A long read optimized *de novo* transcriptome pipeline reveals novel ocular developmentally regulated gene isoforms and disease targets

Vinay S. Swamy[1], Temesgen D. Fufa[2], Robert B. Hufnagel[2], and David M. McGaughey[1,]

[1] Bioinformatics Group, Ophthalmic Genetics & Visual Function Branch, National Eye Institute, Institutes of Health

[2] Medical Genetics and Ophthalmic Genomics Unit, National Eye Institute, National Institutes of Health

Correspondence: David M. McGaughey <mcgaugheyd@mail.nih.gov>

Abstract

De novo transcriptome construction from short-read RNA-seq is a common method for reconstructing mRNA transcripts within a given sample. However, the precision of this process is unclear as it is difficult to obtain a ground-truth measure of transcript expression. With advances in third generation sequencing, full length transcripts of whole transcriptomes can be accurately sequenced to generate a ground-truth transcriptome. We generated long-read PacBio and short-read Illumina RNA-seq data from a human induced pluripotent stem cell- derived retinal pigmented epithelium (iPSC-RPE) cell line. We use long-read data to identify simple metrics for assessing *de novo* transcriptome construction and optimize a short-read based *de novo* transcriptome construction pipeline. We apply this this pipeline to construct transcriptomes for 340 short-read RNA-seq samples originating from healthy adult and fetal human retina, cornea, and RPE. We identify hundreds of novel gene isoforms and examine their significance in the context of ocular development and disease.

Introduction

The transcriptome is defined as the set of unique RNA transcripts expressed in a biological system. A single gene can have multiple distinct transcripts, or isoforms, and there are multiple biological processes that drive the formation of these isoforms including alternative promoter usage, alternative splicing, and alternative polyadenylation. Gene isoforms can have distinct and critical functions in biological processes like development, cell differentiation, and cell migration (Dykes et al., 2018), (Trapnell et al., 2010), (Mitra et al., 2020). Alternative usage of isoforms has also been implicated in multiple diseases including cancer, cardiovascular disease, Alzheimer's disease and diabetic retinopathy (Vitting-Seerup and Sandelin, 2017), (Neagoe Ciprian et al., 2002), (Mills et al., 2013), (Perrin et al., 2005).

Accurate annotation of gene isoforms is fundamental for understanding their biological impact. For example, while the Gencode human comprehensive transcript annotation (release 28) contains 82335 protein coding and 121500 noncoding transcripts across 19901 genes and 38480 pseudogenes, but this annotation is incomplete (Frankish et al., 2019), (Zhang et al., 2020). Some of the first high throughput methods to find novel gene isoforms used short-read (~100bp) RNA-seq to identify novel exon-exon junctions and novel exon boundaries based soley on RNA-seq coverage (Nagalakshmi et al., 2008). More recently, several groups have developed specialized tools to use RNA-seq to reconstruct the whole transcriptome of a biological sample, dubbed *de novo* transcriptome construction (Haas et al., 2013),(Trapnell et al., 2010), (Pertea et al., 2015).

De novo transcriptome construction uses short-read RNA-seq to reconstruct full-length mRNA transcripts. However, a large number of samples are necessary to overcome the noise and short-read lengths of this type of data. Because of increasingly inexpensive sequencing cost, datasets of the necessary size are now available. For example, one of the most comprehensive *de novo* transcriptome projects to date is CHESS, which uses the GTEx data set to construct *de novo* transcriptomes in over 9000 RNA-seq samples from 44 distinct body locations to create a comprehensive annotation of mRNA transcripts across the human body (GTEx Consortium et al., 2017), (Pertea et al., 2018). However, since the

54 GTEx dataset does not include samples from any ocular tissues, the CHESS database
55 remains an incomplete annotation of the human transcriptome.

56 Despite the increasing number of tools developed, there is no gold standard to
57 evaluate the precision and sensitivity of *de novo* transcriptome construction on real (not
58 simulated) biological data. Long-read sequencing technologies provide a potential solution
59 to this problem as long-read sequencing can capture full length transcripts and thus, can be
60 used to identify a more comprehensive range of gene isoforms. While previous iterations of
61 long-read sequencing technologies typically had higher error rates, the new PacBio Sequel
62 II system sequences long-reads as accurately as short-read based sequencing (Wenger et
63 al., 2019).

64 We propose that long-read based transcriptomes can serve as a ground truth for
65 evaluating short-read based transcriptomes. In this study, we used PacBio long-read RNA
66 sequencing to inform the construction of short-read transcriptomes. We generated PacBio
67 long-read RNA-seq along with matched Illumina short-read RNA-seq data from a human
68 induced pluripotent stem cell (iPSC)-differentiated retinal pigmented epithelium (RPE) cell
69 line. We then designed a rigorous StringTie-based pipeline that maximizes the concordance
70 between short and long-read *de novo* transcriptomes.

71 Finally, we applied this optimized pipeline to a data set containing 340 human
72 ocular tissue samples compiled from mining previously published, publicly available short-
73 read RNA-seq data (Swamy and McGaughey, 2019). We built transcriptomes for three
74 major ocular tissues: cornea, retina, and RPE, using RNA-seq data from both adult and fetal
75 tissues to create a high-quality pan-eye transcriptome. In addition to ocular samples, we
76 used a subset of the GTEx data set to construct transcriptomes for tissues in 44 other
77 locations across the body.

78 We used our gold-standard informed pan-eye *de novo* transcriptome to reveal
79 hundreds of novel gene isoforms in the eye and analyze their potential impact on ocular
80 biology and disease. We provide transcript annotation derived from our *de novo*
81 transcriptomes as a resource to other researchers through an R package.

