



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

Merel Stermerdink, 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](https://doi.org/10.1101/gr.280060.124)

P<P	Published online March 4, 2025 in advance of the print journal.
Accepted Manuscript	Peer-reviewed and accepted for publication but not copyedited or typeset; accepted manuscript is likely to differ from the final, published version.
Creative Commons License	This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see https://genome.cshlp.org/site/misc/terms.xhtml). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at http://creativecommons.org/licenses/by-nc/4.0/ .
Email Alerting Service	Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or click here .

Advance online articles have been peer reviewed and accepted for publication but have not yet appeared in the paper journal (edited, typeset versions may be posted when available prior to final publication). Advance online articles are citable and establish publication priority; they are indexed by PubMed from initial publication. Citations to Advance online articles must include the digital object identifier (DOIs) and date of initial publication.

To subscribe to *Genome Research* go to:
<https://genome.cshlp.org/subscriptions>

Published by Cold Spring Harbor Laboratory Press

1 Deciphering the largest disease-associated transcript isoforms in 2 the human neural retina with advanced long-read sequencing 3 approaches

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

5
6 Merel Stemerink^{1,2}, Tabea Riepe^{3,4}, Nick Zomer⁴, Renee Salz³, Michael Kwint⁴, Jaap Oostrik¹, Raoul
7 Timmermans⁴, Barbara Ferrari⁵, Stefano Ferrari⁵, Alfredo Dueñas Rey^{6,7}, Emma Delanote^{6,7},
8 Suzanne E. de Bruijn^{1,4}, Hannie Kremer^{1,2,4}, Susanne Roosing⁴, Frauke Coppieters^{6,7,8}, Alexander
9 Hoischen^{4,9}, Frans P. M. Cremers⁴, Peter A.C. 't Hoen³, Erwin van Wijk^{1*}, Erik de Vrieze^{1*§}

10
11 * These authors contributed equally to this work

12 § Corresponding author: erik.devrieze@radboudumc.nl

13
14 ¹ Department of Otorhinolaryngology, Radboud University Medical Center, Nijmegen, 6525 GA, The
15 Netherlands.

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

18 ³ Department of Medical BioSciences, Radboud University Medical Center, Nijmegen, 6525 GA, The
19 Netherlands.

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

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

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

23 ⁷ Department of Biomolecular Medicine, Ghent University, Ghent, 9000, Belgium.

24 ⁸ Department of Pharmaceutics, Ghent University, Ghent, 9000, Belgium.

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

27

28 **ABSTRACT**

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

48

49 INTRODUCTION

50 The human retina is a complex multicellular tissue that plays a crucial role in visual function by
51 converting light into the electrical signals that are interpreted by the brain. Understanding the
52 molecular composition of the human retina, particularly the wide variety of transcript isoforms
53 expressed there, is essential for comprehending disease mechanisms and designing effective
54 treatment strategies for inherited retinal diseases (IRDs) (Braun *et al.*, 2013). A key factor
55 contributing to the diversity of transcript isoforms expressed in the retina is alternative splicing, a
56 biological process that allows a single gene to encode multiple transcript isoforms leading to tissue-
57 specific differences in gene expression and function. This process involves the use of alternative
58 transcription initiation and termination sites, intron retention, exon skipping, and alternative splice
59 donor and acceptor sites. The retina is a highly specialized tissue that is known to be enriched for
60 these tissue-specific alternative splicing events (Cao *et al.*, 2011; Liu and Zack, 2013). While previous
61 studies using RNA short-read sequencing provided valuable insights, they often fall short in fully
62 characterizing the diverse array of transcript isoforms present in the human retina (Ciampi *et al.*,
63 2022; Murphy *et al.*, 2016; Ruiz-Ceja *et al.*, 2023; Sarantopoulou *et al.*, 2021).

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

75 that have been identified (*MYO7A* (Weil *et al.*, 1995), *USH1C* (Verpy *et al.*, 2000), *CDH23* (Bolz *et al.*,
76 2001), *PCDH15* (Ahmed *et al.*, 2001), *SANS* (Weil *et al.*, 2003), *CIB2* (Riazuddin *et al.*, 2012), *USH2A*
77 (*Eudy et al.*, 1998), *ADGRV1* (Weston *et al.*, 2004), *WHRN* (Ebermann *et al.*, 2007), *CLRN1* (Adato *et*
78 *al.*, 2002) and *ARSG* (Abad-Morales *et al.*, 2020)). With the exception of *CIB2*, the full-length isoforms
79 of the Usher syndrome-associated genes surpass the average human transcript length of 2 kb, with 5
80 of them exceeding the mean transcript length of RETNET genes (4.7 kb). Although the association of
81 *CIB2* with Usher syndrome type IJ (USH1J) has been called into question (Booth *et al.*, 2018),
82 functional studies leave room for the involvement of *CIB2* in Usher syndrome (Linnert *et al.*, 2023;
83 Sethna *et al.*, 2021). Therefore, *CIB2* is included in our analyses as these results could add to the
84 discussion whether it qualifies as an Usher syndrome-associated gene.

85 Obtaining a comprehensive understanding of the Usher syndrome-associated transcript
86 isoforms in the human retina is crucial, as it facilitates the development of therapeutic strategies and
87 enables accurate classification of genetic variants linked to the disorder. To this end, we aimed to
88 provide an overview of the Usher syndrome-associated transcript isoforms in the human neural
89 retina, utilizing the PacBio long-read mRNA Iso-Seq dataset from our previous study (Riepe *et al.*,
90 2024), and supplementing it with two additional datasets aimed at capturing full-length reads from
91 the longest known transcript isoforms. Our previously generated dataset was obtained using the
92 standard PacBio Iso-Seq workflow, optimized for transcripts centered around 2 kb – the average
93 human transcript length – making it suitable for genome-wide studies as we performed in Riepe *et*
94 *al.* (2024). However, as nearly all Usher syndrome-associated genes have transcripts exceeding 2 kb,
95 we aimed to generate an additional Iso-Seq dataset using an optimized PacBio long transcript
96 workflow to enable the sequencing of larger transcripts up to > 10 kb. Despite our efforts, this
97 workflow remained insufficient to enrich for the longest *USH2A* (18.9 kb) and *ADGRV1* (19.6 kb)
98 transcript isoforms. In an attempt to further enhance the capture of full-length transcripts for these
99 two genes, we also employed an ‘indirect targeted enrichment’ approach using the Samplix Xdrop
100 System (Madsen *et al.*, 2020), followed by PacBio long-read sequencing. By integrating the data from

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

104

105

106 RESULTS

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

123

124 **Exploring Usher syndrome-associated transcript isoforms through Iso-Seq PacBio standard- and** 125 **optimized long transcript workflows.**

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

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

141 In Riepe *et al.* (2024), the IsoQuant algorithm was utilized to analyze and categorize the
142 PacBio Iso-Seq standard workflow samples, resulting in dataset 1. A second IsoQuant analysis was
143 conducted on the combined reads from the three PacBio standard workflow samples and the retina
144 sample prepared following the optimized long transcript workflow, resulting in the IsoQuant dataset
145 2. However, the incorporation of the PacBio long transcript workflow sample in the combined
146 IsoQuant analysis negatively impacted the number of identified unique Full Splice Match (FSM)
147 transcripts. The increase in unique Novel In Catalogue (NIC) and Novel Not In Catalogue (NNIC)
148 transcripts suggests that many reads are reallocated in the transcript models predicted by IsoQuant.
149 (Supplemental_Fig_S10). The long transcript workflow results in an enrichment for intron retention
150 events (Figure 2C), which we suspect causes the observed difference between the IsoQuant
151 transcript models. This is exemplified in Supplemental_Fig_S11, showing the IsoQuant transcript
152 models for *MYO7A* based on dataset 1 and 2, of which only those proposed based on dataset 1 were
153 supported by the raw reads. To reliably identify the largest possible transcripts, we reanalyzed the
154 data from the long transcript workflow separately from the standard workflow data (dataset 3). We
155 furthermore lowered the consensus read requirement to 0 (CCS0) and manually curated the reads in
156 IGV.
157

