Deciphering the largest disease-associated transcript isoforms in the human neural retina with advanced long-read sequencing approaches

Running title: **sequencing the largest retinal disease transcripts**

Merel Stemerdink[1, 2], Tabea Riepe[3, 4], Nick Zomer [4], Renee Salz [3], Michael Kwint [4], Jaap Oostrik[1], Raoul Timmermans [4], Barbara Ferrari [5], Stefano Ferrari [5], Alfredo Dueñas Rey[6, 7], Emma Delanote[6, 7], Suzanne E. de Bruijn [1, 4], Hannie Kremer[1, 2, 4], Susanne Roosing [4], Frauke Coppieters[6, 7, 8], Alexander Hoischen [4, 9], Frans P. M. Cremers [4], Peter A.C. 't Hoen[3], Erwin van Wijk[1*], Erik de Vrieze[1*$]

[*] These authors contributed equally to this work

[$] Corresponding author: erik.devrieze@radboudumc.nl

[1] Department of Otorhinolaryngology, Radboud University Medical Center, Nijmegen, 6525 GA, The Netherlands.

[2] Donders Institute for Brain, Cognition and Behaviour, Radboud University Medical Center, Nijmegen, 6525 GA, The Netherlands.

[3] Department of Medical BioSciences, Radboud University Medical Center, Nijmegen, 6525 GA, The Netherlands.

[4] Department of Human Genetics, Radboud University Medical Center, Nijmegen, 6525 GA, The Netherlands.

[5] Fondazione Banca degli Occhi del Veneto, Zelarino – Venice, 30174, Italy.

[6] Center for Medical Genetics, Ghent University Hospital, Ghent, 9000, Belgium.

[7] Department of Biomolecular Medicine, Ghent University, Ghent, 9000, Belgium.

[8] Department of Pharmaceutics, Ghent University, Ghent, 9000, Belgium.

[9] Department of Internal Medicine and Radboud Center for Infectious Diseases (RCI), Radboud University Medical Center, Nijmegen, 6525 GA, The Netherlands.

ABSTRACT

Sequencing technologies have long limited the comprehensive investigation of large transcripts associated with inherited retinal diseases (IRDs) like Usher syndrome, which involves 11 associated genes with transcripts up to 19.6 kb. To address this, we used PacBio long-read mRNA isoform sequencing (Iso-Seq) following standard library preparation and an optimized workflow to enrich for long transcripts in the human neural retina. While our workflow achieved sequencing of transcripts up to 15 kb, this was insufficient for Usher syndrome-associated genes *USH2A* and *ADGRV1*, with transcripts of 18.9 kb and 19.6 kb, respectively. To overcome this, we employed the Samplix Xdrop System for indirect target enrichment of cDNA, a technique typically used for genomic DNA capture. This method facilitated the successful capture and sequencing of *ADGRV1* transcripts as well as full-length 18.9 kb *USH2A* transcripts. By combining algorithmic analysis with detailed manual curation of sequenced reads, we identified novel isoforms characterized by an alternative 5' transcription start site, the inclusion of previously unannotated exons or alternative splicing events across the 11 Usher syndrome-associated genes. These findings have significant implications for genetic diagnostics and therapeutic development. The analysis applied here on Usher syndrome-associated transcripts exemplifies a valuable approach that can be extended to explore the transcriptomic complexity of other IRD-associated genes in the complete transcriptome dataset generated within this study. Additionally, we demonstrated the adaptability of the Samplix Xdrop system for capturing cDNA, and the optimized methodologies described can be expanded to facilitate the enrichment of large transcripts from various tissues of interest.

INTRODUCTION

The human retina is a complex multicellular tissue that plays a crucial role in visual function by converting light into the electrical signals that are interpreted by the brain. Understanding the molecular composition of the human retina, particularly the wide variety of transcript isoforms expressed there, is essential for comprehending disease mechanisms and designing effective treatment strategies for inherited retinal diseases (IRDs) (Braun *et al.*, 2013). A key factor contributing to the diversity of transcript isoforms expressed in the retina is alternative splicing, a biological process that allows a single gene to encode multiple transcript isoforms leading to tissue-specific differences in gene expression and function. This process involves the use of alternative transcription initiation and termination sites, intron retention, exon skipping, and alternative splice donor and acceptor sites. The retina is a highly specialized tissue that is known to be enriched for these tissue-specific alternative splicing events (Cao *et al.*, 2011; Liu and Zack, 2013). While previous studies using RNA short-read sequencing provided valuable insights, they often fall short in fully characterizing the diverse array of transcript isoforms present in the human retina (Ciampi *et al.*, 2022; Murphy *et al.*, 2016; Ruiz-Ceja *et al.*, 2023; Sarantopoulou *et al.*, 2021).

The PacBio long-read mRNA isoform sequencing (Iso-Seq) technology (Wang *et al.*, 2016) provides deeper insights into the transcriptome complexity caused by alternative splicing events. This technology eliminates the need for *de novo* transcript assembly, and consequently provides greater certainty in the identification of alternative splicing events, as it can generate full-length transcripts up to 10 kb in length. While the length of the average human transcript is approximately 2 kb, many transcripts associated with IRDs are considerably longer, and are therefore at (or beyond) the limit of what can be reliably investigated using PacBio Iso-Seq technology. The longest two annotated transcripts of IRD-associated genes are *USH2A* (18.9 kb) and *ADGRV1* (19.6 kb), both associated with Usher syndrome, which is an autosomal recessively inherited disorder characterized by the combination of sensorineural hearing loss and the progressive loss of visual function due to retinitis pigmentosa (RP). The disorder is clinically and genetically diverse, with 11 associated genes

that have been identified (*MYO7A (Weil et al., 1995), USH1C (Verpy et al., 2000), CDH23 (Bolz et al., 2001), PCDH15 (Ahmed et al., 2001), SANS (Weil et al., 2003), CIB2 (Riazuddin et al., 2012), USH2A (Eudy et al., 1998), ADGRV1 (Weston et al., 2004), WHRN (Ebermann et al., 2007), CLRN1 (Adato et al., 2002)* and *ARSG (Abad-Morales et al., 2020)*). With the exception of *CIB2*, the full-length isoforms of the Usher syndrome-associated genes surpass the average human transcript length of 2 kb, with 5 of them exceeding the mean transcript length of RETNET genes (4.7 kb). Although the association of *CIB2* with Usher syndrome type IJ (USH1J) has been called into question (Booth *et al.*, 2018), functional studies leave room for the involvement of *CIB2* in Usher syndrome (Linnert *et al.*, 2023; Sethna *et al.*, 2021). Therefore, *CIB2* is included in our analyses as these results could add to the discussion whether it qualifies as an Usher syndrome-associated gene.

Obtaining a comprehensive understanding of the Usher syndrome-associated transcript isoforms in the human retina is crucial, as it facilitates the development of therapeutic strategies and enables accurate classification of genetic variants linked to the disorder. To this end, we aimed to provide an overview of the Usher syndrome-associated transcript isoforms in the human neural retina, utilizing the PacBio long-read mRNA Iso-Seq dataset from our previous study (Riepe *et al.*, 2024), and supplementing it with two additional datasets aimed at capturing full-length reads from the longest known transcript isoforms. Our previously generated dataset was obtained using the standard PacBio Iso-Seq workflow, optimized for transcripts centered around 2 kb – the average human transcript length – making it suitable for genome-wide studies as we performed in Riepe *et al.* (2024). However, as nearly all Usher syndrome-associated genes have transcripts exceeding 2 kb, we aimed to generate an additional Iso-Seq dataset using an optimized PacBio long transcript workflow to enable the sequencing of larger transcripts up to > 10 kb. Despite our efforts, this workflow remained insufficient to enrich for the longest *USH2A* (18.9 kb) and *ADGRV1* (19.6 kb) transcript isoforms. In an attempt to further enhance the capture of full-length transcripts for these two genes, we also employed an 'indirect targeted enrichment' approach using the Samplix Xdrop System (Madsen *et al.*, 2020), followed by PacBio long-read sequencing. By integrating the data from

these three sequencing workflows and employing a combined strategy of algorithmic analysis and manual curation, we aimed to obtain a comprehensive overview of the Usher syndrome-associated transcript isoforms present in the human neural retina.

RESULTS

Three human neural retina samples were used for PacBio long-read mRNA Iso-Seq, to gain an overview of the Usher syndrome-associated transcript isoforms expressed in the human retina. By integrating the data from three distinct sequencing workflows and combining algorithmic analysis with manual curation (Figure 1), we obtained a comprehensive overview of the landscape of Usher syndrome-associated transcript isoforms in the human neural retina. Table 1 summarizes the previously annotated Usher syndrome-associated transcripts identified in the human retina by IsoQuant analysis (Prjibelski *et al.*, 2023). Additionally, it highlights frequently observed novel transcript isoforms and events identified by Isoquant, and validated by Oxford Nanopore Technology (ONT) long-read mRNA sequencing of independent retina samples. A detailed overview of minor events observed following the manual curation of sequenced reads using the BAM files in the Integrative Genomics Viewer (IGV) is presented in Supplemental_Table_S1. The findings for *MYO7A* (Figure 3), *WHRN* (Figure 4), *USH2A* (Figure 5) and *ADGRV1* (Figure 6) illustrate the different types of observations in the dataset, such as the identification of novel, previously unidentified isoforms, novel coding exons, or regions sensitive to pseudoexon (PE) inclusions. For the remaining Usher syndrome-associated genes, an overview of the algorithmic output and manual curation results has been generated and presented in Supplemental_Fig_S1 – Supplemental_Fig_S8.

Exploring Usher syndrome-associated transcript isoforms through Iso-Seq PacBio standard- and optimized long transcript workflows.

We previously generated a PacBio long-read mRNA Iso-Seq dataset from three human neural retina samples (dataset 1) for which we conducted a genome-wide integration with proteomic and genomic data to construct the proteogenomic atlas presented in Riepe *et al.* (2024). Transcripts identified in this dataset had an average read length of 2.6 kb. Because this is shorter than most Usher syndrome-associated genes we optimized the PacBio long transcript workflow to enrich for larger transcript sizes (dataset 2 and 3). Supplemental_Table_S2 illustrates the successful enrichment for long

transcripts as the average subread length was increased while a similar amount of reads was retrieved with both workflows. The optimized PacBio long transcript workflow produced HiFi (CCS3) reads of over 15 kb with an average size of 4.7 kb (Figure 2A). Transcripts for all 11 Usher syndrome-associated genes were identified in the samples prepared according both the PacBio standard and long transcript workflow, and the optimized long transcript workflow yielded an increased number of reads from Usher genes with transcripts that exceed the average GENCODE transcript length of 2.4kb (Figure 2B, Supplemental_Fig_S9). The limited overlap in read length distribution between the datasets underscores the complementary nature of the standard- and long transcript workflows (Figure 2A).

In Riepe *et al.* (2024), the IsoQuant algorithm was utilized to analyze and categorize the PacBio Iso-Seq standard workflow samples, resulting in dataset 1. A second IsoQuant analysis was conducted on the combined reads from the three PacBio standard workflow samples and the retina sample prepared following the optimized long transcript workflow, resulting in the IsoQuant dataset 2. However, the incorporation of the PacBio long transcript workflow sample in the combined IsoQuant analysis negatively impacted the number of identified unique Full Splice Match (FSM) transcripts. The increase in unique Novel In Catalogue (NIC) and Novel Not In Catalogue (NNIC) transcripts suggests that many reads are reallocated in the transcript models predicted by IsoQuant. (Supplemental_Fig_S10). The long transcript workflow results in an enrichment for intron retention events (Figure 2C), which we suspect causes the observed difference between the IsoQuant transcript models. This is exemplified in Supplemental_Fig_S11, showing the IsoQuant transcript models for *MYO7A* based on dataset 1 and 2, of which only those proposed based on dataset 1 were supported by the raw reads. To reliably identify the largest possible transcripts, we reanalyzed the data from the long transcript workflow separately from the standard workflow data (dataset 3). We furthermore lowered the consensus read requirement to 0 (CCS0) and manually curated the reads in IGV.

158

159 **Iso-Seq identifies a novel predominant *MYO7A* transcript isoform with an alternative 5'**
160 **transcription start site**

161 The *MYO7A* gene encodes the unconventional myosin VIIa motor protein, which is expressed in the
162 outer hair cells of the cochlea (Hasson *et al.*, 1995), the retinal pigmented epithelium (RPE), and in
163 the photoreceptor cells (Liu *et al.*, 1997; Udovichenko *et al.*, 2002). Recent literature describes two
164 human *MYO7A* transcript isoforms that are expressed in the RPE and the neural retina (Figure 3A,
165 3B), which differ in the length of exon 35 (Gilmore *et al.*, 2023). IsoQuant analysis confirmed the
166 presence of both isoforms in the human retina (Figure 3A), with the transcript harboring the short
167 form of exon 35 being the most abundant (4.81 ± 4.04 TPM for ENST00000458637.6 vs. 0.88 ± 0.86
168 TPM for ENST00000409709.9) (Figure 3C), consistent with the findings of Gilmore *et al.* (2023).
169 Moreover, IsoQuant also identified novel transcript isoforms (transcript20052.Chr11.nic;
170 transript20055.Chr11.nic) characterized by an alternative 5' transcription start site (TSS) coupled
171 with an extended 3'-untranslated region (UTR), which have not been previously reported in man.
172 Notably, a similar transcript isoform with a comparable TSS has been identified in mice (Li *et al.*,
173 2020). Manual curation of sequenced transcripts using the BAM-files in IGV revealed substantial read
174 support for this alternative TSS across both the standard- and long transcript workflow samples. The
175 use of this alternative TSS was confirmed by ONT long-read mRNA sequencing on three independent
176 retina samples (Supplemental_Fig_S12). Protein domain analysis using SMART (Letunic *et al.*, 2021)
177 (Figure 3D) and structural modeling with AlphaFold2 (Jumper *et al.*, 2021) (Figure 3E) revealed
178 differences between the canonical MYO7A isoform and the protein isoform encoded by transcripts
179 with the alternative TSS (transcript20052.Chr11.nic). The canonical isoform encodes a protein of
180 2215 amino acids, whereas transcript20052.Chr11.nic is predicted to encode a protein of 2245 amino
181 acids, with a larger N-terminal tail and a predicted low complexity region. To quantify the relative
182 expression of these two transcript isoforms, we performed a qPCR using primers specifically
183 targeting the alternative 5' sequence, canonical 5' sequence, and canonical 3' sequence. The results

suggest that transcripts produced from the alternative TSS are the most abundant *MYO7A* transcript isoform in the human neural retina (Figure 3F). Further manual curation of sequencing reads uncovered skipping or inclusion of certain novel exons, which are summarized in Supplemental_Table_S1. These splicing events are incidental occurrences, except for previously reported 5' truncation of exon 35 and the observed retention of introns 30 and 37 that is found in approximately 25% of the sequenced transcripts.