Results

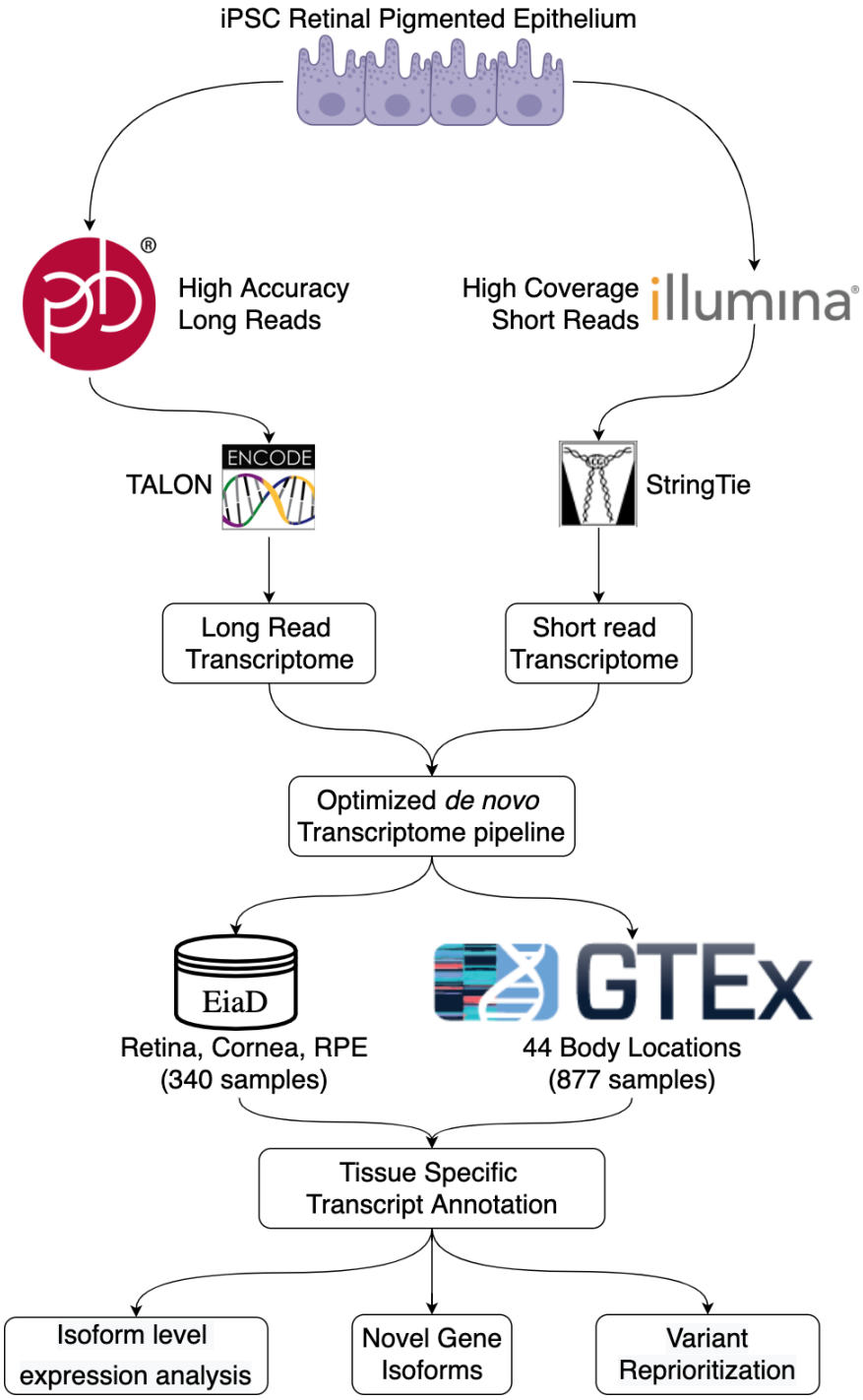


Figure 1. Workflow for long-read informed *de novo* transcriptome construction and analysis

Long-read PacBio RNA sequencing guides short-read *de novo* transcriptome construction

To evaluate the accuracy of short-read transcriptome construction, we first generated PacBio long-read RNA-seq data and Illumina short-read RNA-seq data from iPSC-RPE (Fig 1). These cells were differentiated using an optimized protocol, and thus minimal biological variation is expected (Blenkinsop et al., 2015), (Maruotti et al., 2015). We used these sequencing data to construct a long-read transcriptome and a short-read transcriptome. In our long-read transcriptome we found 1163239 distinct transcripts, and in our short-read transcriptome 366888 distinct transcripts