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 transcript20055.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

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

190

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

192 The existing knowledge of *WHRN* transcript isoforms, encoding whirlin proteins, has predominantly
193 been derived from research on mutant mouse models (Belyantseva *et al.*, 2005; Ebrahim *et al.*, 2016;
194 Mburu *et al.*, 2003). Studies by Mburu *et al.* (2003) and Belyantseva *et al.* (2005) have suggested the
195 presence of human retinal *WHRN* transcript isoforms based on cDNA clones, of which only the full-
196 length isoform (ENST00000362057.4) and a C-terminal isoform (ENST0000674048.1) were verified in
197 the human retina using RT-PCR (van Wijk *et al.*, 2006) (Figure 4A, 4B). IsoQuant analysis confirmed
198 the expression of the full-length reference isoform of the *WHRN* gene in the human retina but did
199 not detect the previously reported C-terminal isoform (ENST0000674048.1). However, the IsoQuant
200 analysis did identify an alternative C-terminal *WHRN* transcript isoform starting with a distinct
201 noncoding exon (ENST00000265134.10), although its expression is relatively low (0.3 ± 0.5 TPM)
202 (Figure 4C). An N-terminal *WHRN* isoform was proposed based on the murine *Whrn* transcripts, but
203 not yet experimentally validated in man (Belyantseva *et al.*, 2005; Mburu *et al.*, 2003). IsoQuant
204 indeed reveals the presence of the N-terminal transcript isoform ENST00000374057.3 in 2 out of the
205 3 retinal samples.

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

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

215 Finally, two novel exons were identified within intron 7: designated exon 7A and 7B. While
216 exon 7A was exclusively observed as a noncoding exon and always co-occurred with exon 7B, exon
217 7B was additionally detected as an independent protein-coding exon in transcript13724.Chr9.nic.
218 Manual inspection of the sequencing data also revealed the widespread presence of exon 7B in the
219 Iso-Seq transcripts, which was further corroborated in an independent ONT long-read mRNA
220 sequencing dataset. This 33-nucleotide-long exon 7B is not present in the GENCODE reference
221 annotations for the human *WHRN* gene but does resemble a corresponding 33-nucleotide exon
222 found in the mouse *Whrn* consensus sequence. Structural modeling with AlphaFold2 revealed that
223 this novel exon 7B encodes an additional alpha helix of 11 amino acids in the center of the protein
224 (Figure 4E). A comparison of the TPM levels of the *WHRN* reference transcript (5.5 ± 3.5 TPM for
225 ENST00000362057.4) with those of the exon 7B-containing transcript (13.6 ± 8.9 TPM for
226 transcript13724.Chr9.nnic), suggests that the latter may represent the predominant retinal isoform.
227 Quantitative PCR confirms that inclusion of exon 7B occurs in the majority of *WHRN* transcripts
228 (Figure 4F). Because biallelic pathogenic variants in *WHRN* can cause USH2D, we queried WGS data
229 of our in-house cohort of unsolved IRD patients (de Bruijn *et al.*, 2023) for rare variants in exon 7B,
230 but found no candidate pathogenic variants in this exon.

231

232 **Identification of pseudoexon-prone regions and the successful enrichment of full-length *USH2A*** 233 **isoform B transcripts using Samplix Xdrop Sort.**

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

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

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

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

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

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

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

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

298

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

300 With the largest annotated transcript spanning 19.6 kb, *ADGRV1* - previously known as *MASS1*,
301 *VLGR1*, and *GPR98* - is the largest Usher syndrome-associated gene, and variants in this gene are
302 responsible for Usher syndrome type 2C (Weston *et al.*, 2004). Three distinct human *ADGRV1*
303 transcript isoforms have been previously reported: *VLGR1a*, *VLGR1b*, and *VLGR1c*
304 (Supplemental_Fig_S6). *VLGR1b* represents the transcript encoding the longest isoform, composed of
305 90 exons and encoding a protein of 6,306 amino acids. Similar challenges to those encountered with
306 the large coding sequence of *USH2A* were observed for *ADGRV1*, which prevented the PacBio
307 standard- and long transcript workflows and ONT long-read mRNA sequencing, from sequencing the
308 *VLGR1b* transcript. Furthermore, the IsoQuant analysis did also not identify any transcript isoforms
309 corresponding to the shorter *VLGR1a* and *VLGR1c* transcripts, despite their smaller sizes not posing a
310 barrier to sequencing (Supplemental_Fig_S6). Manual curation of the sequenced reads revealed that
311 the 5' extension of exon 65, which defines the start of the *VLGR1a* transcript, was present in reads
312 across all sequenced samples. Notably, only the sample prepared following the PacBio long transcript
313 workflow contained reads that perfectly spanned from the 5' extension of exon 65 to exon 90,

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

324 In an effort to capture full-length *ADGRV1* (19.6 kb) transcripts encoding the longest isoform
325 *VLGR1b*, we employed the targeted enrichment approach using the Samplix Xdrop Sort system, by
326 utilizing three detecting sequences targeting the 5', middle, and 3' regions of *VLGR1b* transcripts.
327 qPCR results confirmed the enrichment of *ADGRV1* transcripts (Supplemental_Table_S4). While we
328 cannot conclude that the detection sequences were able to capture the full-length isoform, manual
329 curation of sequenced reads revealed that the 5' detection sequence (targeting exons 16-17)
330 provided the best coverage, spanning exons 3 to 77 and 80 to 90, with an average coverage of 818
331 reads covering the sequenced exons. This suggests the existence of the full-length *ADGRV1* transcript
332 isoform encoding *VLGR1b* in the human retina. Although the absence of reads mapping against exons
333 78-79 in the 5' enriched sample may hint at natural exon skipping, no reads were found that
334 indicated that exon 77 is immediately followed by exon 80. Furthermore, we did not observe any
335 indication for the skipping of exons 78-79 in the data obtained with the PacBio standard- and long
336 transcript workflows. Finally, data obtained with Samplix Xdrop indicate the presence of an even
337 shorter N-terminal transcript isoform that is similar to murine *Vlgr1e* (Figure S6).

338

339

340 **DISCUSSION**

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

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

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

366 identified two retinal *MYO7A* transcript isoforms differing in exon 35 length, with the truncated exon
367 35 variant being the predominant form (Gilmore *et al.*, 2023). This PCR-based targeted methodology
368 is incapable of identifying isoforms characterized by potentially novel transcription initiation and
369 termination sites. In contrast, the PacBio Iso-Seq allowed us to identify a novel *MYO7A* transcript
370 with an alternative 5' TSS. While Gilmore *et al.* (2023) have suggested that both previously known
371 *MYO7A* transcript isoforms should be considered when designing gene therapies, our findings
372 indicate that the isoforms with a novel 5' TSS were the most predominant in the human retina.
373 Notably, the variation in exon 35 length as described by Gilmore *et al.* is also found in this novel
374 isoform. While we do not dispute the relevance of this variation for therapy development, it might
375 be even more important to incorporate the novel TSS isoform we identified in the design of retinal
376 gene augmentation therapies. Furthermore, the novel, previously unannotated *MYO7A* 5' TSS should
377 be incorporated into diagnostic screening pipelines as it may harbor pathogenic variants of potential
378 clinical significance.

