020 MSCA ITN grant (813490 Star T to EB and FC); Solve-RET (EJPRD19-234 to EB); and StarT (813490 to AR). EB is a Senior Clinical Investigator (1802220N) of the Research Foundation-Flanders (FWO) and a member of ERN-EYE (Framework Partnership Agreement No 739534-ERN-EYE). Funding for open access charge: Netherlands X-omics initiative. The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by the NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The data used for the analyses described in this manuscript were obtained from the GTEx Portal on 12 June 2023 [GTEx Analysis Release v.8 (dbGaP Accession phs000424. v8. p2)].

Acknowledgments

Figure 1A, Figure 3C were created with BioRender.com. We thank the Radboud Technology Center Genomics for their assistance with the PacBio long-read sequencer, and Rob W.J. Collin, Alex Garanto, Alex Hoischen and Lisenka E.L.M. Vissers for valuable advice on long-read RNA sequencing.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2024.1451024/full#supplementary-material

References

Abugessaisa, I., Noguchi, S., Hasegawa, A., Kondo, A., Kawaji, H., Carninci, P., et al. (2019). refTSS: a reference data set for human and mouse transcription start sites. *J. Mol. Biol.* 431, 2407–2422. doi:10.1016/J.JMB.2019.04.045

Aísa-Marín, I., García-Arroyo, R., Mirra, S., and Marfany, G. (2021). The alter retina: alternative splicing of retinal genes in health and disease. *Int. J. Mol. Sci.* 22, 1855. doi:10.3390/ijms22041855

Albert, S., Garanto, A., Sangermano, R., Khan, M., Bax, N. M., Hoyng, C. B., et al. (2018). Identification and rescue of splice defects caused by two neighboring deep-intronic ABCA4 mutations underlying stargardt disease. *Am. J. Hum. Genet.* 102, 517–527. doi:10.1016/j.ajhg.2018.02.008

Bacchi, N., Casarosa, S., and Denti, M. A. (2014). Splicing-correcting therapeutic approaches for retinal dystrophies: where endogenous gene regulation and specificity matter. *Invest Ophthalmol. Vis. Sci.* 55, 3285–3294. doi:10.1167/IOVS.14-14544

Ben-Yosef, T. (2022). Inherited retinal diseases. *Int. J. Mol. Sci.* 23, 13467. doi:10.3390/IJMS232113467

Cao, H., Wu, J., Lam, S., Duan, R., Newnham, C., Molday, R. S., et al. (2011). Temporal and tissue specific regulation of RP-associated splicing factor genes PRPF3, PRPF31 and PRPC8—implications in the pathogenesis of RP. *PLoS One* 6, e15860. doi:10.1371/JOURNAL.PONE.0015860

Chen, X., Schulz-Trieglaff, O., Shaw, R., Barnes, B., Schlesinger, F., Källberg, M., et al. (2016). Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* 32, 1220–1222. doi:10.1093/bioinformatics/btv710

Cheng, C. L., and Molday, R. S. (2013). Interaction of 4.1G and cGMP-gated channels in rod photoreceptor outer segments. *J. Cell. Sci.* 126, 5725–5734. doi:10.1242/jcs.137679

Ciampi, L., Mantica, F., López-Blanch, L., Permanyer, J., Rodriguez-Marín, C., Zang, J., et al. (2022). Specialization of the photoreceptor transcriptome by Srrm3-dependent microexons is required for outer segment maintenance and vision. *Proc. Natl. Acad. Sci. U. S. A.* 119, e2117090119. doi:10.1073/pnas.2117090119

Corton, M., Avila-Fernández, A., Campello, L., Sánchez, M., Benavides, B., López-Molina, M. I., et al. (2016). Identification of the photoreceptor transcriptional Co-Repressor SAMD11 as novel cause of autosomal recessive retinitis pigmentosa. *Sci. Rep.* 6, 35370. doi:10.1038/srep35370

Daiger, S., Rossiter, B., Greenberg, J., Christoffels, A., Hide, W., P Daiger, S., et al. (1998). Data services and software for identifying genes and mutations causing retinal degeneration. *Invest Ophthalmol. Vis. Sci.* 39.