Iso-Seq reveals a novel coding exon and *WHRN* retina specific transcript isoforms.

The existing knowledge of *WHRN* transcript isoforms, encoding whirlin proteins, has predominantly been derived from research on mutant mouse models (Belyantseva *et al.*, 2005; Ebrahim *et al.*, 2016; Mburu *et al.*, 2003). Studies by Mburu *et al.* (2003) and Belyantseva *et al.* (2005) have suggested the presence of human retinal *WHRN* transcript isoforms based on cDNA clones, of which only the full-length isoform (ENST00000362057.4) and a C-terminal isoform (ENST0000674048.1) were verified in the human retina using RT-PCR (van Wijk *et al.*, 2006) (Figure 4A, 4B). IsoQuant analysis confirmed the expression of the full-length reference isoform of the *WHRN* gene in the human retina but did not detect the previously reported C-terminal isoform (ENST0000674048.1). However, the IsoQuant analysis did identify an alternative C-terminal *WHRN* transcript isoform starting with a distinct noncoding exon (ENST00000265134.10), although its expression is relatively low (0.3 ± 0.5 TPM) (Figure 4C). An N-terminal *WHRN* isoform was proposed based on the murine *Whrn* transcripts, but not yet experimentally validated in man (Belyantseva *et al.*, 2005; Mburu *et al.*, 2003). IsoQuant indeed reveals the presence of the N-terminal transcript isoform ENST00000374057.3 in 2 out of the 3 retinal samples.

Multiple IsoQuant transcript isoforms exhibited intron 4 retention, introducing a premature stop codon, and therefore these transcripts are predicted to encode a truncated protein containing only the first two PDZ domains (Figure 4D). Manual curation of sequenced reads also confirmed intron 4 retention in the majority of the sequenced reads. Similarly, intron 4 retention was observed

in ONT long-read mRNA sequencing data obtained from three independent retinal samples. Quantification of intron 4-containing transcripts using qPCR indicates that 70% of the expressed *WHRN* transcripts exhibited retention of intron 4 (Figure 4F). In the absence of a universally present *WHRN* transcript region, we included a primer pair targeting *WHRN* exons 8-9 to amplify a segment present in nearly all isoforms, which we designated as 'total' *WHRN*.

Finally, two novel exons were identified within intron 7: designated exon 7A and 7B. While exon 7A was exclusively observed as a noncoding exon and always co-occurred with exon 7B, exon 7B was additionally detected as an independent protein-coding exon in transcript13724.Chr9.nic. Manual inspection of the sequencing data also revealed the widespread presence of exon 7B in the Iso-Seq transcripts, which was further corroborated in an independent ONT long-read mRNA sequencing dataset. This 33-nucleotide-long exon 7B is not present in the GENCODE reference annotations for the human *WHRN* gene but does resemble a corresponding 33-nucleotide exon found in the mouse *Whrn* consensus sequence. Structural modeling with AlphaFold2 revealed that this novel exon 7B encodes an additional alpha helix of 11 amino acids in the center of the protein (Figure 4E). A comparison of the TPM levels of the *WHRN* reference transcript (5.5 ± 3.5 TPM for ENST00000362057.4) with those of the exon 7B-containing transcript (13.6 ± 8.9 TPM for transcript13724.Chr9.nnic), suggests that the latter may represent the predominant retinal isoform. Quantitative PCR confirms that inclusion of exon 7B occurs in the majority of *WHRN* transcripts (Figure 4F). Because biallelic pathogenic variants in *WHRN* can cause USH2D, we queried WGS data of our in-house cohort of unsolved IRD patients (de Bruijn *et al.*, 2023) for rare variants in exon 7B, but found no candidate pathogenic variants in this exon.

Identification of pseudoexon-prone regions and the successful enrichment of full-length *USH2A* isoform B transcripts using Samplix Xdrop Sort.

The *USH2A* gene gives rise to two distinct transcripts and protein isoforms. Isoform A is the shorter variant, composed of 1546 amino acids encoded by a 21-exon mRNA transcript (Eudy *et al.*, 1998).

Conversely, isoform B represents the larger isoform, composed of 5202 amino acids and encoded by a transcript consisting of 72 exons (Van Wijk *et al.*, 2004) (Figure 5A, 5B). IsoQuant analysis confirmed the presence of isoform A transcripts (Figure 5A). Additionally, our IsoQuant analysis revealed transcripts (transcript51429.Chr1.nnic and transcript51430.Chr1.nnic) that demonstrate significant diversity in the 5' region. This entails both the use of alternative TSS that are slightly upstream from the annotated TSS, and the use of alternative splice sites in exons 1 and 2. The latter leading to the inclusion of additional amino acids at the 5' end of the encoded open reading frame (ORF), while maintaining the same reading frame and sharing the same 3' UTR as isoform A. Notably, this diversity in TSS was also observed in the independent ONT long-read mRNA sequencing dataset.

IsoQuant analysis revealed transcript51485.Chr1.nic, spanning exons 1-15, as the most abundant transcript (15.35 ± 2.85 TPM; Figure 5C). Manual inspection of the BAM files in IGV confirmed substantial read support for this alternative isoform. However, the presence of a 12-adenine stretch at the 3' end of this transcript suggests possible internal priming of the oligo(dT) primer. Additionally, its shorter length relative to known usherin isoforms may have resulted in preferential amplification and loading on the PacBio SMRT-cells, potentially explaining its higher TPM values, therefore raising uncertainty about whether it represents a *bona fide* transcript isoform or an experimental artifact. Additional N-terminal *USH2A* transcripts were identified that spanned different exon ranges, including exons 1-9 (transcript 51552.Chr1.nnic), exons 1-26 (transcript51403.Chr1.nic), and exons 1-3 (transcript51634.Chr1.nic), as well as a transcript covering exons 16-21 (transcript 51591.Chr1.nic) (Figure 5A). The authenticity of these transcripts can be questioned as the PacBio standard workflow returned numerous partial reads, as can be observed in BAM files, and was not able to capture the full-length *USH2A* transcripts encoding isoform B.

Although also unable to capture the full-length *USH2A* transcripts encoding isoform B, the value of the optimized long transcript workflow is clearly demonstrated through the manual examination of the sequenced reads (Figure 5D). While tiling of reads resulting from the PacBio standard workflow samples did not fully cover the complete *USH2A* isoform B transcript, the reads

from the sample prepared following the optimized long transcript workflow together were able to cover all 72 exons of *USH2A*. These data did not reveal any evidence of natural exon skipping (NES); a phenomenon where specific exons are excluded from the mature transcripts in healthy tissues, as observed for the *ABCA4* gene by Tomkiewicz (2024). Furthermore, numerous reads support the presence of isoform A transcripts and the observed variation in the 5' region. While individual reads do not clearly indicate whether this 5' variation is specific to isoform A, we did identify reads extending beyond exon 21 that are linked to the variation in the 5' region, suggesting it may also occur in the larger isoform B transcripts.

Manual curation of sequenced transcripts furthermore revealed sporadic inclusion of intronic sequences, for example the inclusion of 87 bp of intron 20 that was previously identified as pseudoexon 20 (PE20) by Reurink *et al.* (2023). In addition to the previously reported *USH2A* PE8 and PE20, our investigation also revealed the occasional incorporation of other uncharacterized cryptic exons. These cryptic exons are denoted by arrows in the overview provided in Figure 5D, with genomic positions provided in Supplemental_Table_S3. As demonstrated by Reurink *et al.* (2023), pathogenic deep-intronic variants can induce inclusion of PEs harboring an in-frame stop codon across all *USH2A* transcripts. This prompted us to evaluate WGS data from our in-house cohort of unsolved IRD patients (de Bruijn *et al.*, 2023). However, we did not identify any candidate pathogenic deep-intronic variants surrounding the identified sporadic cryptic exons.

Since both the PacBio and ONT long-read sequencing approaches failed to sequence the full-length *USH2A* isoform B transcripts, we employed a targeted enrichment approach on cDNA using the Samplix Xdrop Sort technology in an effort to capture these *USH2A* transcripts. The Samplix Xdrop Sort is a technique designed for single-molecule enrichment of genomic DNA. For the first time, we here demonstrated the utility of this approach to capture cDNA as well. By using three detection sequences targeting either 5' - middle and 3' sites of the longest known transcript encoding usherin isoform B, we were able to enrich for *USH2A* transcripts, as evidenced by the qPCR results (Supplemental_Table_S4). The enriched pool of transcripts was then subjected to HiFi long-

read sequencing and subsequent cDNA analysis. Manual curation of the obtained sequencing reads revealed that the detection sequence targeting exons 30-31 enabled us to capture and sequence full-length transcripts encoding usherin isoform B for the first time. Although cDNA amplification approach associated with this method yields shorter reads, they are tiled along the full-length *USH2A* transcript encoding isoform B with an average coverage of 73 reads for each of the 72 exons. Because all these reads are from a single capture region, and therefore must be connected to this region, this finding provides conclusive evidence for the presence of full-length *USH2A* transcripts encoding isoform B in the human retina (Figure 5D). Similar to the reads obtained with the PacBio and ONT workflows, this capture-based sequencing approach also does not indicate the presence of alternative splicing events such as natural exon skipping.

The 19.6kb *ADGRV1* transcript defies the limits of long-read mRNA sequencing approaches.

With the largest annotated transcript spanning 19.6 kb, *ADGRV1* - previously known as *MASS1, VLGR1*, and *GPR98* - is the largest Usher syndrome-associated gene, and variants in this gene are responsible for Usher syndrome type 2C (Weston *et al.*, 2004). Three distinct human *ADGRV1* transcript isoforms have been previously reported: *VLGR1a*, *VLGR1b*, and *VLGR1c* (Supplemental_Fig_S6). *VLGR1b* represents the transcript encoding the longest isoform, composed of 90 exons and encoding a protein of 6,306 amino acids. Similar challenges to those encountered with the large coding sequence of *USH2A* were observed for *ADGRV1*, which prevented the PacBio standard- and long transcript workflows and ONT long-read mRNA sequencing, from sequencing the *VLGR1b* transcript. Furthermore, the IsoQuant analysis did also not identify any transcript isoforms corresponding to the shorter *VLGR1a* and *VLGR1c* transcripts, despite their smaller sizes not posing a barrier to sequencing (Supplemental_Fig_S6). Manual curation of the sequenced reads revealed that the 5' extension of exon 65, which defines the start of the *VLGR1a* transcript, was present in reads across all sequenced samples. Notably, only the sample prepared following the PacBio long transcript workflow contained reads that perfectly spanned from the 5' extension of exon 65 to exon 90,

demonstrating the presence of the *VLGR1a* transcript in the human retina. Additionally, manual curation of the sequenced transcripts revealed that approximately half of the reads contained the 83 bp truncation at the 3' end of exon 31, as also observed in the N-terminal *VLGR1c* transcript isoform. While no sequencing reads spanned exon 1 to exon 31, the observed utilization of the splice site resulting in the 83 bp truncation of exon 31 suggests the presence of the *VLGR1c* transcripts in the human retina (Figure 6). Further manual inspection of the PacBio Iso-Seq data, and independent ONT long-read mRNA sequencing data revealed that nearly half of the reads contained an in-frame novel exon situated in intron 39, which has been designated as exon 39A (33 bp in size). We again queried the WGS data of our in-house cohort of unsolved IRD patients to identify whether pathogenic variants could be present in exon 39A, but found no candidate pathogenic variants in this exon.

In an effort to capture full-length *ADGRV1* (19.6 kb) transcripts encoding the longest isoform VLGR1b, we employed the targeted enrichment approach using the Samplix Xdrop Sort system, by utilizing three detecting sequences targeting the 5', middle, and 3' regions of *VLGR1b* transcripts. qPCR results confirmed the enrichment of *ADGRV1* transcripts (Supplemental_Table_S4). While we cannot conclude that the detection sequences were able to capture the full-length isoform, manual curation of sequenced reads revealed that the 5' detection sequence (targeting exons 16-17) provided the best coverage, spanning exons 3 to 77 and 80 to 90, with an average coverage of 818 reads covering the sequenced exons. This suggests the existence of the full-length *ADGRV1* transcript isoform encoding *VLGR1b* in the human retina. Although the absence of reads mapping against exons 78-79 in the 5' enriched sample may hint at natural exon skipping, no reads were found that indicated that exon 77 is immediately followed by exon 80. Furthermore, we did not observe any indication for the skipping of exons 78-79 in the data obtained with the PacBio standard- and long transcript workflows. Finally, data obtained with Samplix Xdrop indicate the presence of an even shorter N-terminal transcript isoform that is similar to murine *Vlgr1e* (Figure S6).

DISCUSSION

Given the limitations of current sequencing approaches to characterize the largest transcripts expressed in the human neural retina, we conducted PacBio long-read RNA sequencing following the standard library preparation and an optimized workflow to enrich for long transcripts in the human neural retina. Both the standard- and the optimized workflow revealed several novel findings for Usher syndrome-associated transcripts. However, these workflows were insufficient for sequencing the longest two Usher syndrome-associated genes *USH2A* and *ADGRV1*, with transcripts of 18.9 kb and 19.6 kb, respectively. Therefore, we employed the Samplix Xdrop System for indirect targeted enrichment of these transcripts, which enabled the successful capture and sequencing of *ADGRV1* transcripts as well full-length 18.9 kb *USH2A* transcripts.