379 Additionally, sporadic intron retention events were observed across all 11 Usher syndrome-
380 associated genes, likely indicating ongoing splicing in these transcripts. The high frequency of intron
381 30 and 37 retention in the *MYO7A* transcripts, and the intron 4 retention in *WHRN* are particularly
382 interesting. While intron splicing does not follow a strict 5'-3' order or depend on intron length
383 (Pandya-Jones and Black, 2009; Singh and Padgett, 2009), Gazzoli *et al.* (2016) demonstrated that
384 inefficient multi-step splicing of introns can lead to the formation of "exon blocks" - groups of
385 consecutive exons that tend to be spliced together as a unit. The observed retention of introns 30
386 and 37 in the *MYO7A* transcripts may indicate the creation of exon blocks encompassing exons 31-
387 37. The interplay between intron retention and exon blocks may be relevant for the design of
388 antisense oligonucleotide (ASO) therapies as for example proposed for the *DMD* exon block 45-55 to
389 achieve exon skipping with fewer ASO molecules (Gazzoli *et al.*, 2016). Although multi exon skipping
390 does not seem particularly feasible for transcripts encoding myosin motor proteins such as *MYO7A*,
391 our findings illustrate the value of including partially spliced transcripts in the analysis of Iso-Seq

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

395 Another notable instance of intron retention was observed in *WHRN* transcripts, where
396 nearly half of the sequenced transcripts exhibited retention of intron 4. While intron retention is
397 generally thought to disrupt protein production due to the introduction of premature stop codons
398 that trigger nonsense-mediated decay (NMD), Boutz *et al.* (2015) described a class of intron-retaining
399 transcripts that are polyadenylated, sequestered in the nucleus, and resistant to NMD. These
400 retained introns may undergo subsequent splicing and export to the cytoplasm for translation,
401 potentially enabling cells to rapidly adapt to environmental changes or stress. Although we cannot
402 confirm the nuclear origin of *WHRN* intron 4 retaining transcripts in our dataset, their amplification
403 with an Oligo(dT) primer indicates they are polyadenylated mature mRNAs. Alternatively, these
404 intron 4-retaining transcripts may be translated into a truncated whirlin protein, containing only the
405 first two PDZ domains, comparable to the proposed murine N-terminal truncated protein isoforms
406 (Belyantseva *et al.*, 2005; Mathur *et al.*, 2015; Mburu *et al.*, 2003). Unfortunately, antibodies to
407 detect the N-terminus of whirlin are not available to determine if this transcript is translated.
408 Nevertheless, the high prevalence of these transcripts suggests their functional relevance, either
409 through encoding a truncated protein or via mechanisms akin to those described by Boutz *et al.*
410 (2015).

411

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

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

431

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

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

451

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

464

465 **MATERIAL AND METHODS**

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

467 Human donor eyes, deemed unsuitable for corneal transplantation, were used in this study. Written
468 consent was obtained from the donors' next of kin in accordance with the guidelines of the Italian
469 National Transplant Centre (Centro Nazionale Trapianti, Rome, Italy). The procurement and
470 processing of these tissues followed Italian law and adhered to the principles of the Declaration of
471 Helsinki and the guidelines of the European Eye Bank Association. Three human neural retinal
472 samples were obtained from non-visually impaired individuals through the Fondazione Banca degli
473 Occhi del Veneto (Venice, Italy). Prior to this study, we performed whole-genome sequencing on
474 DNA isolated from these retinal samples to screen for and exclude any known genetic variants
475 associated with inherited retinal disorders. The eyes were enucleated within 2-12 hours postmortem,
476 and the retinal extraction followed established protocols (Niyadurupola *et al.*, 2011; Osborne *et al.*,
477 2016). After cornea removal, the eyeball was dissected at the *ora serrata*; iris, lens, and vitreous
478 body were removed before carefully detaching the retina from sclera and retinal pigment epithelium
479 by cutting at the optic nerve head. Subsequently, retinal samples were snap-frozen in cryovials using
480 liquid nitrogen and shipped on dry ice. Detailed information about the donors is presented in
481 Supplemental_Table_S5.

482

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

484 To preserve the integrity of long mRNA molecules, all samples were handled with care during RNA
485 isolation and library preparation, avoiding vortexing or vigorous pipetting to prevent shearing of
486 mRNA molecules. Total RNA was extracted from the human neural retina samples by adding 500 μ L
487 of TRIzol reagent to each sample, which was then homogenized in a 2 mL tube containing a sterile
488 glass bead using a TissueLyser (QIAGEN, Aarhus, Denmark) in two cycles at 30 Hz for 30 seconds.
489 Following a 5-minute incubation at room temperature (RT), 100 μ L chloroform was added to the
490 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 μ L glycogen
492 (5 μ g/ μ l) 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 (Agilent 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

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

527

528 **PacBio long-read Iso-Seq data analysis**

529 Four PacBio long read-RNA Iso-Seq samples (three samples prepared following the PacBio standard
530 workflow, 1 sample prepared following the optimized PacBio long-transcript workflow) were
531 analyzed with IsoQuant (Prjibelski *et al.*, 2023). Three separate datasets were created (Figure 1): the
532 first dataset included the IsoQuant analysis of the three standard workflow samples, like we
533 previously performed in Riepe *et al.* (2024). The second dataset combined the analysis of PacBio
534 standard workflow retina samples 1, 2, and 3 with the retina sample prepared following the
535 optimized long transcript workflow. These two dataset used CCS3 reads. The third dataset consists
536 only of the long-transcript workflow sample, analyzed with less strict filtering settings and using CCS0
537 reads under the assumption that the longest transcripts might not reach CCS3 as frequently as
538 average-sized transcript. The sequencing data resulting from the PacBio Iso-Seq approaches, that
539 were used for IsoQuant analyses are available via European Genome-Phenome Archive (EGA) with
540 identifier EGAD5000000720. Detailed properties of the IsoQuant analyses can be found in the
541 Supplemental_Code document.

542

543

544 **Data analysis and transcript visualization**

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

550

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

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

566

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

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

596 cartridge was loaded with the sample and additional reagents in the following order: 5 μ l of Xdrop
597 Blank Oil Droplets (Samlipx apS, Birkerød, DK) were loaded in well #Out, 600 μ l of DE Sorting Buffer
598 (Samlipx apS, Birkerød, DK) (diluted in a 1:1 ratio using nuclease free water) was loaded in well #B1,
599 300 μ l of diluted DE Sorting Buffer was loaded in well #B2 and lastly 300 μ l of sample in staining
600 buffer was pipetted in well #In. The cartridge was sealed using a gasket and left to settle in the Xdrop
601 Sort for 5 minutes, after which the Sorting program was ran for the selected lanes. Thresholds for
602 detection and sorting were determined based on the displayed signal of droplets and background
603 fluorescence. As these values differed per sample, the guidelines for picking a threshold as
604 established in the Xdrop Sort Manual (Samlipx apS, Birkerød, DK) were consulted. Directly after
605 sorting, droplets were removed from the #Out well of the cartridge. After incubation for 5 minutes at
606 room temperature, the aqueous phase was removed in order to start the process of breaking the
607 droplets to prepare the DNA for the subsequent multiple displacement amplification. The droplets
608 were washed twice using 200 μ l of Droplet Sorting Wash Buffer (Samlipx apS, Birkerød, DK), after
609 which all the wash buffer was carefully removed. Subsequently, 20 μ l of Droplet Break Solution
610 (Samlipx apS, Birkerød, DK) was added as well as 1 μ l of Droplet Break Color (Samlipx apS, Birkerød,
611 DK). After vortexing, the clear break solution phase could be removed, leaving the colored aqueous
612 phase containing the enriched DNA.