David, J. K., Maden, S. K., Wood, M. A., Thompson, R. F., and Nellore, A. (2022). Retained introns in long RNA-seq reads are not reliably detected in sample-matched short reads. *Genome Biol.* 23, 240. doi:10.1186/s13059-022-02789-6

de Bruijn, S. E., Rodenburg, K., Corominas, J., Ben-Yosef, T., Reurink, J., Kremer, H., et al. (2023). Optical genome mapping and revisiting short-read genome sequencing data reveal previously overlooked structural variants disrupting retinal disease–associated genes. *Genet. Med.* 25, 100345. doi:10.1016/J.GIM.2022.11.013

Den Hollander, A. I., Koenekoop, R. K., Yzer, S., Lopez, I., Arends, M. L., Voesenek, K. E. J., et al. (2006). Mutations in the CEP290 (NPHP6) gene are a frequent cause of Leber congenital amaurosis. *Am. J. Hum. Genet.* 79, 556–561. doi:10.1086/507318

Farkas, M. H., Grant, G. R., White, J. A., Sousa, M. E., Consugar, M. B., and Pierce, E. A. (2013). Transcriptome analyses of the human retina identify unprecedented transcript diversity and 3.5 Mb of novel transcribed sequence via significant alternative splicing and novel genes. *BMC Genomics* 14, 486. doi:10.1186/1471-2164-14-486

Fokkema, I. F. A. C., Taschner, P. E. M., Schaafsma, G. C. P., Celli, J., Laros, J. F. J., and den Dunnen, J. T. (2011). LOVD v.2.0: the next generation in gene variant databases. *Hum. Mutat.* 32, 557–563. doi:10.1002/HUMU.21438

Frankish, A., Diekhans, M., Ferreira, A. M., Johnson, R., Jungreis, I., Loveland, J., et al. (2019). GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* 47, D766–D773. doi:10.1093/NAR/GKY955

Goloborodko, A. A., Levitsky, L. I., Ivanov, M. V., and Gorshkov, M. V. (2013). Pyteomics - a python framework for exploratory data analysis and rapid software prototyping in proteomics. *J. Am. Soc. Mass Spectrom.* 24, 301–304. doi:10.1007/s13361-012-0516-6

Hanany, M., Allon, G., Kimchi, A., Blumenfeld, A., Newman, H., Pras, E., et al. (2018). Carrier frequency analysis of mutations causing autosomal-recessive-inherited retinal diseases in the Israeli population. *Eur. J. Hum. Genet.* 26, 1159–1166. doi:10.1038/S41431-018-0152-0

Jayasinghe, R. G., Cao, S., Gao, Q., Wendl, M. C., Vo, N. S., Reynolds, S. M., et al. (2018). Systematic analysis of splice-site-creating mutations in cancer. *Cell. Rep.* 23, 270–281.e3. doi:10.1016/j.celrep.2018.03.052

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589. doi:10.1038/S41586-021-03819-2

Khan, M., Cornelis, S. S., Pozo-Valero, M. D., Whelan, L., Runhart, E. H., Mishra, K., et al. (2020). Resolving the dark matter of ABCA4 for 1054 Stargardt disease probands through integrated genomics and transcriptomics. *Genet. Med.* 22, 1235–1246. doi:10.1038/s41436-020-0787-4

Kubo, S., Yamamoto, H., Kajimura, N., Omori, Y., Maeda, Y., Chaya, T., et al. (2021). Functional analysis of Samd11, a retinal photoreceptor PRC1 component, in establishing rod photoreceptor identity. *Sci. Rep.* 11, 4180. doi:10.1038/s41598-021-83781-1

Landrum, M. J., Lee, J. M., Benson, M., Brown, G. R., Chao, C., Chitipiralla, S., et al. (2018). ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* 46, D1062–D1067. doi:10.1093/NAR/GKX1153

Leung, S. K., Jeffries, A. R., Castanho, I., Jordan, B. T., Moore, K., Davies, J. P., et al. (2021). Full-length transcript sequencing of human and mouse cerebral cortex identifies widespread isoform diversity and alternative splicing. *Cell. Rep.* 37, 110022. doi:10.1016/j.celrep.2021.110022

Levitsky, L. I., Klein, J. A., Ivanov, M. V., and Gorshkov, M. V. (2018). Pyteomics 4.0: five years of development of a Python proteomics framework. *J. Proteome Res.* 18, 709–714. doi:10.1021/acs.jproteome.8b00717

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760. doi:10.1093/bioinformatics/btp324