Our focus on the identification of Usher syndrome-associated transcripts in the human retina enabled a more detailed examination compared to genome-wide studies that rely heavily on advanced tools and algorithms to manage large-scale data. By narrowing our scope to Usher genes, we were able to sift through novel transcripts with greater precision through a combination of algorithmic and manual analysis. For the algorithmic analysis of sequencing data, we utilized the IsoQuant algorithm, which was shown to have the lowest rate of false positive isoforms compared to alternatives like SQANTI3 and TALON (Pardo-Palacios *et al.*, 2024; Prjibelski *et al.*, 2023). However, integrating the long transcript workflow dataset into a combined IsoQuant analysis with the standard workflow datasets posed challenges in isoform classification. This was likely due to the enrichment of intron retaining transcripts in the long transcript dataset. We therefore reanalyzed the long transcript dataset separately, and combined the IsoQuant analysis with manual curation of sequenced reads to reliably identify the largest possible transcripts.

Our integrated strategy, combining algorithmic and manual analysis, uncovered novel isoforms and alternative splicing events across the 11 Usher syndrome-associated genes, which we highlighted for *MYO7A, WHRN, USH2A*, and *ADGRV1*. For instance, we discovered a novel predominant *MYO7A* transcript isoform. Previous research using a targeted PCR approach had only

identified two retinal *MYO7A* transcript isoforms differing in exon 35 length, with the truncated exon 35 variant being the predominant form (Gilmore *et al.*, 2023). This PCR-based targeted methodology is incapable of identifying isoforms characterized by potentially novel transcription initiation and termination sites. In contrast, the PacBio Iso-Seq allowed us to identify a novel *MYO7A* transcript with an alternative 5' TSS. While Gilmore *et al.* (2023) have suggested that both previously known *MYO7A* transcript isoforms should be considered when designing gene therapies, our findings indicate that the isoforms with a novel 5' TSS were the most predominant in the human retina. Notably, the variation in exon 35 length as described by Gilmore et al. is also found in this novel isoform. While we do not dispute the relevance of this variation for therapy development, it might be even more important to incorporate the novel TSS isoform we identified in the design of retinal gene augmentation therapies. Furthermore, the novel, previously unannotated *MYO7A* 5' TSS should be incorporated into diagnostic screening pipelines as it may harbor pathogenic variants of potential clinical significance.

Additionally, sporadic intron retention events were observed across all 11 Usher syndrome-associated genes, likely indicating ongoing splicing in these transcripts. The high frequency of intron 30 and 37 retention in the *MYO7A* transcripts, and the intron 4 retention in *WHRN* are particularly interesting. While intron splicing does not follow a strict 5'-3' order or depend on intron length (Pandya-Jones and Black, 2009; Singh and Padgett, 2009), Gazzoli *et al.* (2016) demonstrated that inefficient multi-step splicing of introns can lead to the formation of "exon blocks" - groups of consecutive exons that tend to be spliced together as a unit. The observed retention of introns 30 and 37 in the *MYO7A* transcripts may indicate the creation of exon blocks encompassing exons 31-37. The interplay between intron retention and exon blocks may be relevant for the design of antisense oligonucleotide (ASO) therapies as for example proposed for the *DMD* exon block 45-55 to achieve exon skipping with fewer ASO molecules (Gazzoli *et al.*, 2016). Although multi exon skipping does not seem particularly feasible for transcripts encoding myosin motor proteins such as *MYO7A*, our findings illustrate the value of including partially spliced transcripts in the analysis of Iso-Seq

results. With several ASO-based splicing modulation therapies under development for IRDs, some also based on multiple exon skipping (Schellens *et al.*, 2023), our dataset could help uncover exon blocks in other retinal disease associated genes.

Another notable instance of intron retention was observed in *WHRN* transcripts, where nearly half of the sequenced transcripts exhibited retention of intron 4. While intron retention is generally thought to disrupt protein production due to the introduction of premature stop codons that trigger nonsense-mediated decay (NMD), Boutz *et al.* (2015) described a class of intron-retaining transcripts that are polyadenylated, sequestered in the nucleus, and resistant to NMD. These retained introns may undergo subsequent splicing and export to the cytoplasm for translation, potentially enabling cells to rapidly adapt to environmental changes or stress. Although we cannot confirm the nuclear origin of *WHRN* intron 4 retaining transcripts in our dataset, their amplification with an Oligo(dT) primer indicates they are polyadenylated mature mRNAs. Alternatively, these intron 4-retaining transcripts may be translated into a truncated whirlin protein, containing only the first two PDZ domains, comparable to the proposed murine N-terminal truncated protein isoforms (Belyantseva *et al.*, 2005; Mathur *et al.*, 2015; Mburu *et al.*, 2003). Unfortunately, antibodies to detect the N-terminus of whirlin are not available to determine if this transcript is translated. Nevertheless, the high prevalence of these transcripts suggests their functional relevance, either through encoding a truncated protein or via mechanisms akin to those described by Boutz *et al.* (2015).

The PacBio Iso-Seq standard and long-transcript workflows were unable to sequence transcripts encoding the currently known largest isoforms of usherin and ADGRV1, but did identify the presence of *USH2A* transcripts encoding usherin isoform A, and several shorter *ADGRV1* transcripts analogous to murine *Vlgr1d* and *Vlgr1e* that all exceed the average GENCODE transcript length of 2.4 kb. Moreover, the long-transcript workflow provided improved coverage across the largest isoforms of these genes. Manual examination of the *USH2A* reads revealed the sporadic inclusion of intronic

sequences, which we have termed cryptic exons. The occurrence of these cryptic exons may be a result of ongoing splicing, as previous studies using ultra-deep sequencing have demonstrated that recursive splicing can take place in which introns are spliced out in multiple steps which can lead to the generation of a cryptic exon as an intermediate product (Gazzoli *et al.*, 2016). Furthermore, it has been shown that cryptic exons can mark regions susceptible to pathogenic pseudoexon-inducing variants (Braun *et al.*, 2013). For example, the sporadic inclusion of 87 bp of *USH2A* intron 20, previously identified as PE20, was also observed in our dataset. Reurink *et al.* (2023) showed that the c.4397-3890A>G pathogenic variant resulted in consistent inclusion of PE20, containing an in-frame stop codon (p.Ala1465_Ala1466ins*5), across all *USH2A* transcripts. This example illustrates how deep intronic variants can lead to the permanent inclusion of potentially pathogenic PEs. An ASO-based splicing correction therapy was developed to target this aberrant PE20 inclusion. This underscores the diagnostic potential of a detailed Iso-Seq data analysis, which could inform the development of targeted diagnostic panels and therapeutic strategies.

Despite using our PacBio long-transcript workflow, we were unable to sequence the currently known largest transcript isoforms of *USH2A* and *ADGRV1*. We chose to utilize PacBio Iso-Seq for our study due to its reputation for offering higher sequencing accuracy compared to Oxford Nanopore Technologies (ONT) (Tvedte *et al.*, 2021). While the ONT platform presents advantages, such as adaptive sampling for target enrichment and direct RNA or cDNA sequencing, which may improve transcript capture in future studies, an independent ONT dataset from human retina samples also failed to capture the full-length transcripts for *USH2A* and *ADGRV1*, highlighting the current challenges faced by both sequencing platforms. Considering these challenges, we utilized the Samplix Xdrop Sort system for an 'indirect target enrichment' (Madsen *et al.*, 2020) on human retina cDNA to enrich for *USH2A* and *ADGRV1* transcripts. This facilitated the successful capture and sequencing of *ADGRV1* transcripts as well as cDNA molecules of *USH2A* encoding the largest isoform B, thereby demonstrating the presence of this complete cDNA molecule (18.9 kb) for the first time. In addition

to its original purpose of genomic DNA enrichment, this marks the first demonstration of applying the Samplix Xdrop Sort system to enrich cDNA samples. However, a limitation of this technique is that the captured cDNA molecules require an amplification step using multiple displacement amplification (MDA), which introduces branched transcripts. Consequently, an enzymatic digestion step is necessary prior to sequencing. Ideally, the workflow would enable full-length amplification of the captured molecules followed by PacBio long-read sequencing, but this remains an area for future improvement.

Collectively, our findings underscore the importance of employing integrated sequencing and analysis approaches to capture the full complexity of the transcriptome in specialized tissues like the human neural retina. Through a comprehensive analysis of Usher syndrome-associated transcript isoforms in the human retina, we revealed a more intricate isoform landscape than previously understood. This analysis led to the discovery of novel isoforms and splicing events, with significant implications for diagnostics and therapeutic development. While functional studies are necessary to validate the transcript isoforms and elucidate their roles in retinal physiology and pathology, our work highlights the power of combining algorithmic and detailed manual analyses to achieve a deeper understanding at the individual gene level. We therefore advocate for the utilization of the human neural retina sequencing datasets we generated here for similar comprehensive studies, as these datasets have the potential to yield valuable insights into other genes and significantly advance our understanding of complex transcriptomic landscapes.

MATERIAL AND METHODS

Tissue collection for PacBio Iso-Seq long-read mRNA sequencing

Human donor eyes, deemed unsuitable for corneal transplantation, were used in this study. Written consent was obtained from the donors' next of kin in accordance with the guidelines of the Italian National Transplant Centre (Centro Nazionale Trapianti, Rome, Italy). The procurement and processing of these tissues followed Italian law and adhered to the principles of the Declaration of Helsinki and the guidelines of the European Eye Bank Association. Three human neural retinal samples were obtained from non-visually impaired individuals through the Fondazione Banca degli Occhi del Veneto (Venice, Italy). Prior to this study, we performed whole-genome sequencing on DNA isolated from these retinal samples to screen for and exclude any known genetic variants associated with inherited retinal disorders. The eyes were enucleated within 2-12 hours postmortem, and the retinal extraction followed established protocols (Niyadurupola *et al.*, 2011; Osborne *et al.*, 2016). After cornea removal, the eyeball was dissected at the *ora serrata*; iris, lens, and vitreous body were removed before carefully detaching the retina from sclera and retinal pigment epithelium by cutting at the optic nerve head. Subsequently, retinal samples were snap-frozen in cryovials using liquid nitrogen and shipped on dry ice. Detailed information about the donors is presented in Supplemental_Table_S5.

RNA isolation and PacBio Iso-Seq library preparation following the standard workflow

To preserve the integrity of long mRNA molecules, all samples were handled with care during RNA isolation and library preparation, avoiding vortexing or vigorous pipetting to prevent shearing of mRNA molecules. Total RNA was extracted from the human neural retina samples by adding 500 µL of TRIzol reagent-to each sample, which was then homogenized in a 2 mL tube containing a sterile glass bead using a TissueLyser (QIAGEN, Aarhus, Denmark) in two cycles at 30 Hz for 30 seconds. Following a 5-minute incubation at room temperature (RT), 100 µL chloroform was added to the samples which were then mixed and incubated for another 3 minutes at RT before being centrifuged

491 at 12,000 g for 15 minutes. The resulting aqueous phase was collected and mixed with 1 µL glycogen
492 (5 µg/µl) and an equal volume of isopropanol. This mixture was incubated at 20°C for 75 minutes
493 followed by centrifugation at 12,000 g for 30 minutes at 4°C. Subsequently, the supernatant was
494 removed and the remaining RNA pellet was dissolved in MQ water and further purified and DNase
495 treated using the Nucleospin RNA Clean-up Kit (Macherey-Nagel, Düren, Germany) according to the
496 manufacturer's instructions. The total isolated RNA quantity was measured with a Qubit fluorometer
497 (Thermo Fisher Scientific, Waltham, MA, USA). Additionally, RNA integrity number (RIN) values were
498 determined using a 2100 Bioanalyzer (Aligent Technologies, Santa Clara, CA, USA). The RIN values of
499 all three samples exceeded 7.0 (Supplemental_Table_S5), and 300 ng of RNA input was utilized to
500 generate Iso-Seq SMRTbell libraries following the Iso-Seq-Express-Template-Preparation protocol PN
501 101-763-800 version 2.0 (Pacific Biosciences, California, USA). These libraries were prepared using
502 the standard workflow, suitable for samples with a predominant transcript size of approximately 2
503 kb, and the SMRTbell® library binding kit 2.1 (Pacific Biosciences, California, USA). The samples were
504 not labeled with a barcode, and 500 ng of cDNA was used for the subsequent steps in the procedure.
505 The quantification and assessment of the SMRTbell library for each sample were conducted using
506 Qubit (Thermo Fisher Scientific, Waltham, MA, USA) and an Agilent Bioanalyzer 2100 employing HS
507 RNA screentape. The on-plate loading concentration of the final Iso-Seq SMRTbell libraries was set at
508 80 pM, and sequencing was carried out on a Sequel IIe system (Pacific Biosciences, California, USA)
509 with a movie time of 24 hours.
510
511 **PacBio Iso-Seq library preparation following adjusted workflow optimized for large transcripts**
512 300 ng of RNA sample 2 was used to generate the Iso-Seq SMRTbell library using the "Iso-Seq-
513 Express-Template-Preparation" protocol PN 101-763-800 version 2.0 (Pacific Biosciences, California,
514 USA). The library was prepared using the PacBio long transcript workflow suitable for samples with a
515 transcript size larger than 3 kb utilizing the SMRTbell® library binding kit 2.1. To further enhance the
516 enrichment for large transcript sizes, an additional size selection step was included using diluted

AMPure PB beads. This was performed after the standard purification of amplified cDNA for long transcripts (page 6 – 7 of protocol PN 101-763-800 version 2) and after performing an additional 5 PCR cycles as outlined in Appendix 1 (page 11). Next, instead of the recommended ProNex bead purification (page 12), the additional size-selection was carried out using a 3.3X ratio of diluted AMPure PB beads according to the PacBio Procedure "Using AMPure PB Beads for Size-Selection" protocol PN 101-854-900 version 2.0 (Pacific Biosciences, California, USA). The concentration and size distribution of the resulting SMRTbell library was evaluated using Qubit (Thermo Fisher Scientific, Waltham, MA, USA) and an Agilent Bioanalyzer 2100 with HS RNA screentape. The on-plate loading concentration of the final Iso-Seq SMRTbell libraries was set at 120 pM, and sequencing was carried out on a Sequel IIe system (Pacific Biosciences, California, USA) with a movie time of 30 hours.