613

614 Since the DNA amount after sorting was insufficient for direct sequencing, the captured transcripts
615 were amplified using multiple displacement amplification (MDA), employing non-sequence-specific
616 primers. Additionally, the Xdrop Sort was used to package single transcripts in single emulsion (SE)
617 droplets to avoid bias towards smaller transcripts. For this purpose, 10 μ l of enriched DNA from each
618 sample was transferred to separate PCR-tubes. Furthermore, 1 pg of unenriched cDNA dissolved in
619 10 μ l nuclease free water was taken along as a positive control, as well as a non-template control. To
620 each of the samples, 1 μ l of SE MDA enzyme (Samlipx apS, Birkerød, DK), 4 μ l of SE MDA mix (5x)
621 (Samlipx apS, Birkerød, DK) and 5 μ l of nuclease free water were added on ice. Subsequently, 20 μ l of

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

632

633 Enrichment validation was performed using qPCR to compare enriched samples with the original, un-
634 enriched cDNA pool. The amplified 1 pg of unenriched cDNA and non-template control from the
635 MDA served as positive and negative controls, respectively. qPCR primers, designed to target 100-
636 125 bp regions close to the targeted regions of the dPCR primers, were used in technical replicates to
637 negate bias from unintended target capture. Primer compositions and targeted exons are detailed in
638 Supplemental_Table_S6. Each qPCR reaction contained 1 μ l of template DNA, 10 μ l of Gotaq 2x
639 Master Mix (Promega Corporation, Madison, USA), 1 μ l of forward and reverse qPCR primers, and 7
640 μ l of nuclease-free water. Reactions were run on a QuantStudio 3 Real-Time PCR System (Thermo
641 Fisher Scientific, Waltham, MA, USA) with the following program: 50°C for 1 second, 95°C for 10
642 minutes, 40 cycles of 95°C for 15 seconds and 60°C for 30 seconds, then 95°C for 15 seconds, 60°C for
643 1 minute, and finally 95°C for 15 seconds.

644

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

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

648 with nuclease free water to a total volume of 25.5 μ l. To each sample, 3 μ l NEBuffer 2 (New England
649 Biolabs, Ipswich, MA, USA) and 1.5 μ l of T7 Endonuclease (New England Biolabs, Ipswich, MA, USA)
650 was added before incubation at 37°C for 15 minutes in a thermal cycler. Afterwards, 20 μ l of TE
651 buffer (pH8) was added to each of the samples and a custom bead cleanup was performed. To do so,
652 40 μ l of Ampure PB beads were resuspended and pelleted on a magnet. After the supernatant had
653 been completely removed the beads were washed with nuclease free water twice and resuspended
654 in an equal volume of the custom bead buffer consisting of the following components: 20 μ l of 1M
655 Tris-HCL (Sigma-Aldrich, Saint Louis, USA), 4 μ l 0.5M EDTA pH8 (Thermo Fisher Scientific, Waltham,
656 MA, USA), 640 μ l 5M NaCl (Thermo Fisher Scientific, Waltham, MA, USA), 550 μ l 40% PEG8000
657 (Sigma-Aldrich, Saint Louis, USA) and 778 μ l nuclease free water. Of this custom bead suspension, 35
658 μ l was added to each sample and incubated at room temperature for 20 minutes. Subsequently, the
659 samples were pelleted on a magnet and the supernatant was removed. Samples were then washed
660 with 70% ethanol and dried for 30 seconds. Afterwards, samples were removed from the magnetic
661 rack and resuspended in nuclease free water before being incubated for 1 minute at 50°C and 5
662 minutes at room temperature. After spinning the samples down, the beads were pelleted on the
663 magnetic rack until the eluate was clear and the purified DNA could be collected. A total of 500 ng of
664 MDA-amplified and purified DNA was used to create multiplexed SMRTbell amplicon libraries,
665 following the manufacturer's instructions for the SMRTbell prep kit 3.0 102-359-000 (Pacific
666 Biosciences, California, USA). If the sample had less than 500 ng of DNA, linearized plasmids not
667 containing the region of interest were added to spike the samples. The samples were then prepared
668 for sequencing using the Sequel IIe Binding Kit 3.2 (Pacific Biosciences, California, USA), following the
669 recommended protocols from SMRT Link 11.0.0.146107. Finally, 115 μ l of the final mix was loaded
670 per well and long read sequencing was performed using the Sequel IIe system (Pacific Biosciences,
671 California, USA). Following sequencing, the reads were processed following a cDNA analysis:
672 subreads were demultiplexed using lima V.2.5.0, and combined to create a consensus sequence
673 using CCS V.6.3.0. These reads were filtered RQ 0.99 to obtain HiFi reads, and HiFi reads were then

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

678

679 **Tissue collection for Oxford Nanopore Technology (ONT) sequencing**

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

691

692 **RNA isolation and Oxford Nanopore Technology (ONT) sequencing**

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

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

701

702

703 **ONT data analysis of selected genes**

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

709

710 **Relative expression of *MYO7A* and *WHRN* isoforms**

711 Quantitative PCR analysis was conducted to determine the relative expression levels of the different
712 *MYO7A*- and *WHRN* transcript isoforms identified in the PacBio Iso-Seq datasets across three human
713 neural retina samples. The same RNA used for preparing the PacBio libraries was employed as the
714 template, with 250 ng of RNA being used for cDNA synthesis using the SuperScript IV Reverse
715 Transcriptase kit (Thermo Fisher Scientific, Waltham, MA, USA, #18090200). Quantitative PCR
716 analysis was performed using GoTaq qPCR Master Mix (Promega), following the manufacturer's
717 protocol. For *MYO7A*, transcript-specific primers were designed and validated to target the 5'
718 canonical start site, the 5' alternative start site, and the 3' end of the identified *MYO7A* isoforms, as
719 well as the reference gene *GUSB*. Amplifications were carried out using the QuantStudio 3 Real-Time
720 PCR system (Applied Biosystems, Waltham, MA, USA), with PCR reactions performed in triplicate.
721 Relative gene expression levels compared to the reference gene *GUSB* were determined using the
722 $2^{-\Delta Ct}$ method.

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

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

730

731 **Analysis of protein isoforms**

732 To predict the 2D protein domain architecture of encoded proteins in silico analyses were performed
733 using the SMART online tool (Letunic *et al.*, 2021). 3D protein structures were modelled using
734 AlphaFold (Jumper *et al.*, 2021) using the Google Colab notebook (v1.5.3) with standard settings. The
735 3D structures were visualized with YASARA (Krieger and Vriend, 2014), and structural alignments
736 were conducted using the MUSTANG algorithm (Konagurthu *et al.*, 2006).

737

738 **DATA ACCESS**

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

747

748 **COMPETING INTEREST STATEMENT**

749 The authors have no competing interest.

750

751 **ACKNOWLEDGEMENTS**

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

758

759 **AUTHOR CONTRIBUTIONS**

760 MS: Investigation, Visualization, Methodology, Project administration, Writing – original draft. TR:
761 Formal Analysis, Data curation, Writing–review and editing. NZ: Investigation, Methodology, Writing–
762 review and editing. RS: Formal Analysis, Writing–review and editing. MK: Investigation, Writing–
763 review and editing. JO: Investigation, Writing–review and editing. RT: Formal Analysis, Writing–
764 review and editing. BF: Resources, Writing–review and editing. SF: Resources, Writing–review and
765 editing. ADR: Resources, Writing–review and editing. ED: Resources, Writing–review and editing. SB:
766 Formal Analysis, Writing–review and editing. HK: Funding acquisition, Writing–review and editing. SR:
767 Writing–review and editing. FC: Resources, Writing–review and editing. AH: Funding acquisition,
768 Methodology, Writing–review and editing. FPMC: Conceptualization, Funding acquisition, Writing–
769 review and editing. P't H: Conceptualization, Funding acquisition, Writing–review and editing. EW:
770 Funding acquisition, Supervision, Conceptualization, Writing–review and editing. EV: Funding
771 acquisition, Conceptualization, Methodology, Supervision, Writing–review and editing.