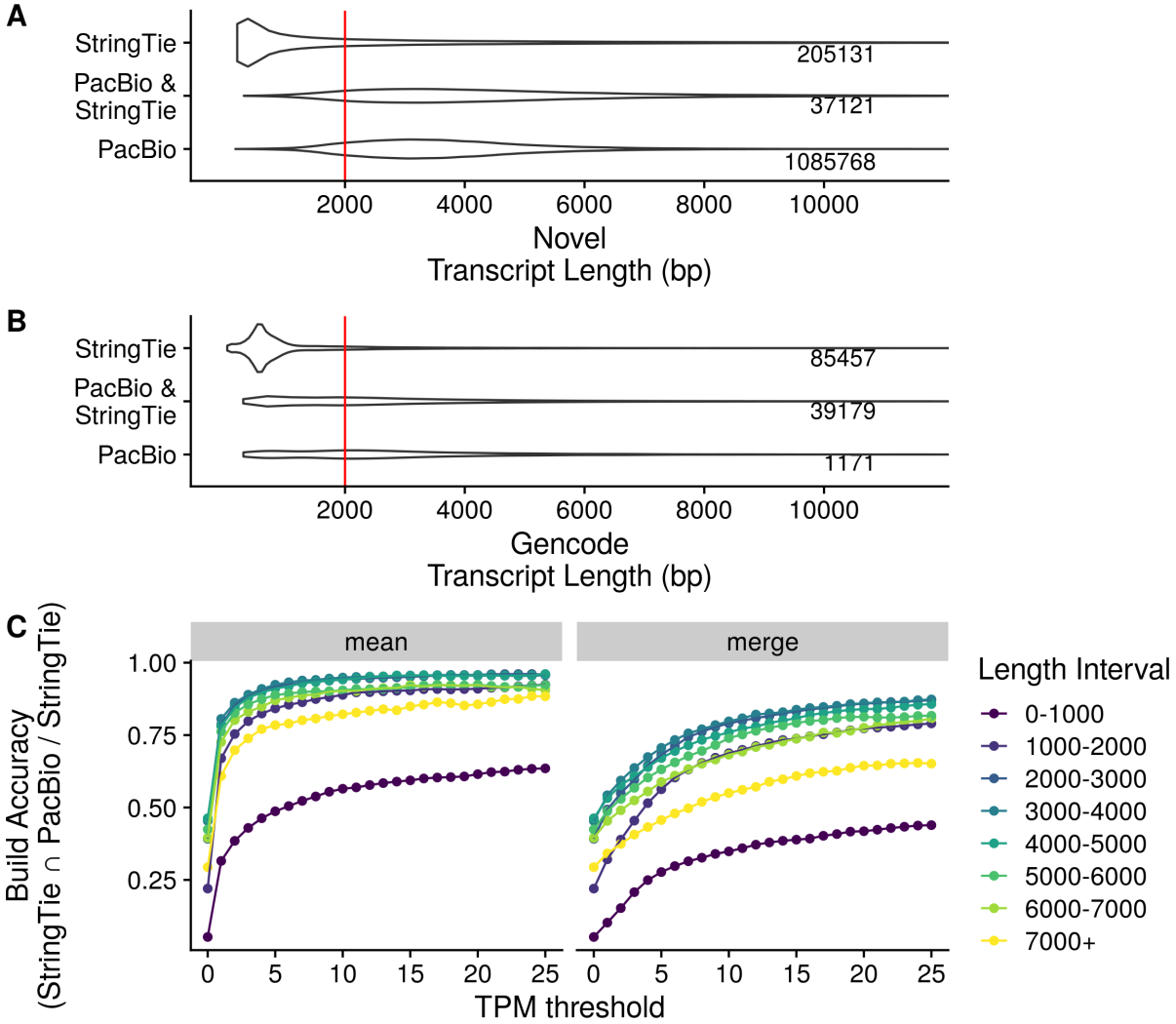


Figure 2. Transcript length and expression dictate transcriptome construction accuracy. A,B) Distributions of novel(A) and previously annotated(B) transcript lengths between PacBio (long-read) and Stringtie (short-read) transcriptomes. Each distribution is labeled with the total number of transcripts in the distribution C) short-read construction accuracy stratified by transcript length at different Transcripts Per Million (TPM)-based transcript exclusion thresholds. The "merge" method follows the protocol for constructing transcriptomes outlined by the StringTie authors and keeps any transcripts expressed above a specific TPM threshold in at least one samples. The "mean" method used by our pipeline keeps transcripts whose average expression across all samples is above a specific TPM threshold.

In our initial comparison between short and long-read transcriptomes, we noticed a low transcriptome construction accuracy (see Methods) of 0.208. When we examined the transcript lengths of each build we saw that the two methods show very different transcript length distributions for both novel and previously annotated transcripts, with the short-read build was comprised mostly of smaller transcripts (Fig 2A). As the PacBio data was generated using two different libraries for 2000 bp and >3000 bp transcripts, we expected an enrichment for longer transcripts in the PacBio data set (Supplemental Figure 2). To assess accuracy relative to transcript length, we grouped transcripts by length in 1000 bp intervals, and compared accuracy between each group. We found that accuracy significantly improves for transcripts longer than 2000 bp. The construction accuracy is 0.426 and 0.137 for transcripts above and below 2000 bp, respectively.

We experimented with various methods to remove spurious transcripts and improve construction accuracy. We first removed transcripts that were expressed <1 TPM in at least one sample as outlined in StringTie's recommended protocol (Pertea et al., 2016). This improved construction accuracy to 0.475 for transcripts longer than 2000bp and 0.212 for transcripts shorter than 2000bp. As this accuracy was still fairly low, we tried different filtering schemes, including experimenting with machine learning-based strategies to identify transcripts that were computational artifacts (data not shown), but we found that the simplest approach with high performance was to retain transcripts that had an average TPM above a specific threshold(Fig 2C). In our downstream pipeline we keep transcripts that have at least an average of 1 TPM across all samples of the same subtissue type as this threshold achieved a build accuracy of 0.772 for transcripts longer than 2000Bp and retained 48470 transcripts within this short-read RPE dataset.

Thousands of novel gene isoforms are detected in human subtissue-specific transcriptomes

Tissue	Source	Samples	Studies	Transcriptome Count
RPE	Adult	48	4	32012
RPE	Fetal	49	7	49967
Retina	Adult	105	8	49714
Retina	Fetal	89	6	66255
Cornea	Adult	43	6	51469
Cornea	Fetal	6	2	59408

Table 1. Ocular sample dataset overview and transcriptome count. Transcriptome count is defined as the number of unique transcripts expressed in a given tissue type

We built transcriptomes from 340 publicly available ocular tissue RNA-seq samples curated in EiaD using an efficient Snakemake pipeline (Köster and Rahmann, 2012). We included both publicly collated non-disease, non-perturbed adult and fetal samples from cornea, retina, and RPE tissues, mined from 29 different studies (Table 1). Our fetal tissues consist of both human fetal tissues and human iPSC-derived tissue, as stem cell-derived tissue has been showed to closely resemble fetal tissue. We inlcude our iPSC-RPE samples originally used to develop our pipeline within this larger set of fetal RPE samples. (Klimanskaya et al., 2004). To more accurately determine the tissue specificity of novel ocular transcripts, we supplemented our ocular data set with 877 samples from 44 body locations across 22 major tissues from the GTEx project and constructed transcriptomes for each of these body locations (GTEx Consortium et al., 2017). We refer to each distinct body location as a subtissue here after.

After initial construction of transcriptomes, we found 183442 previously annotated transcripts and 6241675 novel transcripts detected in at least one of our 1217 samples. We define a novel transcripts as all transcripts whose set of exons and introns do not exactly match that of an annotated transcript within the Gencode, Ensembl, UCSC, and Refseq annotation databases (Frankish et al., 2019), (Zerbino et al., 2018), (O'Leary et al., 2016). After using the filtering methods described above, we merged all subtissue specific transcriptomes into a single final transcriptome which contains 252983 distinct transcripts with 87592 previously annotated and 165391 novel transcripts, and includes 114.9

156 megabases of previously unannotated genomic sequence (Table 1). We refer to the final
157 pan-body transcriptome as the DNTX annotation hereafter.