PacBio long-read Iso-Seq data analysis

Four PacBio long read-RNA Iso-Seq samples (three samples prepared following the PacBio standard workflow, 1 sample prepared following the optimized PacBio long-transcript workflow) were analyzed with IsoQuant (Prjibelski *et al.*, 2023). Three separate datasets were created (Figure 1): the first dataset included the IsoQuant analysis of the three standard workflow samples, like we previously performed in Riepe *et al.* (2024). The second dataset combined the analysis of PacBio standard workflow retina samples 1, 2, and 3 with the retina sample prepared following the optimized long transcript workflow. These two dataset used CCS3 reads. The third dataset consists only of the long-transcript workflow sample, analyzed with less strict filtering settings and using CCS0 reads under the assumption that the longest transcritps might not reach CCS3 as frequenly as average-sized transcript. The sequencing data resulting from the PacBio Iso-Seq approaches, that were used for IsoQuant analyses are available via European Genome-Phenome Archive (EGA) with identifier EGAD50000000720. Detailed properties of the IsoQuant analyses can be found in the Supplemental_Code document.

Data analysis and transcript visualization

RStudio software (v4.3.2) (R Core Team, 2021) was used to obtain read and transcript counts from the IsoQuant output files. GGtranscript package (Gustavsson *et al.*, 2022) was used to visualize transcripts from the GTF output files. ORFs were predicted with Snapgene. Details and R scripts are posted on GitHub (https://github.com/erikdevrieze/USH_retina), and can be found in the Supplemental_Code document.

Targeted Iso-Seq for *USH2A* and *ADGRV1* transcripts using the Samplix Xdrop Sort

The Samplix Xdrop Sort (Samplix ApS, Birkerød, DK) was employed for a targeted Iso-Seq analysis, enriched for *USH2A* and *ADGRV1* transcripts. Of each retina sample, 300 ng of RNA was utilized for cDNA synthesis using the NEBNext Single Cell/Low Input cDNA Synthesis kit (New England Biolabs, Ipswich, MA, USA). For this purpose, the "Preparing Iso-Seq libraries using SMRTbell prep kit 3.0" protocol 102-396-000 version 2.0 (Pacific Biosciences, California, USA) was followed until the cDNA amplification step. The synthesized cDNA of all three samples was pooled and the final DNA concentration of the resulting pool was determined using the Qubit with the single stranded DNA kit (Thermo Fisher Scientific, Waltham, MA, USA) according to the standard workflow. In order to perform the indirect target enrichment, single cDNA molecules were packaged in double emulsion (DE) droplets together with detection sequence primers using the Samplix Xdrop Sort. The detection sequence primers were designed to target a 100-150 bp region located in either the 5', middle or 3' region of full-length USH2A (ENST00000307340.8) and ADGRV1 (ENST00000405460.9) transcripts respectively, as to increase the likelihood of capturing full-length transcripts as well as known and novel shorter isoforms. Primer sequences and targets are provided in Supplemental_Table_S6.

The encapsulation process was performed separately for each of the six dPCR primer sets to ensure targeting of a single region during enrichment. Samples were run in technical duplicates, and two positive controls for amplification and sorting provided by Samplix were included on each Xdrop

DE20 Cartridge (Samplix ApS, Birkerød, Denmark). The input concentration was 2.8 ng of pooled cDNA, to which 20 µl of DE PCR mix (Samplix apS, Birkerød, DK), 0.8 µl of the respective forward and reverse dPCR primers and a volume of nuclease free water was added, resulting in a sample mix with a total volume of 40 µl per sample. For the positive controls, the Samplix positive control DNA and dPCR primers were employed. To ensure proper DE droplet production, the sample and additional reagents were loaded onto the Xdrop DE20 Cartridge in the following order: Firstly, 300 µl of DE PCR buffer (Samplix apS, Birkerød, DK) (diluted in a 1:1 ratio using nuclease free water) was loaded into well #A of the cartridge. Next, 40 µl of diluted DE PCR buffer was loaded on the shelf in well #D. Subsequently, 40 µl of the sample mix was pipetted into well #C. Lastly, 100 µl of DE Droplet Oil (Samplix apS, Birkerød, DK) was loaded into well #B. The cartridge was sealed using a gasket and placed into the Xdrop Sort, after which the DE20 droplet production program was ran. Afterwards, the DE20 droplets were collected from well #D and dispensed into four equal aliquots per sample in PCR-tubes, which were then placed in a thermal cycler in which the following program was ran: 30°C for 5 seconds, 94°C for 3 minutes, 40 cycles of 94°C for 3 seconds followed by 65°C for 30 seconds, ending in a hold at 4°C. Amplification takes place exclusively in droplets containing cDNA molecules in which the detection sequence targeted by the dPCR primers is present. The encapsulated DNA can be stained using an intercalating dye, resulting in a stronger fluorescent signal in the DE20 droplets containing a target transcript, which enables the sorting of these droplets using the Xdrop Sort. To do so, the four aliquots of each sample were pooled again and the aqueous phase was removed before adding 1 ml of DE staining buffer (Samplix apS, Birkerød, DK). Samples were incubated for a duration of 15 minutes at room temperature in the absence of light. An Xdrop DE20 Sort Cartridge (Samplix apS, Birkerød, DK) was sealed with the accompanying sorting foil, ensuring a tight seal around all wells. To prevent residual fluorescence from neighboring lanes, the sorting process was performed in two separate steps: first sorting of the uneven lanes took place, followed by a second sorting run for the even lanes, taking along one of the previously mentioned positive controls in each step. Using the provided lane opener, the lanes to be used in the sorting run were punched open and the

cartridge was loaded with the sample and additional reagents in the following order: 5 µl of Xdrop Blank Oil Droplets (Samplix apS, Birkerød, DK) were loaded in well #Out, 600 µl of DE Sorting Buffer (Samplix apS, Birkerød, DK) (diluted in a 1:1 ratio using nuclease free water) was loaded in well #B1, 300 µl of diluted DE Sorting Buffer was loaded in well #B2 and lastly 300 µl of sample in staining buffer was pipetted in well #In. The cartridge was sealed using a gasket and left to settle in the Xdrop Sort for 5 minutes, after which the Sorting program was ran for the selected lanes. Thresholds for detection and sorting were determined based on the displayed signal of droplets and background fluorescence. As these values differed per sample, the guidelines for picking a threshold as established in the Xdrop Sort Manual (Samplix apS, Birkerød, DK) were consulted. Directly after sorting, droplets were removed from the #Out well of the cartridge. After incubation for 5 minutes at room temperature, the aqueous phase was removed in order to start the process of breaking the droplets to prepare the DNA for the subsequent multiple displacement amplification. The droplets were washed twice using 200 µl of Droplet Sorting Wash Buffer (Samplix apS, Birkerød, DK), after which all the wash buffer was carefully removed. Subsequently, 20 µl of Droplet Break Solution (Samplix apS, Birkerød, DK) was added as well as 1 µl of Droplet Break Color (Samplix apS, Birkerød, DK). After vortexing, the clear break solution phase could be removed, leaving the colored aqueous phase containing the enriched DNA.

Since the DNA amount after sorting was insufficient for direct sequencing, the captured transcripts were amplified using multiple displacement amplification (MDA), employing non-sequence-specific primers. Additionally, the Xdrop Sort was used to package single transcripts in single emulsion (SE) droplets to avoid bias towards smaller transcripts. For this purpose, 10 µl of enriched DNA from each sample was transferred to separate PCR-tubes. Furthermore, 1 pg of unenriched cDNA dissolved in 10 µl nuclease free water was taken along as a positive control, as well as a non-template control. To each of the samples, 1 µl of SE MDA enzyme (Samplix apS, Birkerød, DK), 4 µl of SE MDA mix (5x) (Samplix apS, Birkerød, DK) and 5 µl of nuclease free water were added on ice. Subsequently, 20 µl of

each sample was loaded onto the Xdrop SE85 Cartridge (Samplix apS, Birkerød, DK) according to the manufacturer's instructions, after which 75 µl of SE Droplet Oil (Samplix apS, Birkerød, DK) was administered to each inlet well of the cartridge. The cartridge was then placed into the Xdrop Sort and the SE droplet production program was ran, after which droplets were collected in PCR-tubes. All but 1-2 mm of droplet oil was removed from the bottom of the PCR-tubes, which were then placed in a thermal cycler to be incubated at 30°C for 16 hours, 65°C for 10 minutes and 4°C until release of DNA from the droplets. In order to break the SE droplets, 20 µl of Droplet Break Solution is added to each of the tubes, together with 1 µl of Droplet Break Color. After vortexing, the clear break solution phase could be removed, leaving the colored aqueous phase containing the enriched and amplified DNA.

Enrichment validation was performed using qPCR to compare enriched samples with the original, un-enriched cDNA pool. The amplified 1 pg of unenriched cDNA and non-template control from the MDA served as positive and negative controls, respectively. qPCR primers, designed to target 100-125 bp regions close to the targeted regions of the dPCR primers, were used in technical replicates to negate bias from unintended target capture. Primer compositions and targeted exons are detailed in Supplemental_Table_S6. Each qPCR reaction contained 1 µl of template DNA, 10 µl of Gotaq 2x Master Mix (Promega Corporation, Madison, USA), 1 µl of forward and reverse qPCR primers, and 7 µl of nuclease-free water. Reactions were run on a QuantStudio 3 Real-Time PCR System (Thermo Fisher Scientific, Waltham, MA, USA) with the following program: 50°C for 1 second, 95°C for 10 minutes, 40 cycles of 95°C for 15 seconds and 60°C for 30 seconds, then 95°C for 15 seconds, 60°C for 1 minute, and finally 95°C for 15 seconds.

PacBio library preparation and sequencing - adjusted workflow for Samplix enriched transcripts

To remove the branched structures present in the amplified transcripts generated through the MDA process, an enzymatic digestion was performed. In order to do so, the enriched samples were diluted

with nuclease free water to a total volume of 25.5 µl. To each sample, 3 µl NEBuffer 2 (New England Biolabs, Ipswich, MA, USA) and 1.5 µl of T7 Endonuclease (New England Biolabs, Ipswich, MA, USA) was added before incubation at 37°C for 15 minutes in a thermal cycler. Afterwards, 20 µl of TE buffer (pH8) was added to each of the samples and a custom bead cleanup was performed. To do so, 40 µl of Ampure PB beads were resuspended and pelleted on a magnet. After the supernatant had been completely removed the beads were washed with nuclease free water twice and resuspended in an equal volume of the custom bead buffer consisting of the following components: 20 µl of 1M Tris-HCL (Sigma-Aldrich, Saint Louis, USA), 4 µl 0.5M EDTA pH8 (Thermo Fisher Scientific, Waltham, MA, USA), 640 µl 5M NaCl (Thermo Fisher Scientific, Waltham, MA, USA), 550 µl 40% PEG8000 (Sigma-Aldrich, Saint Louis, USA) and 778 µl nuclease free water. Of this custom bead suspension, 35 µl was added to each sample and incubated at room temperature for 20 minutes. Subsequently, the samples were pelleted on a magnet and the supernatant was removed. Samples were then washed with 70% ethanol and dried for 30 seconds. Afterwards, samples were removed from the magnetic rack and resuspended in nuclease free water before being incubated for 1 minute at 50°C and 5 minutes at room temperature. After spinning the samples down, the beads were pelleted on the magnetic rack until the eluate was clear and the purified DNA could be collected. A total of 500 ng of MDA-amplified and purified DNA was used to create multiplexed SMRTbell amplicon libraries, following the manufacturer's instructions for the SMRTbell prep kit 3.0 102-359-000 (Pacific Biosciences, California, USA). If the sample had less than 500 ng of DNA, linearized plasmids not containing the region of interest were added to spike the samples. The samples were then prepared for sequencing using the Sequel IIe Binding Kit 3.2 (Pacific Biosciences, California, USA), following the recommended protocols from SMRT Link 11.0.0.146107. Finally, 115 µl of the final mix was loaded per well and long read sequencing was performed using the Sequel IIe system (Pacific Biosciences, California, USA). Following sequencing, the reads were processed following a 'cDNA analysis: subreads were demultiplexed using lima V.2.5.0, and combined to create a consensus sequence using CCS V.6.3.0. These reads were filtered RQ 0.99 to obtain HiFi reads, and HiFi reads were then

mapped along the GRCh38 reference genome using pbmm2 V1.8.0 with the –preset Iso-Seq mode. Instead of performing an IsoQuant analysis, we visualized the data in IGV, because despite the enzymatic digestion following the MDA procedure, debranched reads can still contain multiple fragments of the captured transcript and are incompatible with algorithms such as IsoQuant.

Tissue collection for Oxford Nanopore Technology (ONT) sequencing

For the independent ONT validation dataset, rest material from human donors (n=3) without any known or clinical evidence of retinal disease was collected from the tissue banks of either Ghent University Hospital or Antwerp University Hospital. This collection adhered to the ethical standards of the Declaration of Helsinki and received approval from the Ethics Committee of Ghent University Hospital (IRB approval B670201837286). Details about the donors can be found in Supplemental_Table_S7. The eyes were transported in CO2 Independent Medium (Gibco) prior to dissection. To preserve RNA integrity and minimize the effects of autolysis, retinas were only harvested from eyes with a total post-mortem interval of less than 20 hours. Following a visual inspection to ensure no contamination from the retinal pigment epithelium (RPE), the neural retinas were either processed immediately for total RNA isolation or snap-frozen and stored at -80°C for later use.

RNA isolation and Oxford Nanopore Technology (ONT) sequencing

Total RNA was extracted from the post-mortem adult human neural retina samples using the RNeasy Mini kit® (Qiagen), following the manufacturer's instructions. The extracted RNA then underwent DNase treatment (ArcticZymes, Tromsø, Norway) followed by poly(A) capture. Samples of poly(A) mRNA with sufficient quality (RNA Integrity Value, RIN > 8.0) were used for direct-cDNA library preparation using the SQK-DCS109 kit (Oxford Nanopore Technologies (ONT)), with minor modifications to the supplier's protocol. Each prepared library was then loaded onto a FLO-PRO002

flow cell (ONT) and sequenced on an ONT PromethION device for 72 hours. Details regarding the number of reads can be found in Supplemental_Table_S7.