772

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 \pm 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 \pm 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 \pm 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-containing 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.

831

832 **Figure 5: *USH2A* transcript isoforms identified by IsoQuant analysis, manual curation and and Samplix Xdrop**833 **targeted enrichment. A.** The GENCODE reference transcript is depicted at the top in green, followed by the834 known human *USH2A* transcript isoforms in blue (Van Wijk *et al.*, 2004). The *USH2A* IsoQuant transcripts are

835 depicted in red. The light green, blue and red colors indicate the untranslated regions (UTR) and the dark

836 green, blue and red colors indicate the open reading frame (ORF) of each transcript. Differences between the

837 IsoQuant transcript isoforms and the GENCODE reference transcript are highlighted in grey boxes. **B.** Relative838 expression of *USH2A* isoforms based on literature in either the retina or the cochlea. **C.** The Transcripts Per839 Million (based on dataset 1) for each IsoQuant transcript are presented for the three individual samples. **D.**840 Proposed *USH2A* transcript isoforms based on manual curation and and Samplix Xdrop targeted enrichment.841 The GENCODE reference transcript is depicted in green, followed by the proposed *USH2A* transcript isoforms

842 and events based on manual curation of BAM-files using the Integrative Genomics Viewer (IGV) in red, and the

843 proposed transcript isoform following the Samplix Xdrop targeted enrichment in orange. The light green, red

844 and orange colors indicate the untranslated regions (UTR) and the dark green, red and orange colors indicate

845 the open reading frame (ORF) of each transcript. Differences between the proposed transcript isoforms and

846 the GENCODE reference transcript are highlighted in grey boxes. The overview of sporadic incorporation of

847 cryptic exons indicates the presence of PE8 and PE20 as previously described by Reurink *et al.* (2023).

848 Additionally, locations where cryptic exons are occasionally incorporated at sites that are not yet associated to

849 deep-intronic pathogenic variants are indicated with black arrows.

850

851 **Figure 6: *ADGRV1* proposed transcript isoforms from manual curation and Samplix Xdrop targeted**852 **enrichment.** The GENCODE reference transcript is depicted at the top in green, followed by the *ADGRV1*

853 proposed retinal transcript isoforms and events based on manual curation of BAM-files using the Integrative

854 Genomics Viewer (IGV) in red, and proposed transcript isoforms following the Samplix Xdrop targeted

855 enrichment in orange. The light green, red and orange colors indicate the untranslated regions (UTR) and the

856 dark green, red and orange colors indicate the open reading frame (ORF) of each transcript. Differences

857 between the proposed transcript isoforms and the GENCODE reference transcript are highlighted in grey box.

858

859

REFERENCES

- 860 Abad-Morales V, Navarro R, Bures-Jelstrup A, Pomares E. 2020. Identification of a novel
861 homozygous ARSG mutation as the second cause of Usher syndrome type 4. *Am J*
862 *Ophthalmol Case Rep* 19:100736.
- 863 Adato A, Vreugde S, Joensuu T, Avidan N, Hamalainen R, Belenkiy O, Olender T, Bonne-
864 Tamir B, Ben-Asher E, Espinos C et al. . 2002. USH3A transcripts encode clarin-1, a
865 four-transmembrane-domain protein with a possible role in sensory synapses. *Eur J*
866 *Hum Genet* 10(6):339-50.
- 867 Ahmed ZM, Riazuddin S, Bernstein SL, Ahmed Z, Khan S, Griffith AJ, Morell RJ, Friedman
868 TB, Riazuddin S, Wilcox ER. 2001. Mutations of the protocadherin gene PCDH15
869 cause Usher syndrome type 1F. *Am J Hum Genet* 69(1):25-34.
- 870 Belyantseva IA, Boger ET, Naz S, Frolenkov GI, Sellers JR, Ahmed ZM, Griffith AJ,
871 Friedman TB. 2005. Myosin-XVa is required for tip localization of whirlin and
872 differential elongation of hair-cell stereocilia. *Nat Cell Biol* 7(2):148-56.
- 873 Bolz H, von Brederlow B, Ramirez A, Bryda EC, Kutsche K, Nothwang HG, Seeliger M, del
874 CSCM, Vila MC, Molina OP et al. . 2001. Mutation of CDH23, encoding a new
875 member of the cadherin gene family, causes Usher syndrome type 1D. *Nat Genet*
876 27(1):108-12.
- 877 Booth KT, Kahrizi K, Babanejad M, Daghigh H, Bademci G, Arzhanghi S, Zareabdollahi D,
878 Duman D, El-Amraoui A, Tekin M et al. . 2018. Variants in CIB2 cause DFNB48 and
879 not USH1J. *Clin Genet* 93(4):812-821.
- 880 Boutz PL, Bhutkar A, Sharp PA. 2015. Detained introns are a novel, widespread class of post-
881 transcriptionally spliced introns. *Genes Dev* 29(1):63-80.
- 882 Braun TA, Mullins RF, Wagner AH, Andorf JL, Johnston RM, Bakall BB, Deluca AP,
883 Fishman GA, Lam BL, Weleber RG et al. . 2013. Non-exomic and synonymous
884 variants in ABCA4 are an important cause of Stargardt disease. *Hum Mol Genet*
885 22(25):5136-45.
- 886 Cao H, Wu J, Lam S, Duan R, Newnham C, Molday RS, Graziotto JJ, Pierce EA, Hu J. 2011.
887 Temporal and tissue specific regulation of RP-associated splicing factor genes PRPF3,
888 PRPF31 and PRPC8--implications in the pathogenesis of RP. *PLoS One* 6(1):e15860.
- 889 Ciampi L, Mantica F, Lopez-Blanch L, Permanyer J, Rodriguez-Marin C, Zang J, Cianferoni
890 D, Jimenez-Delgado S, Bonnal S, Miravet-Verde S et al. . 2022. Specialization of the
891 photoreceptor transcriptome by Srm3-dependent microexons is required for outer
892 segment maintenance and vision. *Proc Natl Acad Sci U S A* 119(29):e2117090119.
- 893 de Bruijn SE, Rodenburg K, Corominas J, Ben-Yosef T, Reurink J, Kremer H, Whelan L,
894 Plomp AS, Berger W, Farrar GJ et al. . 2023. Optical genome mapping and revisiting
895 short-read genome sequencing data reveal previously overlooked structural variants
896 disrupting retinal disease-associated genes. *Genet Med* 25(3):100345.
- 897 Ebermann I, Scholl HP, Charbel Issa P, Becirovic E, Lamprecht J, Jurklics B, Millan JM,
898 Aller E, Mitter D, Bolz H. 2007. A novel gene for Usher syndrome type 2: mutations
899 in the long isoform of whirlin are associated with retinitis pigmentosa and
900 sensorineural hearing loss. *Hum Genet* 121(2):203-11.
- 901 Ebrahim S, Ingham NJ, Lewis MA, Rogers MJC, Cui R, Kachar B, Pass JC, Steel KP. 2016.
902 Alternative Splice Forms Influence Functions of Whirlin in Mechanosensory Hair Cell
903 Stereocilia. *Cell Rep* 15(5):935-943.