158 We split novel transcripts into two categories: novel isoforms, which are novel
159 variations of known genes, and novel loci, which are previously unreported, entirely novel
160 regions of transcribed sequence (Fig 3B). Novel isoforms are further classified by the
161 novelty of their encoded protein: isoforms with novel open reading frame, novel isoforms
162 with a known ORF, and isoforms with no ORF as noncoding isoforms (Fig 3A). The number
163 of distinct ORFs was significantly less than the number of transcripts, with 43279
164 previously annotated ORFs and 46226 novel ORFs across all subtissues. Furthermore,
165 across all subtissues there was an average of 10393 novel isoforms and 3716 novel ORFs.

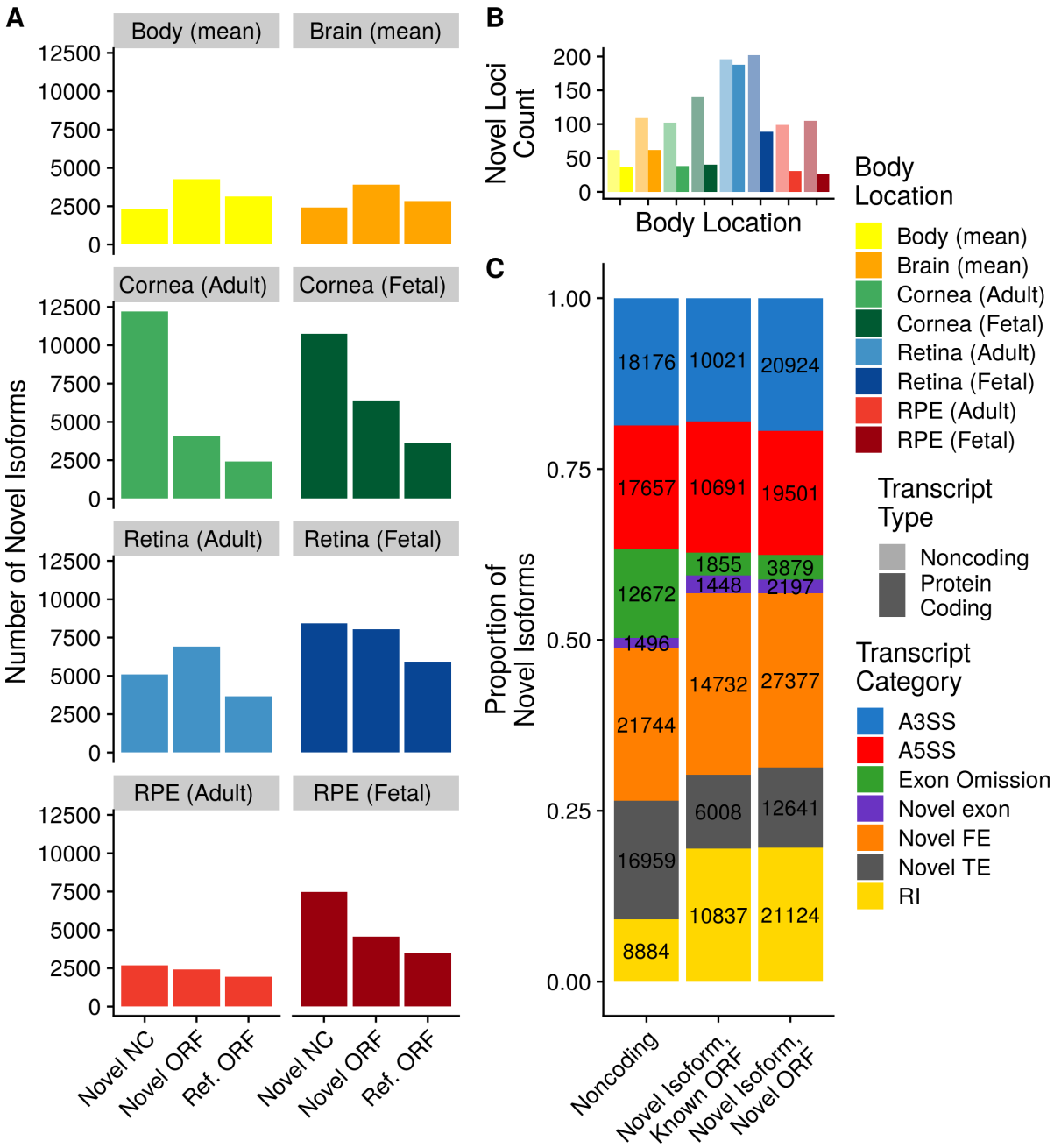


Figure 3. Overview of novel isoforms. A) Number of novel gene isoforms, grouped by transcript type. Brain and body represent an average of 13 and 34 distinct subtissues, respectively. B) Novel protein coding and noncoding loci. Novel exon composition of novel isoforms, by isoform type. Labels indicate number of transcripts. C) Classification of novel exon types, stratified by novel isoform type.