Li, K., Vaudel, M., Zhang, B., Ren, Y., and Wen, B. (2019). PDV: an integrative proteomics data viewer. *Bioinformatics* 35, 1249–1251. doi:10.1093/BIOINFORMATICS/BTY770

Ling, J. P., Wilks, C., Charles, R., Leavey, P. J., Ghosh, D., Jiang, L., et al. (2020). ASCOT identifies key regulators of neuronal subtype-specific splicing. *Nat. Commun.* 11 (1), 137. doi:10.1038/s41467-019-14020-5

Liu, M. M., and Zack, D. J. (2013). Alternative splicing and retinal degeneration. *Clin. Genet.* 84, 142–149. doi:10.1111/CGE.12181

Mehlferber, M. M., Jeffery, E. D., Saquing, J., Jordan, B. T., Sheynkman, L., Murali, M., et al. (2022). Characterization of protein isoform diversity in human umbilical vein endothelial cells via long-read proteogenomics. *RNA Biol.* 19, 1228–1243. doi:10.1080/15476286.2022.2141938

Miller, R. M., Jordan, B. T., Mehlferber, M. M., Jeffery, E. D., Chatzipantsiou, C., Kaur, S., et al. (2022). Enhanced protein isoform characterization through long-read proteogenomics. *Genome Biol.* 23, 69. doi:10.1186/s13059-022-02624-y

Murphy, D., Cieply, B., Carstens, R., Ramamurthy, V., and Stoilov, P. (2016). The musashi 1 controls the splicing of photoreceptor-specific exons in the vertebrate retina. *PLoS Genet.* 12, e1006256. doi:10.1371/journal.pgen.1006256

Murphy, D., Singh, R., Kolandaivelu, S., Ramamurthy, V., and Stoilov, P. (2015). Alternative splicing shapes the phenotype of a mutation in BBS8 to cause nonsyndromic retinitis pigmentosa. *Mol. Cell. Biol.* 35, 1860–1870. doi:10.1128/MCB.00040-15

Navarro, E., Mallén, A., and Hueso, M. (2021). Dynamic variations of 3'UTR length reprogram the mRNA regulatory landscape. *Biomedicines* 9, 1560. doi:10.3390/BIOMEDICINES9111560

Niyadurupola, N., Sidaway, P., Osborne, A., Broadway, D. C., and Sanderson, J. (2011). The development of human organotypic retinal cultures (HORCs) to study retinal neurodegeneration. *Br. J. Ophthalmol.* 95, 720–726. doi:10.1136/bjo.2010.181404

Osborne, A., Hopes, M., Wright, P., Broadway, D. C., and Sanderson, J. (2016). Human organotypic retinal cultures (HORCs) as a chronic experimental model for investigation of retinal ganglion cell degeneration. *Exp. Eye Res.* 143, 28–38. doi:10.1016/j.exer.2015.09.012

Pardo-Palacios, F. J., Wang, D., Reese, F., Diekhans, M., Carbonell-Sala, S., Williams, B., et al. (2023). Systematic assessment of long-read RNA-seq methods for transcript identification and quantification. *bioRxiv* 21, 1349–1363. doi:10.1101/2023.07.25.550582

Parfitt, D. A., Lane, A., Ramsden, C. M., Carr, A. J. F., Munro, P. M., Jovanovic, K., et al. (2016). Identification and correction of mechanisms underlying inherited blindness in human iPSC-derived optic cups. *Cell. Stem Cell.* 18, 769–781. doi:10.1016/J.STEM.2016.03.021

Pinelli, M., Carissimo, A., Cutillo, L., Lai, C. H., Mutarelli, M., Moretti, M. N., et al. (2016). An atlas of gene expression and gene co-regulation in the human retina. *Nucleic Acids Res.* 44, 5773–5784. doi:10.1093/NAR/GKW486

Prjibelski, A. D., Mikheenko, A., Joglekar, A., Smetanin, A., Jarroux, J., Lapidus, A. L., et al. (2023). Accurate isoform discovery with IsoQuant using long reads. *Nat. Biotechnol.* 41, 915–918. doi:10.1038/s41587-022-01565-y

Ratnapriya, R., Sosina, O. A., Starostik, M. R., Kwicklis, M., Kapphahn, R. J., Fritsche, L. G., et al. (2019). Retinal transcriptome and eQTL analyses identify genes associated with age-related macular degeneration. *Nat. Genet.* 51, 606–610. doi:10.1038/S41588-019-0351-9