- 904 Eudy JD, Weston MD, Yao S, Hoover DM, Rehm HL, Ma-Edmonds M, Yan D, Ahmad I,
905 Cheng JJ, Ayuso C et al. . 1998. Mutation of a gene encoding a protein with
906 extracellular matrix motifs in Usher syndrome type IIa. *Science* 280(5370):1753-7.
- 907 Gazzoli I, Pulyakhina I, Verwey NE, Ariyurek Y, Laros JF, t Hoen PA, Aartsma-Rus A.
908 2016. Non-sequential and multi-step splicing of the dystrophin transcript. *RNA Biol*
909 13(3):290-305.
- 910 Gilmore WB, Hultgren NW, Chadha A, Barocio SB, Zhang J, Kutsyr O, Flores-Bellver M,
911 Canto-Soler MV, Williams DS. 2023. Expression of two major isoforms of MYO7A
912 in the retina: Considerations for gene therapy of Usher syndrome type 1B. *Vision Res*
913 212:108311.
- 914 Gustavsson EK, Zhang D, Reynolds RH, Garcia-Ruiz S, Ryten M. 2022. ggtranscript: an R
915 package for the visualization and interpretation of transcript isoforms using ggplot2.
916 *Bioinformatics* 38(15):3844-3846.
- 917 Hasson T, Heintzelman MB, Santos-Sacchi J, Corey DP, Mooseker MS. 1995. Expression in
918 cochlea and retina of myosin VIIa, the gene product defective in Usher syndrome type
919 1B. *Proc Natl Acad Sci U S A* 92(21):9815-9.
- 920 Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K,
921 Bates R, Zidek A, Potapenko A et al. . 2021. Highly accurate protein structure
922 prediction with AlphaFold. *Nature* 596(7873):583-589.
- 923 Konagurthu AS, Whisstock JC, Stuckey PJ, Lesk AM. 2006. MUSTANG: a multiple
924 structural alignment algorithm. *Proteins* 64(3):559-74.
- 925 Krieger E, Vriend G. 2014. YASARA View - molecular graphics for all devices - from
926 smartphones to workstations. *Bioinformatics* 30(20):2981-2.
- 927 Letunic I, Khedkar S, Bork P. 2021. SMART: recent updates, new developments and status in
928 2020. *Nucleic Acids Res* 49(D1):D458-D460.
- 929 Li H. 2021. New strategies to improve minimap2 alignment accuracy. *Bioinformatics*
930 37(23):4572-4574.
- 931 Li S, Mecca A, Kim J, Caprara GA, Wagner EL, Du TT, Petrov L, Xu W, Cui R, Rebutini IT
932 et al. . 2020. Myosin-VIIa is expressed in multiple isoforms and essential for
933 tensioning the hair cell mechanotransduction complex. *Nat Commun* 11(1):2066.
- 934 Linnert J, Knapp B, Guler BE, Boldt K, Ueffing M, Wolfrum U. 2023. Usher syndrome
935 proteins ADGRV1 (USH2C) and CIB2 (USH1J) interact and share a common
936 interactome containing TRiC/CCT-BBS chaperonins. *Front Cell Dev Biol*
937 11:1199069.
- 938 Liu MM, Zack DJ. 2013. Alternative splicing and retinal degeneration. *Clin Genet* 84(2):142-
939 9.
- 940 Liu X, Vansant G, Udovichenko IP, Wolfrum U, Williams DS. 1997. Myosin VIIa, the
941 product of the Usher 1B syndrome gene, is concentrated in the connecting cilia of
942 photoreceptor cells. *Cell Motil Cytoskeleton* 37(3):240-52.
- 943 Madsen EB, Hoijer I, Kvist T, Ameer A, Mikkelsen MJ. 2020. Xdrop: Targeted sequencing
944 of long DNA molecules from low input samples using droplet sorting. *Hum Mutat*
945 41(9):1671-1679.
- 946 Mathur PD, Zou J, Zheng T, Almishaal A, Wang Y, Chen Q, Wang L, Vashist D, Brown S,
947 Park A et al. . 2015. Distinct expression and function of whirlin isoforms in the inner
948 ear and retina: an insight into pathogenesis of USH2D and DFNB31. *Hum Mol Genet*
949 24(21):6213-28.
- 950 Mburu P, Mustapha M, Varela A, Weil D, El-Amraoui A, Holme RH, Rump A, Hardisty RE,
951 Blanchard S, Coimbra RS et al. . 2003. Defects in whirlin, a PDZ domain molecule
952 involved in stereocilia elongation, cause deafness in the whirler mouse and families
953 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, Stemerink 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, Hjelmén 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, Lujendijk 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		Corresponding figures
		Identified Ensembl Spliced Transcripts (ENST)	Validated novel events and previously unidentified transcripts***	
<i>MYO7A</i>	USH1B	ENST00000409709.9 (MANE) ENST00000458637.6	Alternative transcription start site (transcript20055.Chr11.nic)	Figure 3
<i>USH1C</i>	USH1C	ENST00000527020.5 ENST00000318024.9 ENST00000526313.5		Figure S1
<i>CDH23</i>	USH1D	ENST00000224721.12* ENST00000461841.7	Novel in frame exon 11A and skipping of micro exon 12 (transcript11235.Chr10.nnic) Exon 69 skipping	Figure S2
<i>PCDH15</i>	USH1F	ENST00000475158.1 ENST00000644397.2 (MANE) ENST00000373957.7 ENST00000621708.4 ENST00000373955.5		Figure S3
<i>SANS</i>	USH1G	ENST00000614341.5 (MANE)		Figure S4
<i>CIB2</i>	USH1J [#]	-		Figure S5
<i>USH2A</i>	USH2A	ENST00000307340.8 (MANE)**	5' UTR splice events (transcript51429.Chr1.nnic; transcript51430.Chr1.nnic; transcript51439.Chr1.nnic)	Figure 5
<i>ADGRV1</i>	USH2C	ENST00000366942.3 ENST00000638316.1 ENST00000639884.1 ENST00000640109. ENST00000640281.1		Figure 6, S6
<i>WHRN</i>	USH2D	ENST00000362057.4 (MANE) ENST00000374057.3 ENST00000265134.10	Inclusion of novel exon 7B (transcript13724.Chr9.nnic) Intron 4 retention (transcript13718.Chr9.nnic)	Figure 4
<i>CLRN1</i>	USH3A	ENST00000327047.6 (MANE) ENST00000472224.1		Figure S7
<i>ARSG</i>	USH4	ENST00000448504.6 ENST00000578726.1		Figure S8

* 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).