173 Novel isoforms can occur due to an omission of a previously annotated exon,
174 commonly referred as exon skipping or the addition of an unannotated exon which we
175 refer to as a novel exon. We further classified novel exons by the biological process that
176 may be driving their formation: alternative promoter usage driving the addition of novel
177 first exons (FE), alternative polyadenylation driving the addition of novel terminal exons
178 (TE), and alternative splicing driving the formation of all novel exons that are not the first
179 or last exon (Landry et al., 2003), (Tian and Manley, 2017), (Wang et al., 2015). We then
180 split alternatively spliced exons into their commonly seen patterns, alternative 5' splice site
181 (A5SS), alternative 3' splice site (A3SS), and retained introns (RI). Exons whose entire
182 sequence was unannotated and is not a retained intron are fully novel exons. We note that
183 all three of these mechanisms can lead to exon skipping, so for simplicity we grouped all
184 novel isoforms resulting from exon skipping together. We found that the majority of novel
185 exons within our dataset are novel FEs. We noticed that the majority of RI exons lead to
186 novel ORFs, whereas novel isoforms with omitted exons more often lead to noncoding
187 isoforms. (Fig 3C)

188 *De novo* transcriptomes match previously published experimental data better 189 than existing annotation

190 We validated *de novo* transcriptomes using three independent approaches. We first
191 looked for evolutionary conservation since it is commonly accepted as a proxy for
192 functional significance. We used the PhyloP 20 way species alignment, a measure of
193 conservation between species, to calculate the average conservation score for each exon in
194 the DNTX annotation and compared that to the average conservations score for each exon
195 in the Gencode annotation (Pollard et al., 2010). We found that, on average, exons in the
196 DNTX annotation are more conserved than exons in the Gencode annotation (pvalue <2.2e-
197 16) (Supplemental Figure 2A).

198 Next, since we observed an enrichment in novel first and last exons within our data
199 set, we decided to compare the TSS and TES within the DNTX annotation to two well-
200 established annotation databases from FANTOM and the polyA Atlas (Noguchi et al., 2017),
201 (Herrmann et al., 2020). We compared DNTX and Gencode TSS's to CAGE-seq data from the
202 FANTOM consortium; as CAGE-seq is optimized to detect the 5' end of transcripts, we

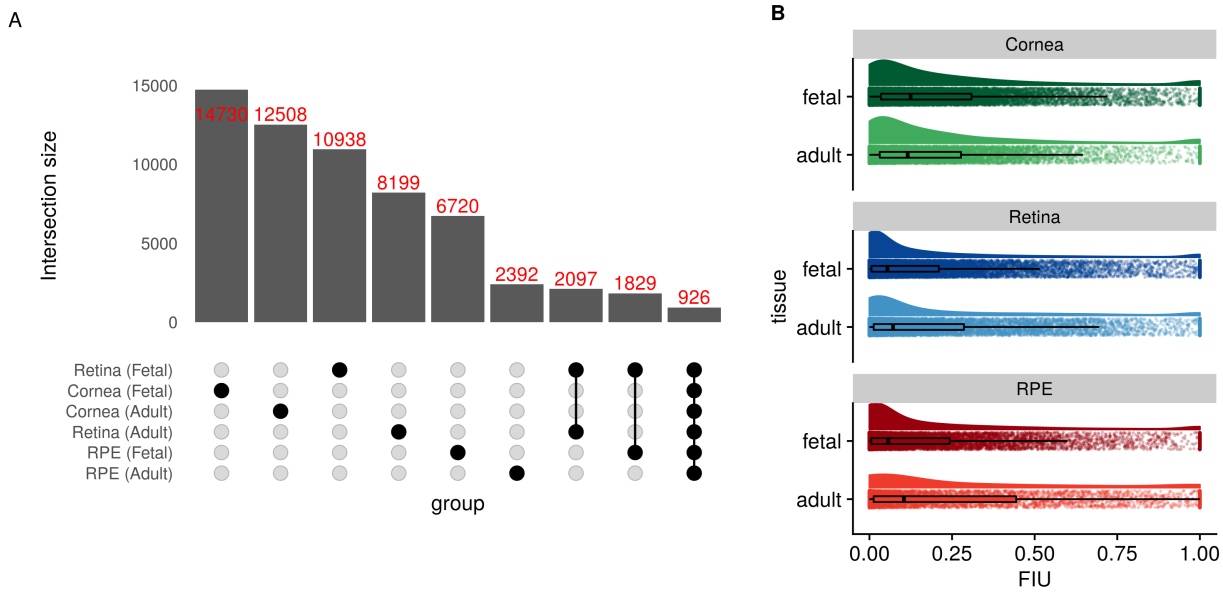
203 reasoned that it can serve as a valid ground truth set to evaluate TSS detection (Takahashi
204 et al., 2012). We calculated the absolute distance of DNTX TSS's to CAGE peaks, and
205 compared them to the absolute distance of Gencode TSS's to CAGE peaks. We found that, on
206 average, DNTX TSS's were closer to CAGE peaks than Gencode TSS's (pvalue <2.2e-
207 16)(Supplemental Figure 2B).

208 Finally, we evaluated TES's using the polyA Atlas, which is comprised of
209 polyadenylation signal annotation generated from aggregating 3' seq data from multiple
210 studies. As 3'-seq data is designed to accurately capture the 3' ends of transcripts, it can
211 similarly serve as a ground truth set to evaluate the accuracy of TES's (Beck et al., 2010).
212 We calculated the absolute distance of DNTX TES's to annotated polyA signals and
213 compared them to the absolute distance of Gencode TES's to polyA signals. We found that
214 on average DNTX TES's are closer to annotated polyadenylation signals than gencode TSS's
215 (pvalue <2.2e-16) (Supplemental Figure 2C)

216 *De novo* transcriptomes reduce overall transcriptome sizes

217 *De novo* transcriptomes removed on average 76.141 % of a subtissue's base
218 transcriptome. We defined base transcriptome for a subtissue as any transcript in the
219 Gencode annotation with non-zero TPM in at least one sample of a given subtissue. This
220 was a large reduction in transcriptome size and we wanted to ensure that we were not
221 unduly discarding data. We quantified transcript expression of each sample using Salmon
222 with two methods: once using the full gencode v28 human transcript annotation, and once
223 using its associated subtissue specific transcriptome. We found that despite the 76.141 %
224 reduction in number of transcripts between the base gencode and *de novo* transcriptomes
225 (Supplemental Figure 3A), the per-sample Salmon mapping rate increased on average by
226 2.041 % indicating that the vast majority of gene expression data is retained within our
227 transcriptome (Supplemental Figure 3B).

228 **Novel Isoforms are identified in ocular tissues**



229

230 Figure 4. Overview of novel gene isoforms in the eye. A) Set intersection of
231 novel isoforms in ocular transcriptomes. B) Boxplots of fraction isoform
232 usage (FIU) overlaid over FIU data points with estimated distribution of
233 data set above each boxplot.