ONT data analysis of selected genes

MinKNOW version 5.1.0 was used to produce FAST5 files, which were subsequently base-called with Guppy version 6.1.5. The reads were then aligned to the Human Reference Genome GRCh38/hg38 using minimap2 (Li, 2021) version 2.24 with the -ax splice flags for spliced alignments. Alignment files were converted to BAM format, sorted, and indexed using SAMtools version 1.15 to produce the dataset investigated here.

Relative expression of *MYO7A* and *WHRN* isoforms

Quantitative PCR analysis was conducted to determine the relative expression levels of the different *MYO7A*- and *WHRN* transcript isoforms identified in the PacBio Iso-Seq datasets across three human neural retina samples. The same RNA used for preparing the PacBio libraries was employed as the template, with 250 ng of RNA being used for cDNA synthesis using the SuperScript IV Reverse Transcriptase kit (Thermo Fisher Scientific, Waltham, MA, USA, #18090200). Quantitative PCR analysis was performed using GoTaq qPCR Master Mix (Promega), following the manufacturer's protocol. For *MYO7A,* transcript-specific primers were designed and validated to target the 5' canonical start site, the 5' alternative start site, and the 3' end of the identified *MYO7A* isoforms, as well as the reference gene *GUSB*. Amplifications were carried out using the QuantStudio 3 Real-Time PCR system (Applied Biosystems, Waltham, MA, USA), with PCR reactions performed in triplicate. Relative gene expression levels compared to the reference gene *GUSB* were determined using the $2^{-\Delta Ct}$ method.

For *WHRN,* we designed and validated transcript-specific primers to target isoforms with intron 4 retention and those containing exon 7B. In the absence of a universally present *WHRN*

transcript region, we included a primer pair targeting *WHRN* exons 8-9 to capture a segment present in nearly all isoforms, designated as 'total' *WHRN*. Expression levels of *WHRN* transcripts were determined using the $2^{-\Delta Ct}$ method, and plotted relative to the expression of the exon 8-9 target representing 'total *WHRN* transcripts' set at 1. The primers used are listed in Supplemental_Table_S8.

Analysis of protein isoforms

To predict the 2D protein domain architecture of encoded proteins in silico analyses were performed using the SMART online tool (Letunic *et al.*, 2021). 3D protein structures were modelled using AlphaFold (Jumper *et al.*, 2021) using the Google Colab notebook (v1.5.3) with standard settings. The 3D structures were visualized with YASARA (Krieger and Vriend, 2014), and structural alignments were conducted using the MUSTANG algorithm (Konagurthu *et al.*, 2006).

DATA ACCESS

The PacBio Iso-Seq data generated for this study, and the ONT long-read mRNA sequencing data used for validation of selected transcripts and events, are publicly accessible through the European Genome-Phenome Archive at www.ega-archive.org, under the accession number EGAD50000000720. Additionally, genome browser tracks of the analyzed PacBio Iso-Seq data can be accessed at https://genome-euro.ucsc.edu/s/tabeariepe/retina_atlas. The original codes have been made publicly accessible through the GitHub repository (https://github.com/cmbi/Neural-Retina-Atlas and https://github.com/erikdevrieze/USH_retina), and can be found in the Supplemental_Code document.

COMPETING INTEREST STATEMENT

The authors have no competing interest.

ACKNOWLEDGEMENTS

This study was financially supported by Stichting UitZicht (2019-16), CUREUsher and Stichting Ushersyndroom. Purchase of the Samplix Xdrop Sort Instrument was enabled by an internal Radboudumc technology innovation grant (to AH). We would like to thank the R&D department of Samplix ApS for their expert guidance in optimizing the research protocol to enable cDNA capturing. Finally, we also thank the Radboud Technology Center Genomics for the library preparation and sequencing of all samples.

AUTHOR CONTRIBUTIONS

MS: Investigation, Visualization, Methodology, Project administration, Writing – original draft. TR: Formal Analysis, Data curation, Writing-review and editing. NZ: Investigation, Methodology, Writing–review and editing. RS: Formal Analysis, Writing–review and editing. MK: Investigation, Writing–review and editing. JO: Investigation, Writing–review and editing. RT: Formal Analysis, Writing–review and editing. BF: Resources, Writing–review and editing. SF: Resources, Writing–review and editing. ADR: Resources, Writing–review and editing. ED: Resources, Writing–review and editing. SB: Formal Analysis, Writing–review and editing. HK: Funding acquisition, Writing–review and editing. SR: Writing–review and editing. FC: Resources, Writing–review and editing. AH: Funding acquisition, Methodology, Writing–review and editing. FPMC: Conceptualization, Funding acquisition, Writing–review and editing. P't H: Conceptualization, Funding acquisition, Writing–review and editing. EW: Funding acquisition, Supervision, Conceptualization, Writing–review and editing. EV: Funding acquisition, Conceptualization, Methodology, Supervision, Writing–review and editing.

773 **Figure legends:**

774 **Figure 1: Overview of the sequencing workflows and subsequent analyses.** The figure illustrates the
775 sequencing workflows and subsequent analysis performed on RNA extracted from three human neural retina
776 samples. The workflows included PacBio long-read mRNA Iso-Seq using both the standard and an optimized
777 long transcript workflow. The analysis was carried out in three distinct datasets: Dataset 1 comprised the
778 standard workflow samples analyzed with IsoQuant, Dataset 2 involved a combined analysis of the reads
779 obtained with standard and optimized long transcript workflows, and Dataset 3 focused solely on reads
780 obtained with from the long transcript workflow. Additionally, an "indirect targeted enrichment" of transcripts
781 for the *USH2A* and *ADGRV1* genes was achieved using the Samplix Xdrop System, followed by PacBio long-read
782 sequencing and cDNA analysis. All reads mapping to Usher syndrome-associated transcript isoforms were
783 manually curated using BAM files in the Integrative Genomics Viewer. An independent Oxford Nanopore
784 Technology (ONT) long-read sequencing dataset of three independent retina samples was used to validate
785 findings.
786
787 **Figure 2: Exploring the Usher syndrome-associated transcript isoform landscape in the human neural retina**
788 **using PacBio long-read mRNA Iso-Seq. A.**The size distribution of sequenced transcripts derived from the
789 standard workflow (blue) and optimized long workflow (red) datasets. For the standard workflow dataset, the
790 mean size distribution across the three sequenced samples is depicted ± standard deviation (SD). **B.**
791 Comparison of Usher syndrome-associated transcript coverage between the standard workflow and optimized
792 long workflow dataset. The Usher genes are arranged in order from smallest to largest coding sequence, with
793 the coding sequence length of the largest known transcript for each gene provided in brackets. For the
794 standard workflow dataset, the mean +/- SD transcript length across the three sequenced samples is presented
795 **C.** Quantification of the percentage of reads displaying intron retention in standard workflow samples 1-3
796 (mean of 3 samples ± SD) versus long workflow sample 4.
797
798 **Figure 3: *MYO7A* transcripts identified by IsoQuant analysis compared to known isoforms from the**
799 **literature. A.** The GENCODE reference transcript is depicted at the top in green, followed by the known human
800 *MYO7A* transcript isoforms in blue (Gilmore *et al.,* 2023) and the murine isoforms in grey (Li *et al.,* 2020). The
801 *MYO7A* IsoQuant transcripts are depicted in red. The light green, blue, grey and red colors indicate the

802 untranslated regions (UTR) and the dark green, blue, grey and red colors indicate the open reading frame (ORF)
803 of each transcript. Differences between the IsoQuant transcript isoforms and the GENCODE reference
804 transcript are highlighted in grey boxes. **B.** Relative expression of *MYO7A* isoforms based on literature in either
805 the retina or the cochlea. **C.** The Transcripts Per Million (based on dataset 1) for each IsoQuant isoform are
806 presented for the three individual samples. **D.** The predicted 2D protein domain architecture of the MYO7A
807 protein isoforms with the canonical 5' start and the alternative 5' start from transcript20052.Chr11.nic. The bar
808 below the 2D protein structures displays the amino acid positions. IQ = isoleucine-glutamine motif, CC1 =
809 Coiled Coil domain, LowC = Low complexity region, MyTH4 = Myosin Tail Homology 4, SH3 = SRC Homology 3
810 domain. **E.** AlphaFold2 3D protein predictions of the MYO7A protein isoforms, modeled from the 5' start to the
811 end of the Myosin motor head domain. **F.** RT-qPCR analysis of the relative expression of the *MYO7A* canonical
812 5' start site, the alternative 5' start, and the 3' end is shown. The locations of the primers for this RT-qPCR are
813 indicated with the arrows on top of the IsoQuant isoforms in Figure 3A.
814
815 **Figure 4: *WHRN* transcript isoforms identified by IsoQuant analysis compared to known isoforms from the**
816 **literature. A.** The GENCODE reference transcript is depicted at the top in green, followed by human *WHRN*
817 transcript isoforms from literature in blue (van Wijk *et al.*, 2006) and the murine transcript isoforms in grey
818 (Belyantseva *et al.*, 2005; Ebrahim *et al.*, 2016; Mburu *et al.*, 2003). The *WHRN* IsoQuant transcripts are
819 depicted in red. The light green, blue, grey and red colors indicate the untranslated regions (UTR) and the dark
820 green, blue, grey and red colors indicate the open reading frame (ORF) of each transcript. Differences between
821 the IsoQuant transcripts and the GENCODE reference transcript are highlighted in grey boxes. **B.** Relative
822 expression of *WHRN* isoforms based on literature in either the retina or the cochlea. **C.** The Transcripts Per
823 Million (based on dataset 1) for each IsoQuant transcript isoform are presented for the three individual
824 samples. **D.** The predicted 2D protein domain architecture of the encoded WHRN protein isoforms. Light blue
825 and green boxes highlight the difference between the WHRN reference isoform and the protein isoform
826 encoded by exon 7B-containg transcript13724.Chr9.nic. **E.** AlphaFold2 3D protein predictions of two WHRN
827 isoforms; reference isoform ENST00000362057.4 in green and transcript13724.Chr9.nic in red, with the alpha
828 helix encoded by the novel exon 7B highlighted in blue. **F.** RT-qPCR analysis of the expression of the *WHRN*
829 transcripts containing exon 7B, and *WHRN* transcripts with intron 4 retention, relative to all *WHRN* transcripts
830 containing exons 8-9.

Figure 5: *USH2A* transcript isoforms identified by IsoQuant analysis, manual curation and and Samplix Xdrop targeted enrichment. A. The GENCODE reference transcript is depicted at the top in green, followed by the known human *USH2A* transcript isoforms in blue (Van Wijk *et al.*, 2004). The *USH2A* IsoQuant transcripts are depicted in red. The light green, blue and red colors indicate the untranslated regions (UTR) and the dark green, blue and red colors indicate the open reading frame (ORF) of each transcript. Differences between the IsoQuant transcript isoforms and the GENCODE reference transcript are highlighted in grey boxes. **B.** Relative expression of *USH2A* isoforms based on literature in either the retina or the cochlea. **C.** The Transcripts Per Million (based on dataset 1) for each IsoQuant transcript are presented for the three individual samples. **D.** Proposed *USH2A* transcript isoforms based on manual curation and and Samplix Xdrop targeted enrichment. The GENCODE reference transcript is depicted in green, followed by the proposed *USH2A* transcript isoforms and events based on manual curation of BAM-files using the Integrative Genomics Viewer (IGV) in red, and the proposed transcript isoform following the Samplix Xdrop targeted enrichment in orange. The light green, red and orange colors indicate the untranslated regions (UTR) and the dark green, red and orange colors indicate the open reading frame (ORF) of each transcript. Differences between the proposed transcript isoforms and the GENCODE reference transcript are highlighted in grey boxes. The overview of sporadic incorporation of cryptic exons indicates the presence of PE8 and PE20 as previously described by Reurink *et al.* (2023). Additionally, locations where cryptic exons are occasionally incorporated at sites that are not yet associated to deep-intronic pathogenic variants are indicated with black arrows.

Figure 6: *ADGRV1* proposed transcript isoforms from manual curation and Samplix Xdrop targeted enrichment. The GENCODE reference transcript is depicted at the top in green, followed by the *ADGRV1* proposed retinal transcript isoforms and events based on manual curation of BAM-files using the Integrative Genomics Viewer (IGV) in red, and proposed transcript isoforms following the Samplix Xdrop targeted enrichment in orange. The light green, red and orange colors indicate the untranslated regions (UTR) and the dark green, red and orange colors indicate the open reading frame (ORF) of each transcript. Differences between the proposed transcript isoforms and the GENCODE reference transcript are highlighted in grey box.

REFERENCES

Abad-Morales V, Navarro R, Bures-Jelstrup A, Pomares E. 2020. Identification of a novel homozygous ARSG mutation as the second cause of Usher syndrome type 4. Am J Ophthalmol Case Rep 19:100736.

Adato A, Vreugde S, Joensuu T, Avidan N, Hamalainen R, Belenkiy O, Olender T, Bonne-Tamir B, Ben-Asher E, Espinos C et al. . 2002. USH3A transcripts encode clarin-1, a four-transmembrane-domain protein with a possible role in sensory synapses. Eur J Hum Genet 10(6):339-50.

Ahmed ZM, Riazuddin S, Bernstein SL, Ahmed Z, Khan S, Griffith AJ, Morell RJ, Friedman TB, Riazuddin S, Wilcox ER. 2001. Mutations of the protocadherin gene PCDH15 cause Usher syndrome type 1F. Am J Hum Genet 69(1):25-34.

Belyantseva IA, Boger ET, Naz S, Frolenkov GI, Sellers JR, Ahmed ZM, Griffith AJ, Friedman TB. 2005. Myosin-XVa is required for tip localization of whirlin and differential elongation of hair-cell stereocilia. Nat Cell Biol 7(2):148-56.