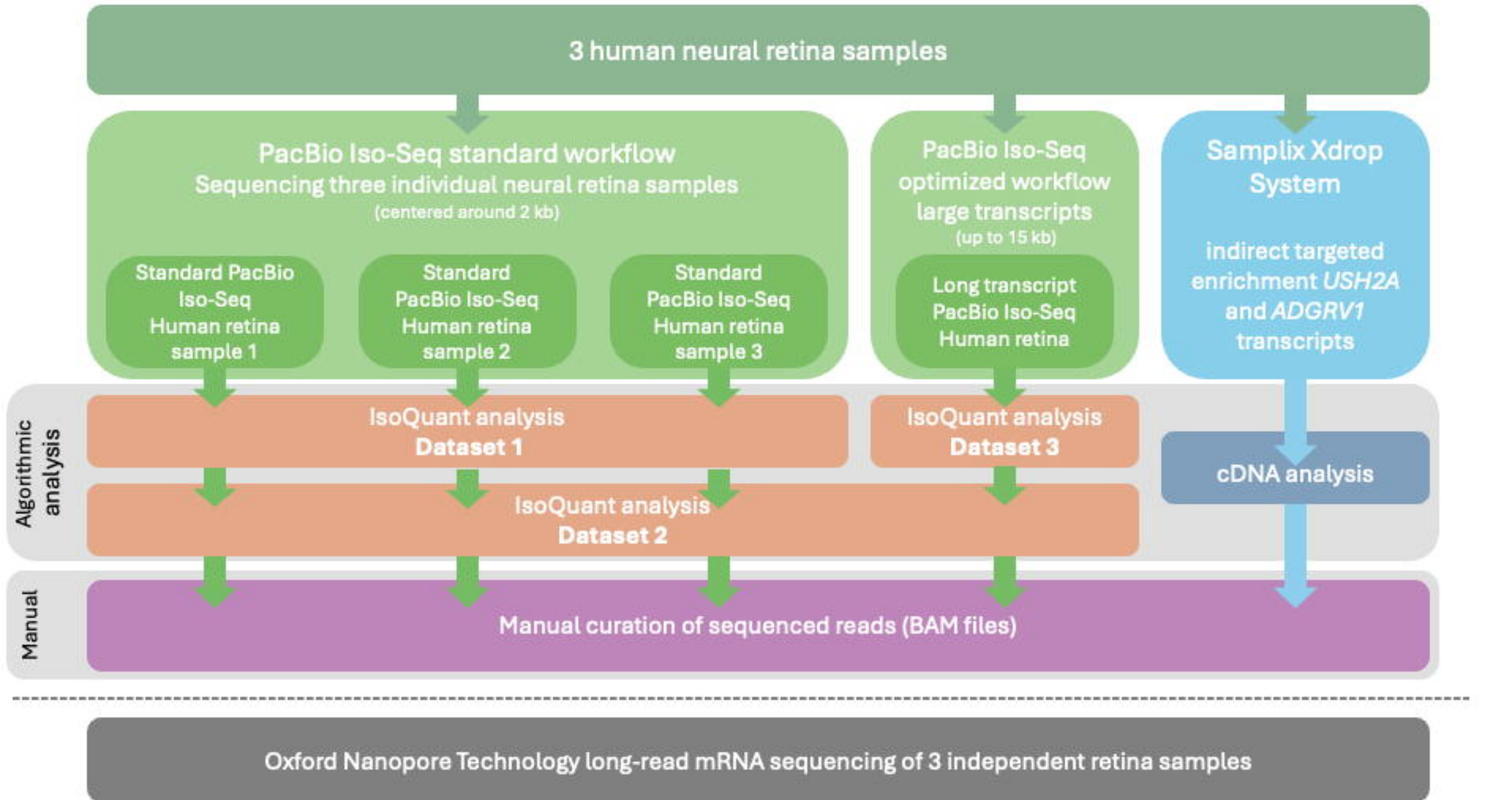
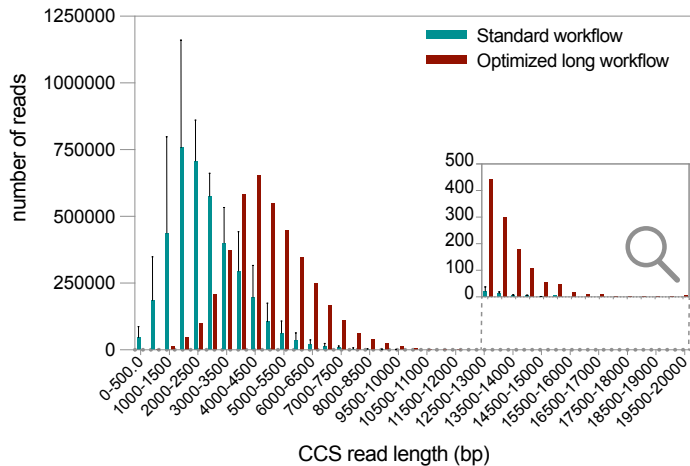
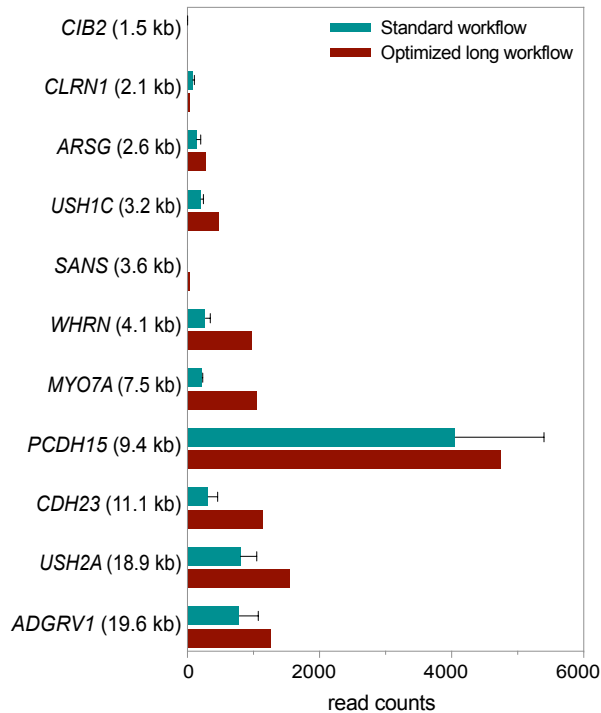


Figure 2

A



B



C

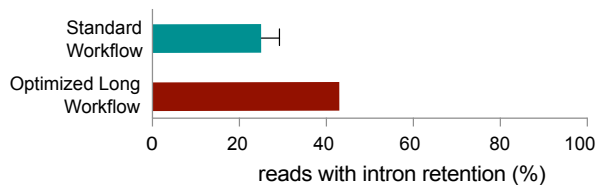
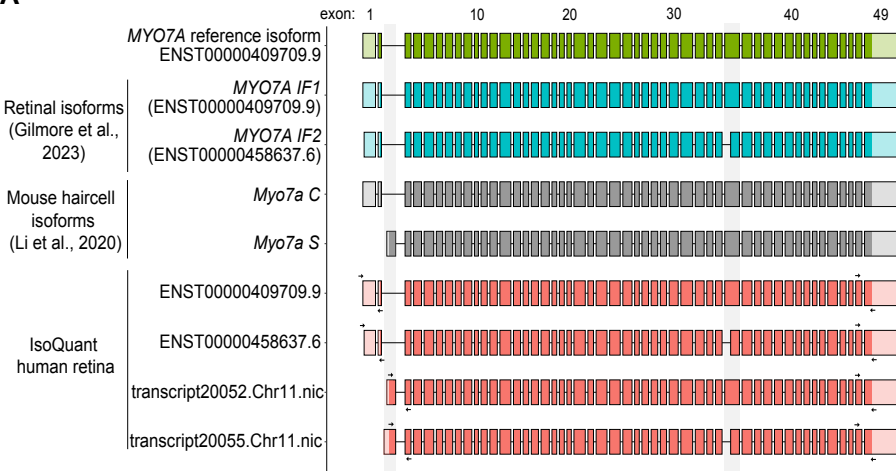
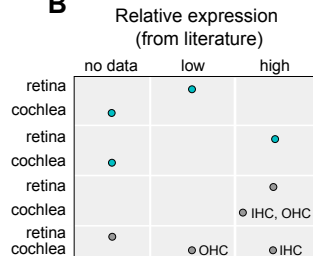


Figure 3

A

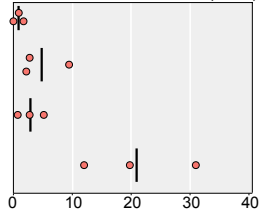


B



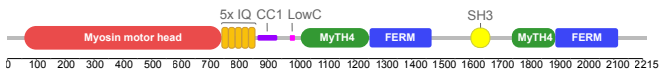
C

MYO7A in human retina (TPM)

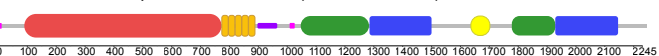


D

MYO7A ENST00000409709.9 (Canonical start)

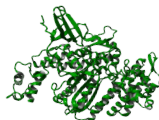


MYO7A transcript20052.Chr11.nic (Alternative start)

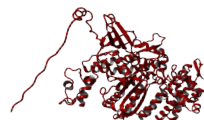


E

MYO7A canonical
5' region
(aa 1 - 742)



MYO7A alternative
5' region
(aa 1 - 772)