234 Using the pan-eye transcriptome, we compared the overlap in constructed novel
235 isoforms across ocular subtissues and found that 77.968 % of novel isoforms are specific to
236 a singular ocular subtissue (Fig 4A). Additionally, fetal-like tissues had more novel isoforms
237 that their adult counterpart. For each novel isoform we then calculated fraction isoform
238 usage (FIU), or the fraction of total gene expression a transcript contributes to its parent
239 gene. We found that, on average, novel isoforms contributed to 20.584 % of their parent
240 gene's expression but in each subtissue we found multiple novel isoforms that contribute
241 to the majority of their parent genes expression (Fig 4B)

242 **Differential usage of gene isoforms occurs during retinal development**

243 Multiple studies have shown that gene isoforms play a significant role in eye
244 development (Bharti et al., 2008), (Mellough et al., 2019). We hypothesized that the DNTX
245 annotation provides additional insight into alternative isoform usage and identifies novel
246 gene isoforms potentially involved in eye development. We used RNA-seq data of the
247 developing retina from Mellough et al, an independent data set that we did not include for

248 transcriptome construction, and used a subset of the DNTX annotation corresponding to
249 fetal retina to quantify transcript expression and identify transcripts with significant
250 changes in expression across retinal development. Transcripts that are differentially
251 expressed (qvalue <.01) and have a mean FIU difference of .25 in at least one comparison of
252 time points are indicative of differential transcript usage (DTU).

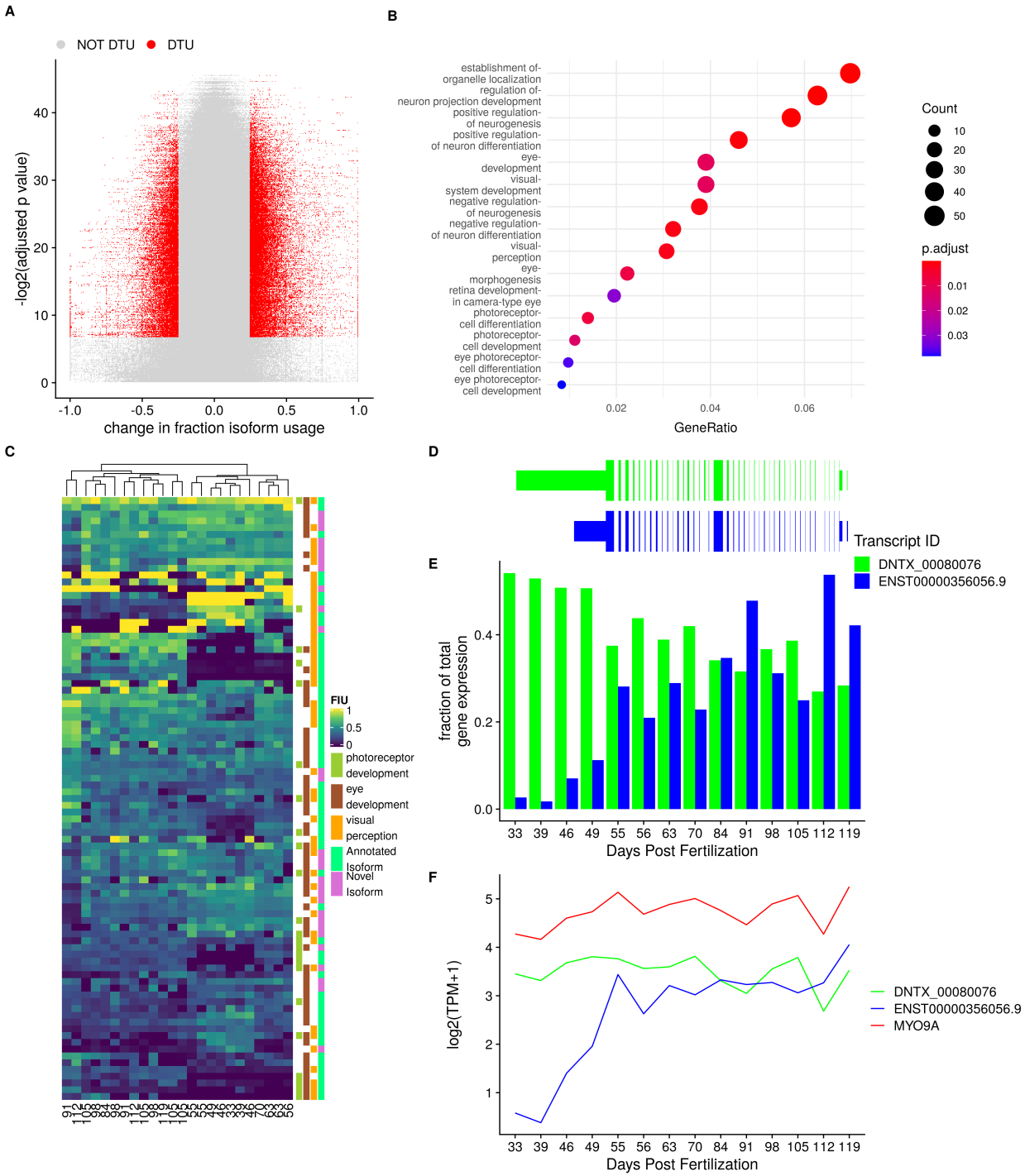


Figure 5 Differential Transcript usage during retinal development. A) Volcano plot of tested transcripts B) Dot plot for gene set enrichment analysis C) Heatmap of hiearchical clustering of transcripts with DTU associated with eye development D) Transcript models for *MYO9A*, a gene undergoing DTU E) FIU change in *MYO9A* FIU across development F) average log-transformed TPM expression of *MYO9A* across retinal development

261 We analyzed 24 samples across 14 developmental days post fertilization and found
262 1717 transcripts across 812 genes displaying DTU (Fig 5A). We found that genes involved
263 in DTU are enriched(qvalue <.05) for genes related to eye and neurological development
264 (Fig 5B), and that hierarchical clustering of DTU transcripts generates an early stage and
265 late stage cluster (Fig 5C). One of these genes, *MYO9A*, is a classical example of DTU. *MYO9A*
266 is associated with the visual perception GO term, plays a role in ocular development, and
267 has been associated with ocular disease (Gorman et al., 1999). While expression of *MYO9A*
268 remains relatively unchanged across development, expression of two of its associated
269 isoforms in fetal retina (Fig 5D) changes dramatically during development: a novel isoform
270 is highly expressed early during development, but switched to the canonical isoform later
271 in development (Fig 5E,F). This novel isoform contains a novel exon within the protein
272 coding region of the isoform as well as novel last exon extending the 3' UTR (Fig 5d). A full
273 list of genes and transcripts displaying DTU is available in Supplemental Data
274 (Supplemental Data 4).