Bolz H, von Brederlow B, Ramirez A, Bryda EC, Kutsche K, Nothwang HG, Seeliger M, del CSCM, Vila MC, Molina OP et al. . 2001. Mutation of CDH23, encoding a new member of the cadherin gene family, causes Usher syndrome type 1D. Nat Genet 27(1):108-12.

Booth KT, Kahrizi K, Babanejad M, Daghagh H, Bademci G, Arzhangi S, Zareabdollahi D, Duman D, El-Amraoui A, Tekin M et al. . 2018. Variants in CIB2 cause DFNB48 and not USH1J. Clin Genet 93(4):812-821.

Boutz PL, Bhutkar A, Sharp PA. 2015. Detained introns are a novel, widespread class of post-transcriptionally spliced introns. Genes Dev 29(1):63-80.

Braun TA, Mullins RF, Wagner AH, Andorf JL, Johnston RM, Bakall BB, Deluca AP, Fishman GA, Lam BL, Weleber RG et al. . 2013. Non-exomic and synonymous variants in ABCA4 are an important cause of Stargardt disease. Hum Mol Genet 22(25):5136-45.

Cao H, Wu J, Lam S, Duan R, Newnham C, Molday RS, Graziotto JJ, Pierce EA, Hu J. 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(1):e15860.

Ciampi L, Mantica F, Lopez-Blanch L, Permanyer J, Rodriguez-Marin C, Zang J, Cianferoni D, Jimenez-Delgado S, Bonnal S, Miravet-Verde S 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(29):e2117090119.

de Bruijn SE, Rodenburg K, Corominas J, Ben-Yosef T, Reurink J, Kremer H, Whelan L, Plomp AS, Berger W, Farrar GJ 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(3):100345.

Ebermann I, Scholl HP, Charbel Issa P, Becirovic E, Lamprecht J, Jurklies B, Millan JM, Aller E, Mitter D, Bolz H. 2007. A novel gene for Usher syndrome type 2: mutations in the long isoform of whirlin are associated with retinitis pigmentosa and sensorineural hearing loss. Hum Genet 121(2):203-11.

Ebrahim S, Ingham NJ, Lewis MA, Rogers MJC, Cui R, Kachar B, Pass JC, Steel KP. 2016. Alternative Splice Forms Influence Functions of Whirlin in Mechanosensory Hair Cell Stereocilia. Cell Rep 15(5):935-943.

Eudy JD, Weston MD, Yao S, Hoover DM, Rehm HL, Ma-Edmonds M, Yan D, Ahmad I, Cheng JJ, Ayuso C et al. . 1998. Mutation of a gene encoding a protein with extracellular matrix motifs in Usher syndrome type IIa. Science 280(5370):1753-7.

Gazzoli I, Pulyakhina I, Verwey NE, Ariyurek Y, Laros JF, t Hoen PA, Aartsma-Rus A. 2016. Non-sequential and multi-step splicing of the dystrophin transcript. RNA Biol 13(3):290-305.

Gilmore WB, Hultgren NW, Chadha A, Barocio SB, Zhang J, Kutsyr O, Flores-Bellver M, Canto-Soler MV, Williams DS. 2023. Expression of two major isoforms of MYO7A in the retina: Considerations for gene therapy of Usher syndrome type 1B. Vision Res 212:108311.

Gustavsson EK, Zhang D, Reynolds RH, Garcia-Ruiz S, Ryten M. 2022. ggtranscript: an R package for the visualization and interpretation of transcript isoforms using ggplot2. Bioinformatics 38(15):3844-3846.

Hasson T, Heintzelman MB, Santos-Sacchi J, Corey DP, Mooseker MS. 1995. Expression in cochlea and retina of myosin VIIa, the gene product defective in Usher syndrome type 1B. Proc Natl Acad Sci U S A 92(21):9815-9.

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Zidek A, Potapenko A et al. . 2021. Highly accurate protein structure prediction with AlphaFold. Nature 596(7873):583-589.

Konagurthu AS, Whisstock JC, Stuckey PJ, Lesk AM. 2006. MUSTANG: a multiple structural alignment algorithm. Proteins 64(3):559-74.

Krieger E, Vriend G. 2014. YASARA View - molecular graphics for all devices - from smartphones to workstations. Bioinformatics 30(20):2981-2.

Letunic I, Khedkar S, Bork P. 2021. SMART: recent updates, new developments and status in 2020. Nucleic Acids Res 49(D1):D458-D460.

Li H. 2021. New strategies to improve minimap2 alignment accuracy. Bioinformatics 37(23):4572-4574.

Li S, Mecca A, Kim J, Caprara GA, Wagner EL, Du TT, Petrov L, Xu W, Cui R, Rebustini IT et al. . 2020. Myosin-VIIa is expressed in multiple isoforms and essential for tensioning the hair cell mechanotransduction complex. Nat Commun 11(1):2066.

Linnert J, Knapp B, Guler BE, Boldt K, Ueffing M, Wolfrum U. 2023. Usher syndrome proteins ADGRV1 (USH2C) and CIB2 (USH1J) interact and share a common interactome containing TRiC/CCT-BBS chaperonins. Front Cell Dev Biol 11:1199069.

Liu MM, Zack DJ. 2013. Alternative splicing and retinal degeneration. Clin Genet 84(2):142-9.

Liu X, Vansant G, Udovichenko IP, Wolfrum U, Williams DS. 1997. Myosin VIIa, the product of the Usher 1B syndrome gene, is concentrated in the connecting cilia of photoreceptor cells. Cell Motil Cytoskeleton 37(3):240-52.

Madsen EB, Hoijer I, Kvist T, Ameur A, Mikkelsen MJ. 2020. Xdrop: Targeted sequencing of long DNA molecules from low input samples using droplet sorting. Hum Mutat 41(9):1671-1679.

Mathur PD, Zou J, Zheng T, Almishaal A, Wang Y, Chen Q, Wang L, Vashist D, Brown S, Park A et al. . 2015. Distinct expression and function of whirlin isoforms in the inner ear and retina: an insight into pathogenesis of USH2D and DFNB31. Hum Mol Genet 24(21):6213-28.

Mburu P, Mustapha M, Varela A, Weil D, El-Amraoui A, Holme RH, Rump A, Hardisty RE, Blanchard S, Coimbra RS et al. . 2003. Defects in whirlin, a PDZ domain molecule involved in stereocilia elongation, cause deafness in the whirler mouse and families with DFNB31. Nat Genet 34(4):421-8.

954 Murphy D, Cieply B, Carstens R, Ramamurthy V, Stoilov P. 2016. The Musashi 1 Controls
955 the Splicing of Photoreceptor-Specific Exons in the Vertebrate Retina. PLoS Genet
956 12(8):e1006256.
957 Niyadurupola N, Sidaway P, Osborne A, Broadway DC, Sanderson J. 2011. The development
958 of human organotypic retinal cultures (HORCs) to study retinal neurodegeneration. Br
959 J Ophthalmol 95(5):720-6.
960 Osborne A, Hopes M, Wright P, Broadway DC, Sanderson J. 2016. Human organotypic
961 retinal cultures (HORCs) as a chronic experimental model for investigation of retinal
962 ganglion cell degeneration. Exp Eye Res 143:28-38.
963 Pandya-Jones A, Black DL. 2009. Co-transcriptional splicing of constitutive and alternative
964 exons. RNA 15(10):1896-908.
965 Pardo-Palacios FJ, Wang D, Reese F, Diekhans M, Carbonell-Sala S, Williams B, Loveland
966 JE, De Maria M, Adams MS, Balderrama-Gutierrez G et al. . 2024. Systematic
967 assessment of long-read RNA-seq methods for transcript identification and
968 quantification. Nat Methods 21(7):1349-1363.
969 Prjibelski AD, Mikheenko A, Joglekar A, Smetanin A, Jarroux J, Lapidus AL, Tilgner HU.
970 2023. Accurate isoform discovery with IsoQuant using long reads. Nat Biotechnol
971 41(7):915-918.
972 R Core Team. 2021. R: A Language and Environment for Statistical Computing.
973 Reurink J, Weisschuh N, Garanto A, Dockery A, van den Born LI, Fajardy I, Haer-Wigman
974 L, Kohl S, Wissinger B, Farrar GJ et al. . 2023. Whole genome sequencing for
975 USH2A-associated disease reveals several pathogenic deep-intronic variants that are
976 amenable to splice correction. HGG Adv 4(2):100181.
977 Riazuddin S, Belyantseva IA, Giese AP, Lee K, Indzhykulian AA, Nandamuri SP, Yousaf R,
978 Sinha GP, Lee S, Terrell D et al. . 2012. Alterations of the CIB2 calcium- and
979 integrin-binding protein cause Usher syndrome type 1J and nonsyndromic deafness
980 DFNB48. Nat Genet 44(11):1265-71.
981 Riepe TV, Stemerdink M, Salz R, Rey AD, de Bruijn SE, Boonen E, Tomkiewicz TZ, Kwint
982 M, Gloerich J, Wessels H et al. . 2024. A proteogenomic atlas of the human neural
983 retina. Front Genet 15:1451024.
984 Ruiz-Ceja KA, Capasso D, Pinelli M, Del Prete E, Carrella D, di Bernardo D, Banfi S. 2023.
985 Definition of the transcriptional units of inherited retinal disease genes by meta-
986 analysis of human retinal transcriptome data. BMC Genomics 24(1):206.
987 Sarantopoulou D, Brooks TG, Nayak S, Mrcela A, Lahens NF, Grant GR. 2021. Comparative
988 evaluation of full-length isoform quantification from RNA-Seq. BMC Bioinformatics
989 22(1):266.
990 Schellens RTW, Broekman S, Peters T, Graave P, Malinar L, Venselaar H, Kremer H, De
991 Vrieze E, Van Wijk E. 2023. A protein domain-oriented approach to expand the
992 opportunities of therapeutic exon skipping for USH2A-associated retinitis pigmentosa.
993 Mol Ther Nucleic Acids 32:980-994.
994 Sethna S, Scott PA, Giese APJ, Duncan T, Jian X, Riazuddin S, Randazzo PA, Redmond TM,
995 Bernstein SL, Riazuddin S et al. . 2021. CIB2 regulates mTORC1 signaling and is
996 essential for autophagy and visual function. Nat Commun 12(1):3906.
997 Singh J, Padgett RA. 2009. Rates of in situ transcription and splicing in large human genes.
998 Nat Struct Mol Biol 16(11):1128-33.
999 Tomkiewicz TZ. 2024. Skipping, elongation, and restoration. A tale of *ABCA4* splicing to
1000 pave the road towards therapeutic applications. Radboud University Nijmegen.
1001 Tvedte ES, Gasser M, Sparklin BC, Michalski J, Hjelmen CE, Johnston JS, Zhao X, Bromley
1002 R, Tallon LJ, Sadzewicz L et al. . 2021. Comparison of long-read sequencing
1003 technologies in interrogating bacteria and fly genomes. G3 (Bethesda) 11(6).

1004 Udovichenko IP, Gibbs D, Williams DS. 2002. Actin-based motor properties of native myosin
1005 VIIa. J Cell Sci 115(Pt 2):445-50.
1006 Van Wijk E, Pennings RJ, te Brinke H, Claassen A, Yntema HG, Hoefsloot LH, Cremers FP,
1007 Cremers CW, Kremer H. 2004. Identification of 51 novel exons of the Usher
1008 syndrome type 2A (USH2A) gene that encode multiple conserved functional domains
1009 and that are mutated in patients with Usher syndrome type II. The American Journal
1010 of Human Genetics 74(4):738-744.
1011 van Wijk E, van der Zwaag B, Peters T, Zimmermann U, Te Brinke H, Kersten FF, Marker T,
1012 Aller E, Hoefsloot LH, Cremers CW et al. . 2006. The DFNB31 gene product whirlin
1013 connects to the Usher protein network in the cochlea and retina by direct association
1014 with USH2A and VLGR1. Hum Mol Genet 15(5):751-65.
1015 Verpy E, Leibovici M, Zwaenepoel I, Liu XZ, Gal A, Salem N, Mansour A, Blanchard S,
1016 Kobayashi I, Keats BJ et al. . 2000. A defect in harmonin, a PDZ domain-containing
1017 protein expressed in the inner ear sensory hair cells, underlies Usher syndrome type
1018 1C. Nat Genet 26(1):51-5.
1019 Wang B, Tseng E, Regulski M, Clark TA, Hon T, Jiao Y, Lu Z, Olson A, Stein JC, Ware D.
1020 2016. Unveiling the complexity of the maize transcriptome by single-molecule long-
1021 read sequencing. Nat Commun 7:11708.
1022 Weil D, Blanchard S, Kaplan J, Guilford P, Gibson F, Walsh J, Mburu P, Varela A, Levilliers
1023 J, Weston MD et al. . 1995. Defective myosin VIIA gene responsible for Usher
1024 syndrome type 1B. Nature 374(6517):60-1.
1025 Weil D, El-Amraoui A, Masmoudi S, Mustapha M, Kikkawa Y, Laine S, Delmaghani S,
1026 Adato A, Nadifi S, Zina ZB et al. . 2003. Usher syndrome type I G (USH1G) is caused
1027 by mutations in the gene encoding SANS, a protein that associates with the USH1C
1028 protein, harmonin. Hum Mol Genet 12(5):463-71.
1029 Weston MD, Luijendijk MW, Humphrey KD, Moller C, Kimberling WJ. 2004. Mutations in
1030 the VLGR1 gene implicate G-protein signaling in the pathogenesis of Usher syndrome
1031 type II. Am J Hum Genet 74(2):357-66.