F

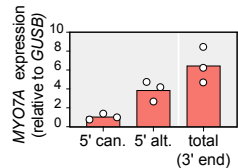


Figure 4

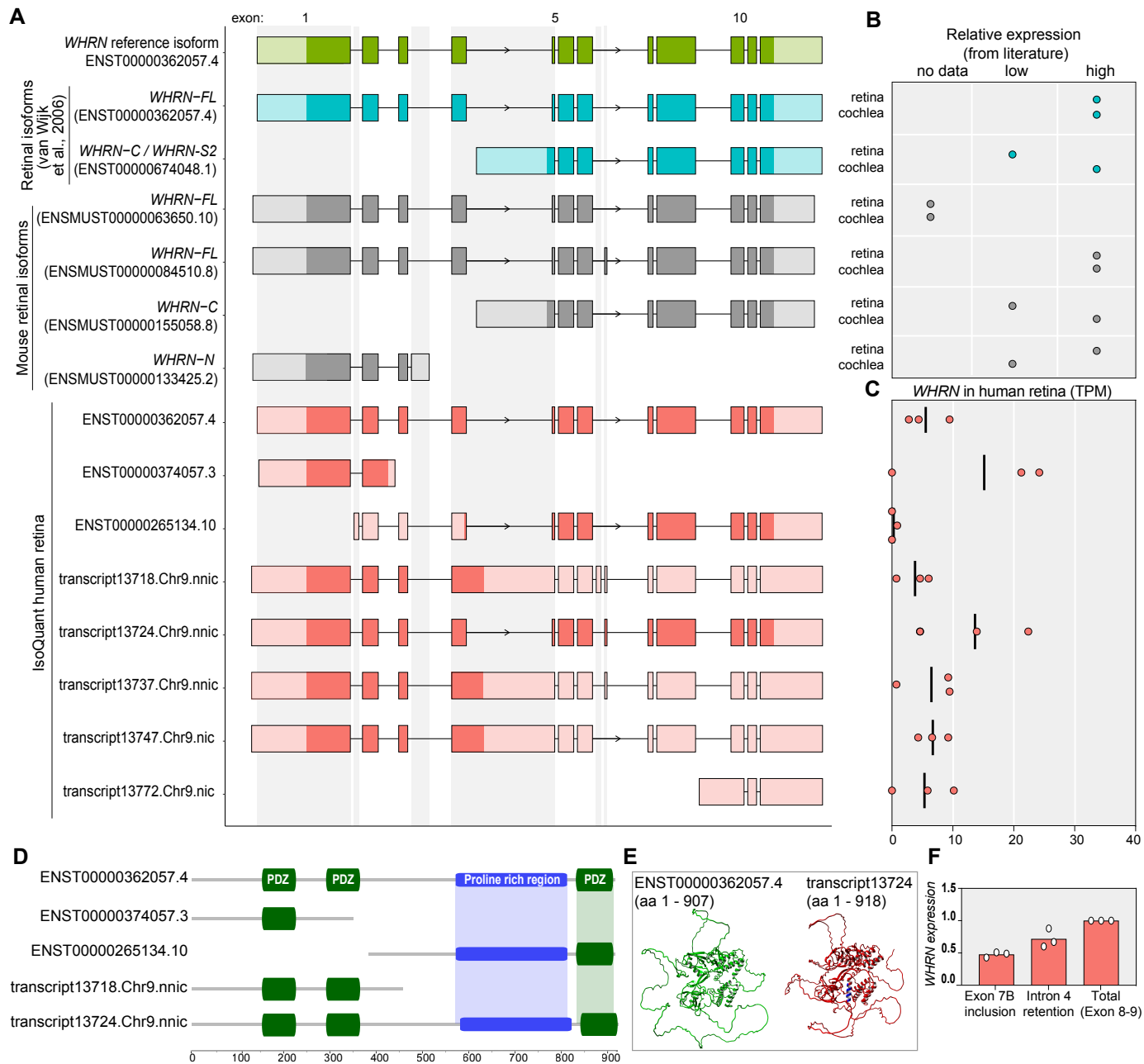


Figure 5

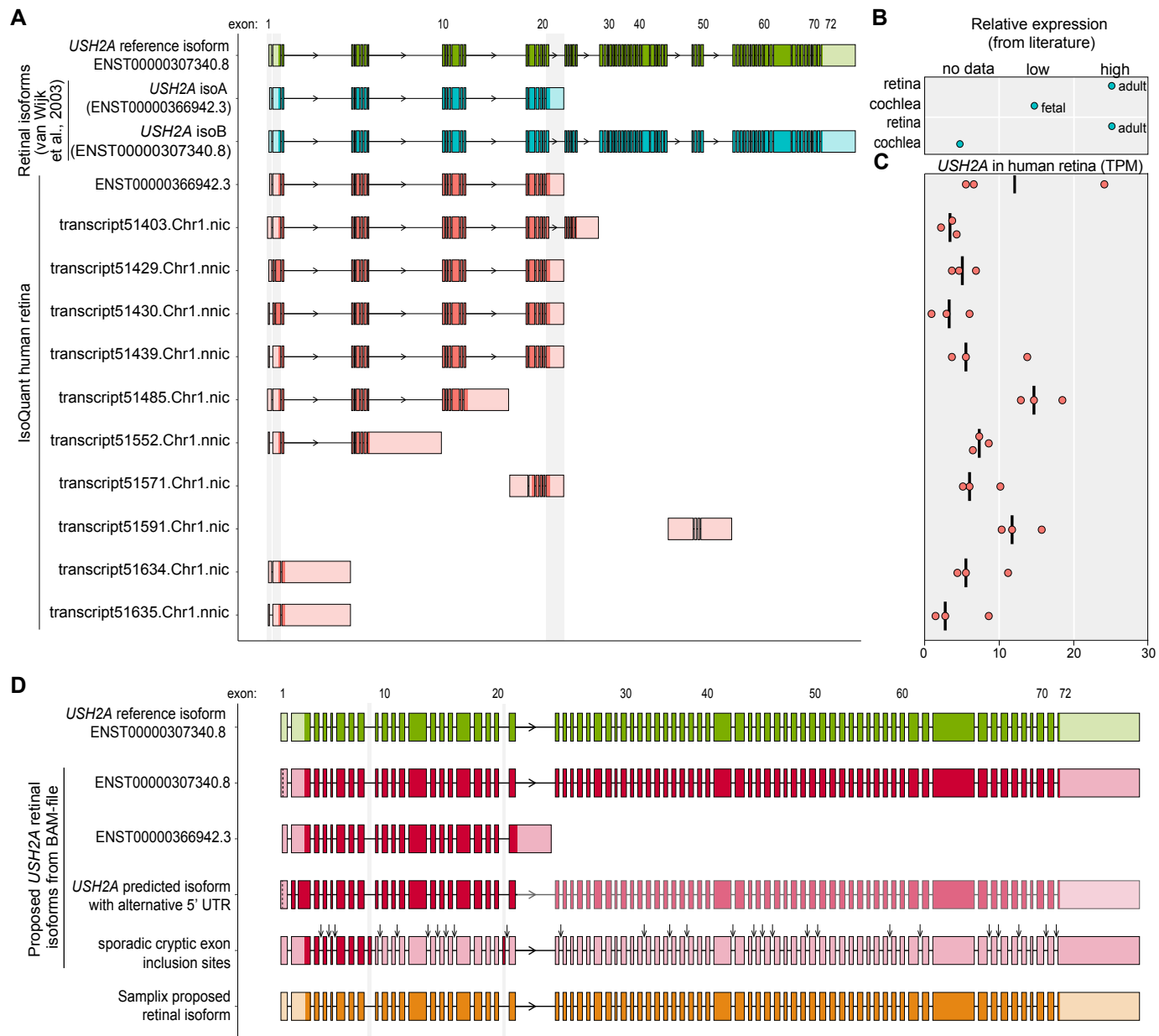


Figure 6