275 ***De novo* transcriptomes allow for a more precise variant prioritization.**

276 The identification of a disease-causing variant through genome sequencing is a
277 common step in diagnosing genetic disease, when disease causing variants cannot be
278 determined from exonic sequencing. Prediction of a variant's biological impact and
279 subsequent variant prioritization is a fundamental step in this process. Many methods for
280 predicting variant effects on protein function or gene expression are based on location
281 within the body of a transcript; for example variants that disrupt splice sites and start/stop
282 codons are considered to be the most damaging, while variants within intronic and
283 intergenic regions have unknown impact or are not classified, and, thus, are not included
284 for further consideration. However, multiple studies have identified pathogenic deep
285 intronic variants for retinal dystrophies (Braun et al., 2013), (Bauwens et al., 2019),
286 (Zernant et al., 2014), (Sangermano et al., 2019), (Jamshidi et al., 2019), (Mayer et al.,
287 2016), (Geoffroy et al., 2018). Pathogenic intronic variants are thought to function by
288 introducing a novel splice site, disrupting regulatory motifs, or altering a tissue-specific
289 transcript. To explore this third possibility, we mapped known pathogenic intronic variants
290 onto novel isoforms within the *de novo* transcriptomes.

Gene Name	Associated Disease	Location (hg19)	Canonical Variant HGVS	Gencode Predicted Consequence	DNTX Predicted Consequence	Published Study
ABCA4	ABCA4-associated maculopathy	Chr1:94481967 C>T	c.5197–557G>T, NM_000350.2	intron variant, downstream gene variant	5 prime UTR variant	Bauwens et al.
		Chr1:94546814 G>C	c.859–540C>G, NM_000350.2	intron variant	non coding transcript exon variant	
	Stargardt disease	Chr1:94484001 C>T	c.5196+1137G>A, NM_000350.2	intron variant, downstream gene variant	5 prime UTR variant	Braun et al. Zernant et al.
		Chr1:94484082 T>G	c.5196+1056A>G, NM_000350.2	intron variant, downstream gene variant	5 prime UTR variant	
		Chr1:94526934 T>G	c.1938-619A>G, NM_000350.2	intron variant, splice region variant, non coding transcript variant	non coding transcript exon variant	Zernant et al.
		Chr1:94527698 G>C	c.1937+435C>G, NM_000350.2	intron variant, upstream gene variant	non coding transcript exon variant	Sangermano et al.
		Chr1:94546780 C>G	c.859-506G>C, NM_000350.2	intron variant	non coding transcript exon variant	
IFT140	Ciliopathy	Chr16:1576595 C>A	c.2577+25G>A, NM_014714.3	upstream gene variant, intron variant, NMD transcript variant, non coding transcript exon variant, non coding transcript variant	missense variant	Geoffroy et al.
PROM1	Cone–rod dystrophy	Chr4:15989860 T>G	c.2077-521A>G, NM_006017.2	intron variant, upstream gene variant	5 prime UTR variant	Mayer et al.
RPGRIP1	RPGRIP1-mediated inherited retinal degeneration	Chr14:21789588 G>A	c.1611+27G>A, NM_020366.3	intron variant, non coding transcript variant, upstream gene variant, synonymous variant, NMD transcript variant, downstream gene variant	5 prime UTR variant	Jamshidi et al.

Table 2. Pathogenic variants previously considered intronic that are on expressed transcripts in the retina *de novo* transcriptome. Canonical human genome variation society (HGVS) annotation is based on transcripts from the RefSeq annnotation. Predicted consequences were generaed with the Variant Effect Predictor(VEP)

We used a list of 129 intronic and noncoding variants previously identified as pathogenic for a retinal dystrophy and predicted the effect of these variants with Ensembl's Variant Effect Predictor using a subset of the DNTX annotation corresponding to fetal and adult retina as the input transcript annotation. We identified ten variants whose predicted effect increased in severity due the presence of a novel gene isoform in a previously intronic region (Table 2). Seven of these variants were in deep intronic hotpsots known for pathogenic variation within the gene ABCA4.

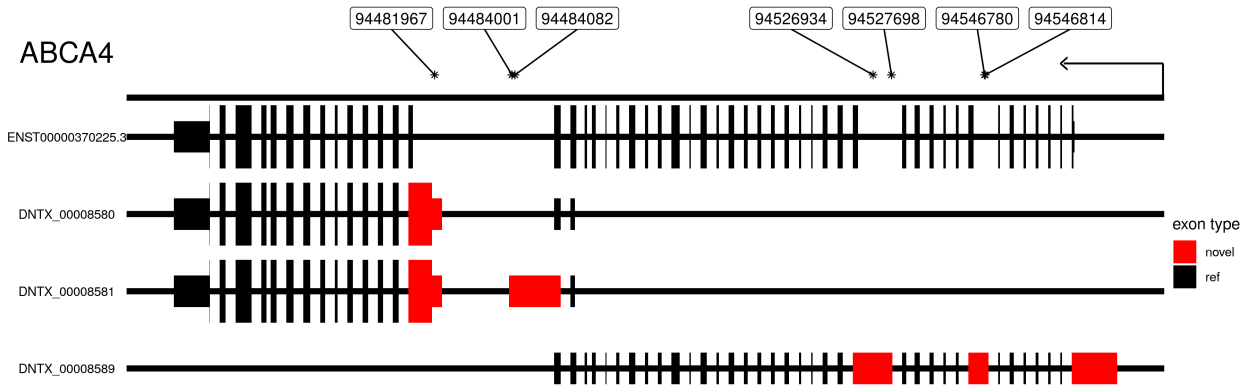


Figure 6. Transcript models for selected Isoforms of *ABCA4* along with location of pathogenic intronic variants. Location is on the hg19 human genome build. Thick lines indicate protein coding regions. Arrow indicates direction of transcription. Introns not drawn to scale