TABLE 1: Identified USH-associated retinal transcript isoforms and prevalent events observed across the majority of transcripts

Gene	USH subtype	Identified retinal transcript isoforms		
		Identified Ensembl Spliced Transcripts (ENST)	*Validated novel events and previously unidentified transcripts****	*Corresponding figures*
MYO7A	USH1B	ENST00000409709.9 (MANE)		Figure 3
		ENST00000458637.6	Alternative transcription start site (transcript20055.Chr11.nic)	
USH1C	USH1C	ENST00000527020.5		Figure S1
		ENST00000318024.9		
		ENST00000526313.5		
CDH23	USH1D	ENST00000224721.12*		Figure S2
		ENST00000461841.7	Novel in frame exon 11A and skipping of micro exon 12 (transcript11235.Chr10.nnic)	
		ENST00000475158.1	Exon 69 skipping	
PCDH15	USH1F	ENST00000644397.2 (MANE)		Figure S3
		ENST00000373957.7		
		ENST00000621708.4		
		ENST00000373955.5		
SANS	USH1G	ENST00000614341.5 (MANE)		Figure S4
CIB2	USH1J[#]	-		Figure S5
USH2A	USH2A	ENST00000307340.8 (MANE)**	5'UTR splice events (transcript51429.Chr1.nnic; transcript51430.Chr1.nnic; transcript51439.Chr1.nnic)	Figure 5
		ENST00000366942.3		
ADGRV1	USH2C	ENST00000638316.1		Figure 6, S6
		ENST00000639884.1		
		ENST00000640109.		
		ENST00000640281.1		
WHRN	USH2D	ENST00000362057.4 (MANE)	Inclusion of novel exon 7B (transcript13724.Chr9.nnic)	Figure 4
		ENST00000374057.3	Intron 4 retention (transcript13718.Chr9.nnic)	
		ENST00000265134.10		
CLRN1	USH3A	ENST00000327047.6 (MANE)		Figure S7
		ENST00000472224.1		
ARSG	USH4	ENST00000448504.6		Figure S8
		ENST00000578726.1		

* Presence of transcript isoform is based solely on manual curation of sequenced reads of the sample prepared following the optimized PacBio long transcript workflow.
** Presence of transcript isoform is based solely on results of Samplix Xdrop targeted enrichment.
*** Events are validated using raw sequencing reads of an Oxford Nanopore Technology (ONT) long-read sequencing dataset of independent retina samples.
[#] The association of *CIB2* with Usher syndrome type IJ (USH1J) has been called into question (Booth *et al.*, 2018).

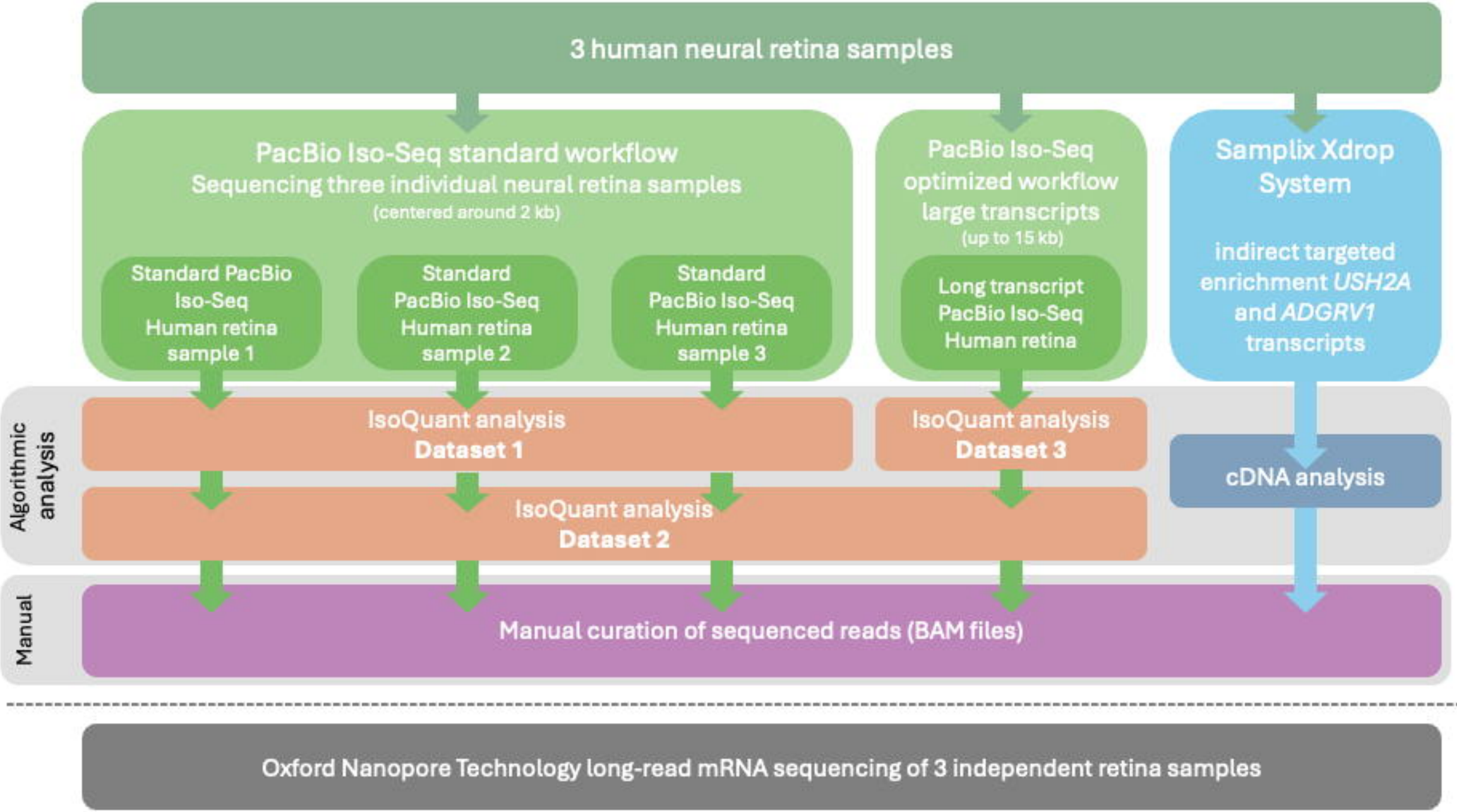

3 human neural retina samples
PacBio Iso-Seq standard workflow
Sequencing three individual neural retina samples
(centered around 2 kb)
Standard PacBio Iso-Seq Human retina sample 1
Standard PacBio Iso-Seq Human retina sample 2
Standard PacBio Iso-Seq Human retina sample 3
PacBio Iso-Seq optimized workflow large transcripts
(up to 15 kb)
Long transcript PacBio Iso-Seq Human retina
Samplix Xdrop System
indirect targeted enrichment USH2A and ADGRV1 transcripts
Algorithmic analysis
IsoQuant analysis Dataset 1
IsoQuant analysis Dataset 3
cDNA analysis
IsoQuant analysis Dataset 2
Manual
Manual curation of sequenced reads (BAM files)
Oxford Nanopore Technology long-read mRNA sequencing of 3 independent retina samples

Figure 2

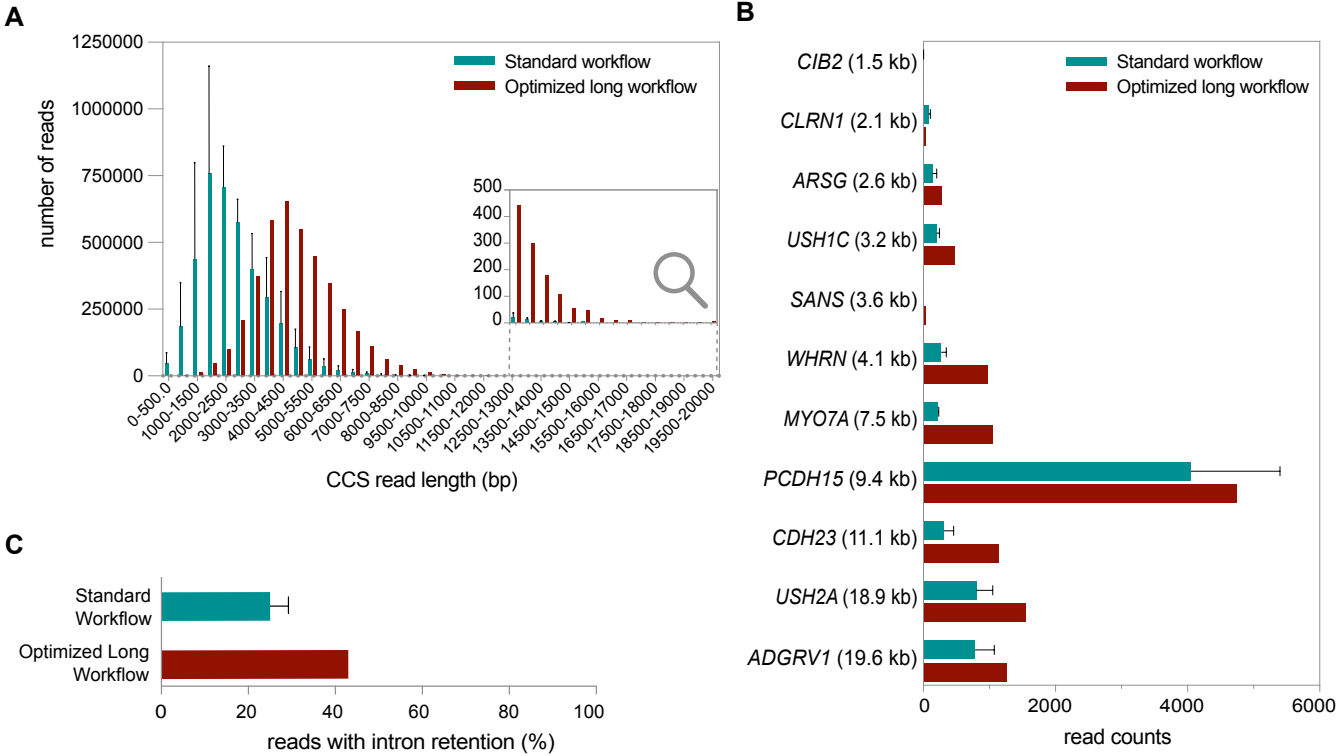

A
number of reads
1250000
1000000
750000
500000
250000
0
Standard workflow
Optimized long workflow
500
400
300
200
100
0
0-500.0
1000-1500
2000-2500
3000-3500
4000-4500
5000-5500
6000-6500
7000-7500
8000-8500
9500-10000
10500-11000
11500-12000
12500-13000
13500-14000
14500-15000
15500-16000
16500-17000
17500-18000
18500-19000
19500-20000
CCS read length (bp)
B
CIB2 (1.5 kb)
CLRN1 (2.1 kb)
ARSG (2.6 kb)
USH1C (3.2 kb)
SANS (3.6 kb)
WHRN (4.1 kb)
MYO7A (7.5 kb)
PCDH15 (9.4 kb)
CDH23 (11.1 kb)
USH2A (18.9 kb)
ADGRV1 (19.6 kb)
Standard workflow
Optimized long workflow
0
2000
4000
6000
read counts
C
Standard Workflow
Optimized Long Workflow
0
20
40
60
80
100
reads with intron retention (%)

Figure 3

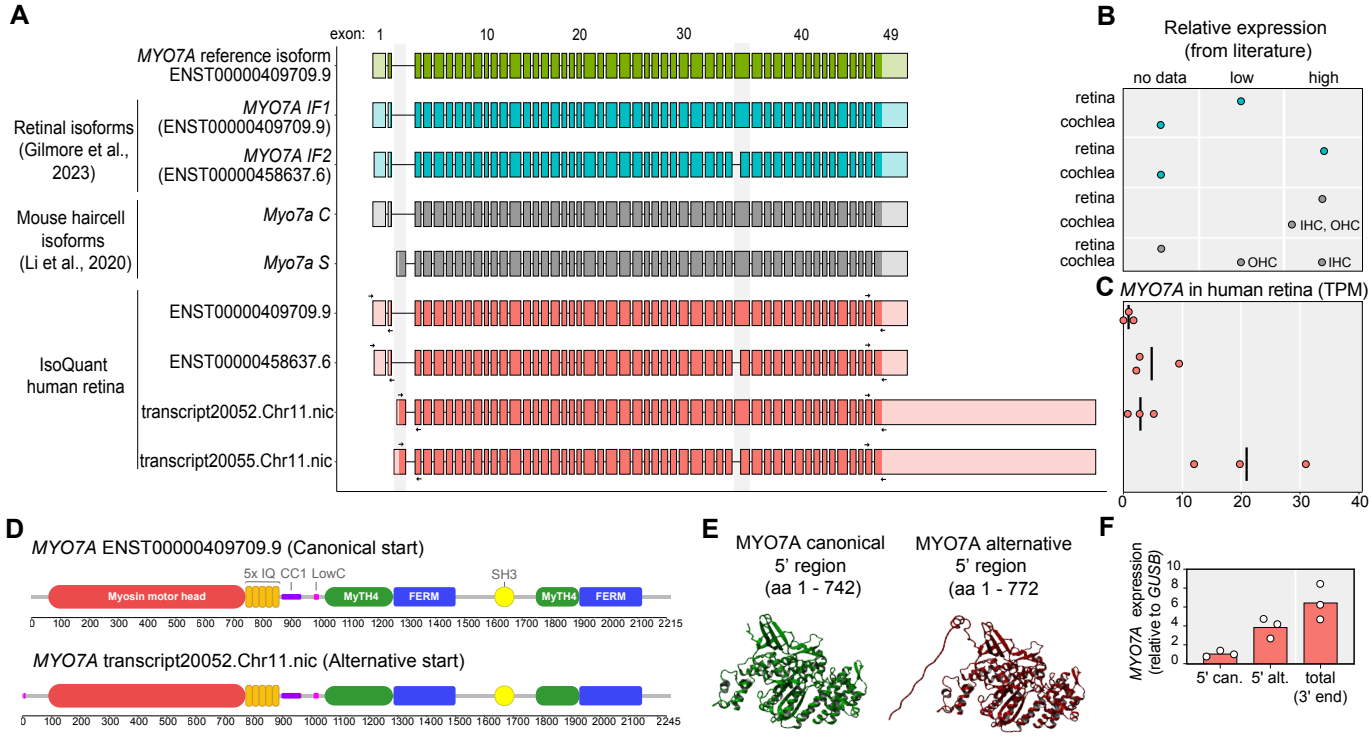

A
exon: 1 10 20 30 40 49
MYO7A reference isoform ENST00000409709.9
Retinal isoforms (Gilmore et al., 2023)
MYO7A IF1 (ENST00000409709.9)
MYO7A IF2 (ENST00000458637.6)
Mouse haircell isoforms (Li et al., 2020)
Myo7a C
Myo7a S
IsoQuant human retina
ENST00000409709.9
ENST00000458637.6
transcript20052.Chr11.nic
transcript20055.Chr11.nic
B
Relative expression (from literature)
no data low high
retina
cochlea
IHC, OHC
OHC
IHC
C
MYO7A in human retina (TPM)
0 10 20 30 40
D
MYO7A ENST00000409709.9 (Canonical start)
5x IQ CC1 LowC
SH3
Myosin motor head
MyTH4
FERM
MyTH4
FERM
0 100 200 300 400 500 600 700 800 900 1000 1100 1200 1300 1400 1500 1600 1700 1800 1900 2000 2100 2215
MYO7A transcript20052.Chr11.nic (Alternative start)
0 100 200 300 400 500 600 700 800 900 1000 1100 1200 1300 1400 1500 1600 1700 1800 1900 2000 2100 2245
E
MYO7A canonical 5' region (aa 1 - 742)
MYO7A alternative 5' region (aa 1 - 772
F
MYO7A expression (relative to GUSB)
0 2 4 6 8 10
5' can.
5' alt.
total (3' end)