Ray, T. A., Cochran, K., Kozlowski, C., Wang, J., Alexander, G., Cady, M. A., et al. (2020). Comprehensive identification of mRNA isoforms reveals the diversity of neural cell-surface molecules with roles in retinal development and disease. *Nat. Commun.* 11, 3328. doi:10.1038/s41467-020-17009-7

Riazuddin, S. A., Iqbal, M., Wang, Y., Masuda, T., Chen, Y., Bowne, S., et al. (2010). A splice-site mutation in a retina-specific exon of BBS8 causes nonsyndromic retinitis pigmentosa. *Am. J. Hum. Genet.* 86, 805–812. doi:10.1016/J.AJHG.2010.04.001

Roller, E., Ivakhno, S., Lee, S., Royce, T., and Tanner, S. (2016). Canvas: versatile and scalable detection of copy number variants. *Bioinformatics* 32, 2375–2377. doi:10.1093/bioinformatics/btw163

Ruiz-Ceja, K. A., Capasso, D., Pinelli, M., Del Prete, E., Carrella, D., di Bernardo, D., et al. (2023). Definition of the transcriptional units of inherited retinal disease genes by meta-analysis of human retinal transcriptome data. *BMC Genomics* 24, 206. doi:10.1186/S12864-023-09300-W

Salz, R., Saraiva-Agostinho, N., Vorsteveld, E., van der Made, C. I., Kersten, S., Stemerdink, M., et al. (2023). SUsPECT: a pipeline for variant effect prediction based on custom long-read transcriptomes for improved clinical variant annotation. *BMC Genomics* 24, 305. doi:10.1186/S12864-023-09391-5

Sarantopoulou, D., Brooks, T. G., Nayak, S., Mrčela, A., Lahens, N. F., and Grant, G. R. (2021). Comparative evaluation of full-length isoform quantification from RNA-Seq. *BMC Bioinforma.* 22, 266. doi:10.1186/s12859-021-04198-1

Schumacker, S. T., Coppage, K. R., and Enke, R. A. (2020). RNA sequencing analysis of the human retina and associated ocular tissues. *Sci. Data* 7, 199. doi:10.1038/s41597-020-0541-4

Sinitcyn, P., Richards, A. L., Weatheritt, R. J., Brademan, D. R., Marx, H., Shishkova, E., et al. (2023). Global detection of human variants and isoforms by deep proteome sequencing. *Nat. Biotechnol.* 2023, 1776–1786. doi:10.1038/s41587-023-01714-x

Smith, L. M., and Kelleher, N. L. (2013). Proteoform: a single term describing protein complexity. *Nat. Methods* 10, 186–187. doi:10.1038/nmeth.2369

Swamy, V. S., Fufa, T. D., Hufnagel, R. B., and McGaughey, D. M. (2020). A long read optimized *de novo* transcriptome pipeline reveals novel ocular developmentally regulated gene isoforms and disease targets. *bioRxiv*. 2020.08.21.261644. doi:10.1101/2020.08.21.261644

Tanguay, R. L., and Gallie, D. R. (1996). Translational efficiency is regulated by the length of the 3' untranslated region. *Mol. Cell. Biol.* 16, 146–156. doi:10.1128/MCB.16.1.146

Tardaguila, M., De La Fuente, L., Marti, C., Pereira, C., Pardo-Palacios, F. J., Del Risco, H., et al. (2018). SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. *Genome Res.* 28, 396–411. doi:10.1101/gr.222976.117

Taylor, J. C., Martin, H. C., Lise, S., Broxholme, J., Cazier, J. B., Rimmer, A., et al. (2015). Factors influencing success of clinical genome sequencing across a broad spectrum of disorders. *Nat. Genet.* 47, 717–726. doi:10.1038/ng.3304

Wang, M., Li, Y., Wang, J., Oh, S. H., and Chen, R. (2024). Integrating short-read and long-read single-cell RNA sequencing for comprehensive transcriptome profiling in mouse retina. *bioRxiv* 20, 581234. doi:10.1101/2024.02.20.581234

Wyman, D., Balderrama-Gutierrez, G., Reese, F., Jiang, S., Rahmanian, S., Zeng, W., et al. (2019). A technology-agnostic long-read analysis pipeline for transcriptome discovery and quantification. *bioRxiv* 672931. doi:10.1101/672931