Figure 4

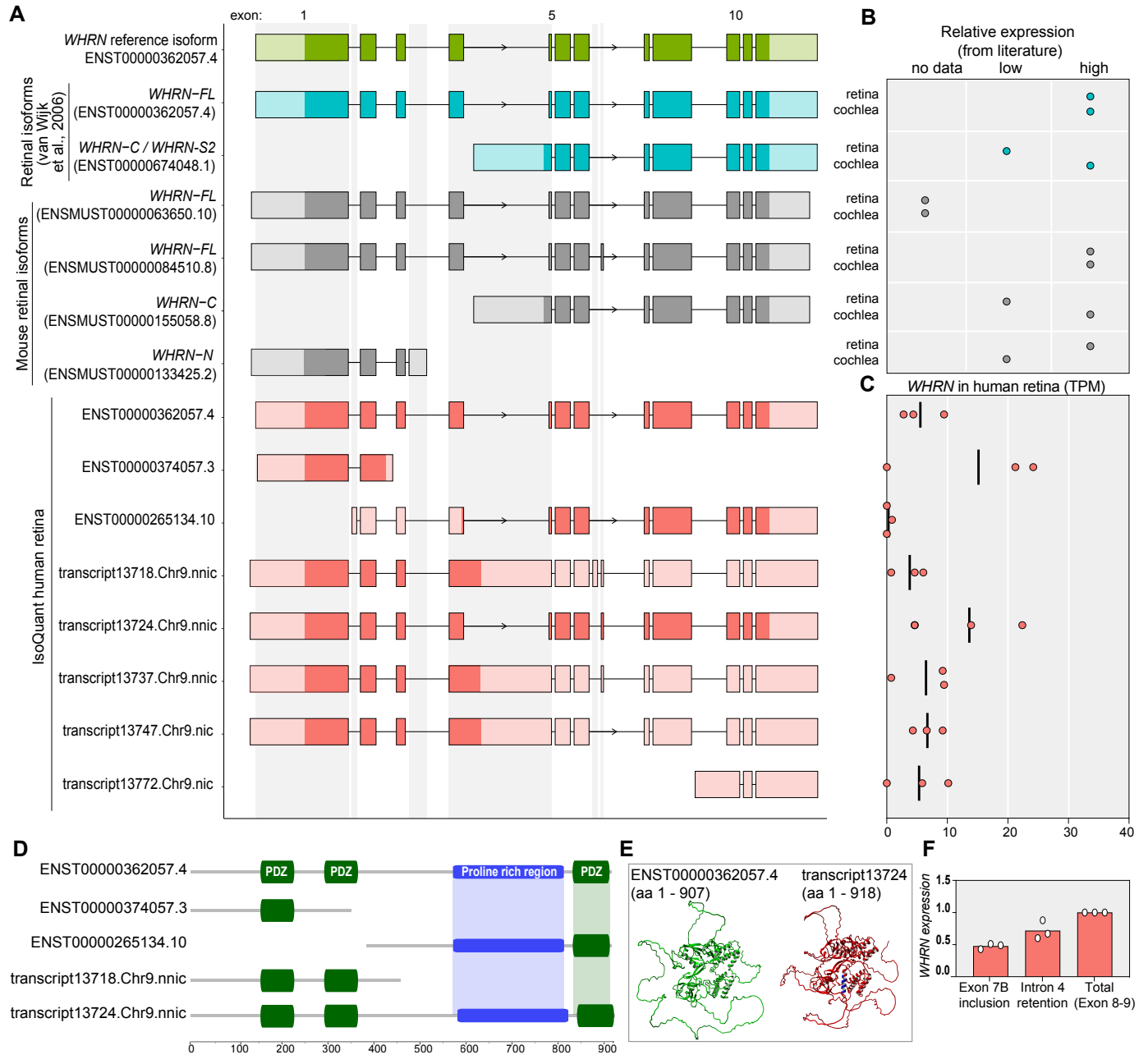

A
exon: 1 5 10
WHRN reference isoform ENST00000362057.4
Retinal isoforms (van Wijk et al., 2006)
WHRN−FL (ENST00000362057.4)
WHRN−C / WHRN-S2 (ENST00000674048.1)
Mouse retinal isoforms
WHRN−FL (ENSMUST00000063650.10)
WHRN−FL (ENSMUST00000084510.8)
WHRN−C (ENSMUST00000155058.8)
WHRN−N (ENSMUST00000133425.2)
IsoQuant human retina
ENST00000362057.4
ENST00000374057.3
ENST00000265134.10
transcript13718.Chr9.nnic
transcript13724.Chr9.nnic
transcript13737.Chr9.nnic
transcript13747.Chr9.nic
transcript13772.Chr9.nic
B
Relative expression (from literature)
no data
low
high
retina
cochlea
C
WHRN in human retina (TPM)
0
10
20
30
40
D
ENST00000362057.4
ENST00000374057.3
ENST00000265134.10
transcript13718.Chr9.nnic
transcript13724.Chr9.nnic
PDZ
PDZ
Proline rich region
PDZ
0
100
200
300
400
500
600
700
800
900
E
ENST00000362057.4 (aa 1 - 907)
transcript13724 (aa 1 - 918)
F
WHRN expression
1.5
1.0
0.5
0.0
Exon 7B inclusion
Intron 4 retention
Total (Exon 8-9)

Figure 5

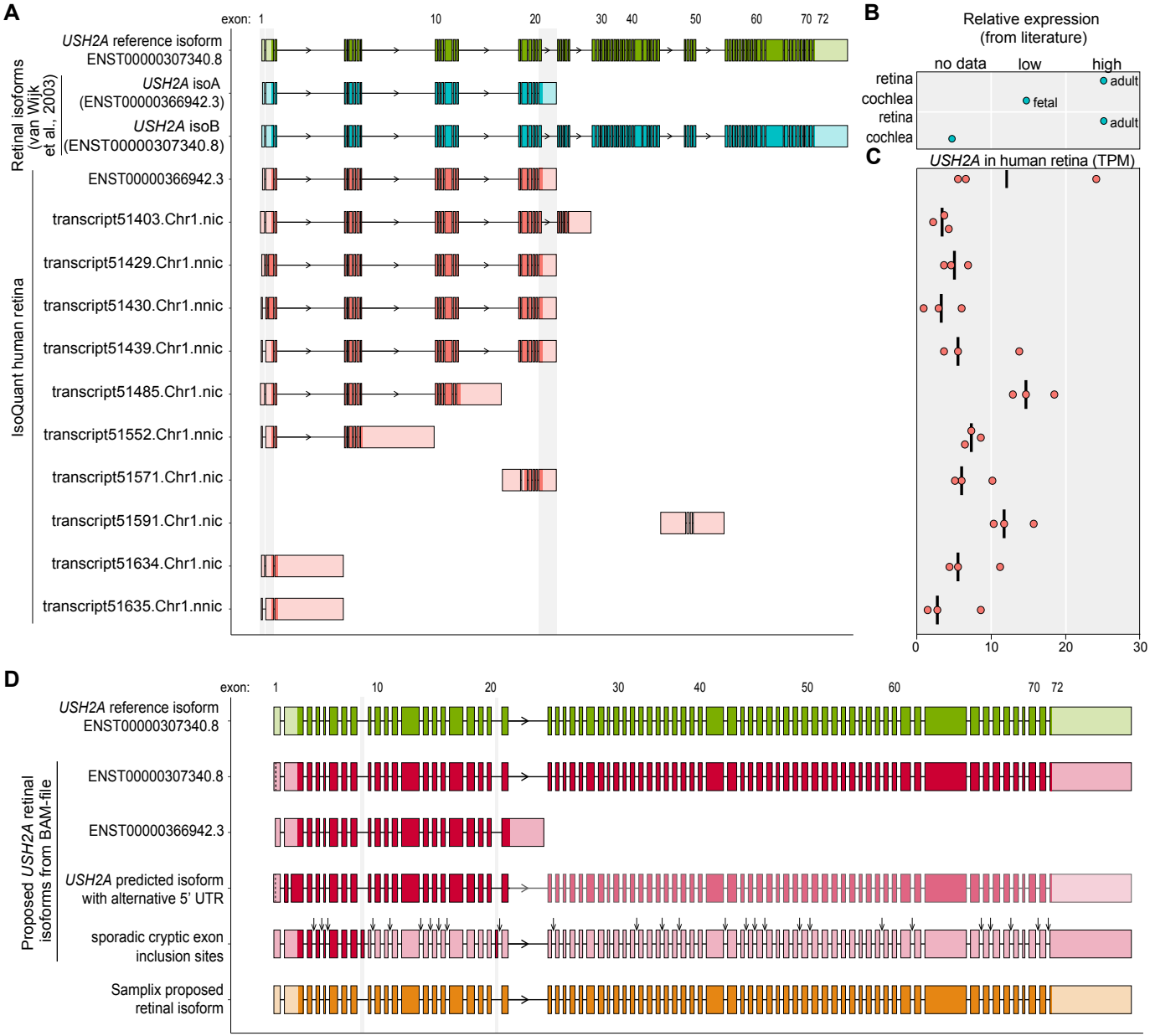
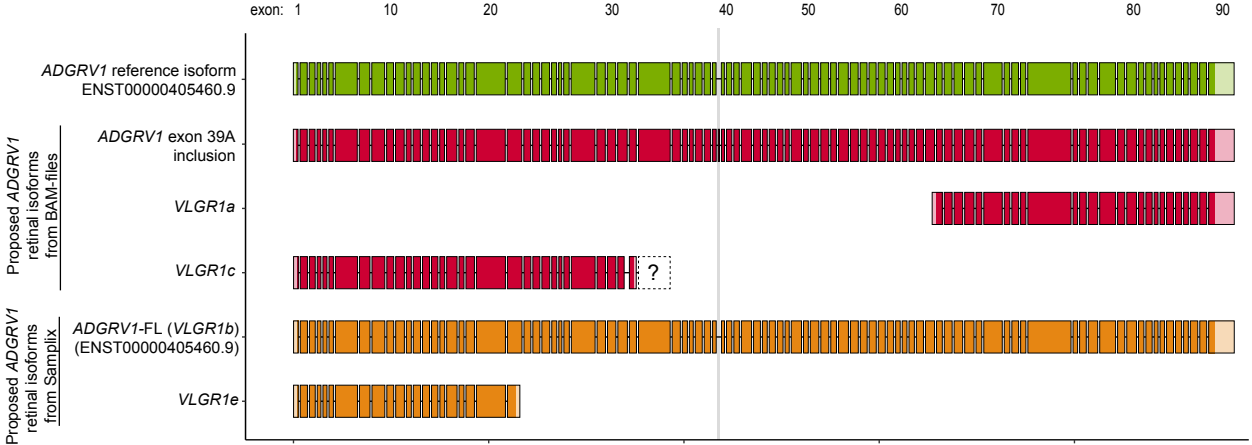

A
exon: 1 10 20 30 40 50 60 70 72
USH2A reference isoform ENST00000307340.8
USH2A isoA (ENST00000366942.3)
USH2A isoB (ENST00000307340.8)
Retinal isoforms (van Wijk et al., 2003)
ENST00000366942.3
transcript51403.Chr1.nic
transcript51429.Chr1.nnic
transcript51430.Chr1.nnic
transcript51439.Chr1.nnic
transcript51485.Chr1.nic
transcript51552.Chr1.nnic
transcript51571.Chr1.nic
transcript51591.Chr1.nic
transcript51634.Chr1.nic
transcript51635.Chr1.nnic
IsoQuant human retina
B
Relative expression (from literature)
no data
low
high
retina
cochlea
retina
cochlea
adult
fetal
adult
C
USH2A in human retina (TPM)
0 10 20 30
D
exon: 1 10 20 30 40 50 60 70 72
USH2A reference isoform ENST00000307340.8
ENST00000307340.8
ENST00000366942.3
USH2A predicted isoform with alternative 5' UTR
sporadic cryptic exon inclusion sites
Samplix proposed retinal isoform
Proposed USH2A retinal isoforms from BAM-file

Figure 6


exon: 1
10
20
30
40
50
60
70
80
90
ADGRV1 reference isoform
ENST00000405460.9
Proposed ADGRV1 retinal isoforms from BAM-files
ADGRV1 exon 39A inclusion
VLGR1a
VLGR1c
?
Proposed ADGRV1 retinal isoforms from Samplix
ADGRV1-FL (VLGR1b)
(ENST00000405460.9)
VLGR1e

Deciphering the largest disease-associated transcript isoforms in the human neural retina with advanced long-read sequencing approaches

Merel Stemerdink, Tabea Riepe, Nick Zomer, et al.

Genome Res. published online March 4, 2025
Access the most recent version at doi:10.1101/gr.280060.124