These variants were spanned by three distinct novel isoforms with two containing open reading frames (ORFs) encoding only the carboxy-terminus of the canonical protein isoform, and one noncoding spanning the proximal half of the canonical isoform (Fig 6). *ABCA4* expression and function has also been observed in RPE (Lenis et al., 2018). However, we did not observe these transcripts in RPE, suggesting that these pathogenic variants are primarily affecting retinal-specific *ABCA4* transcripts. We note that these transcripts have not been experimentally validated.

To further highlight the potential importance of *de novo* transcriptomes for future genetic tests we determined how many genes associated with retinal disease from RetNet have novel isoforms (sph.uth.edu/retnet/). We found that within the set of genes with novel isoforms, there is significant enrichment of retinal disease genes (hypergeometric pvalue = 3.4e-04), with 220 out of 379 RetNet genes having a novel isoform. A full list of these genes is available in the Supplementary data(supplemental data 5).

A companion visualization tool enables easy use of *de novo* transcriptomes

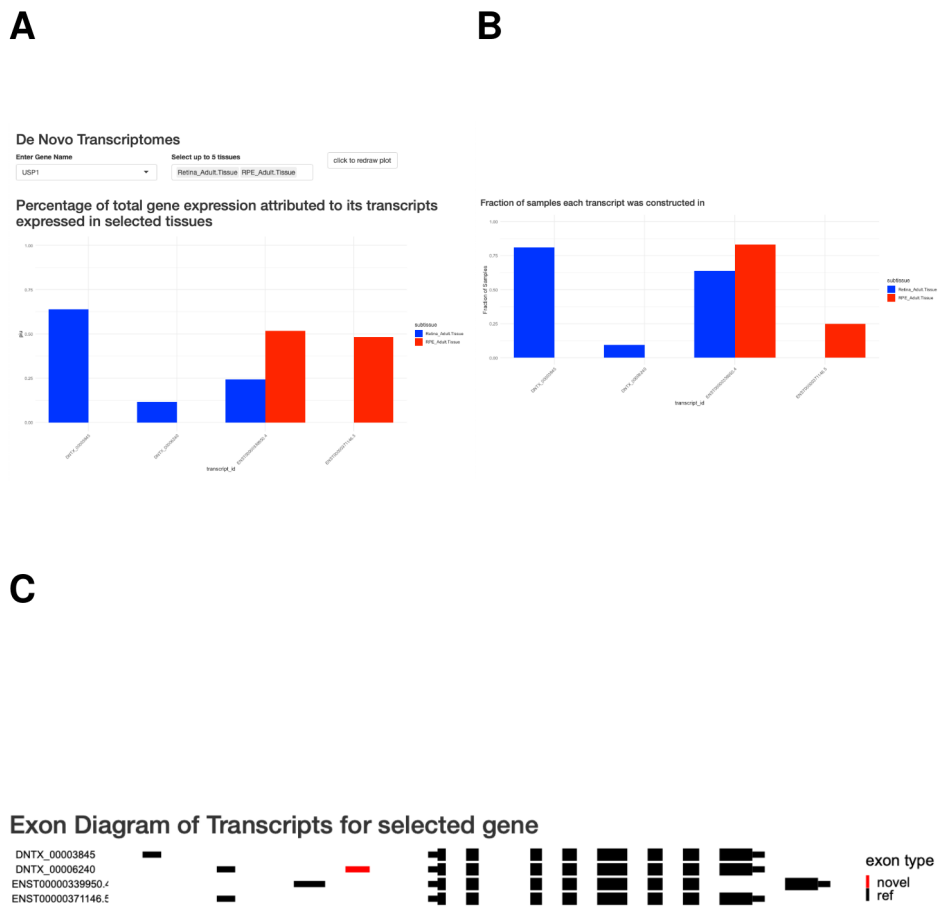


Figure 7. Screenshots from dynamic *de novo* transcriptome visualization tool. A). FIU bar plot for selected gene and subtissue. B). Exon level diagram of transcript body Thicklines represent coding region of transcript. novel exons colored in red. Tooltip contains genomic location and phylop score C) Bargraph of fraction of samples within dataset each transcript was consructed in by tissue.

To make our results easily accessible we designed a R-Shiny app for visualizing and accessing our *de novo* transcriptomes. For each subtissue we show the FIU for each transcript associated with a gene (Fig 7A). We show the exon-intron structure of each transcript and mousing over exons show genomic location overlapping SNPs, and

333 phylogenetic conservation score (Fig 7B). We additionally show a barplot of the fraction of
334 samples each transcript was constructed in (Fig 7C). Users can also download the *de novo*
335 transcriptomes for selected subtissues in GTF and fasta format. Instructions to download
336 and run the app are available at https://github.com/vinay-
337 swamy/ocular_transcriptomes_shiny. While visualization of direct transcript expresion is
338 not a part of this app, it can be viewed in the eyeIntegration app (Swamy and McGaughey,
339 2019) by selected 'DNTX' as the transcript annotation. Finally, we provide all code as a
340 Snakemake workflow and provide a Docker container with all software required for the
341 pipeline available at https://github.com/vinay-swamy/ocular_transcriptomes_pipeline

342 Discussion

343 Motivated by the lack of a comprehensive transcriptome for the eye, we constructed
344 transcriptomes for adult and fetal retina, RPE and cornea. By using long-read RNA-seq data
345 to calibrate our short-read construction pipeline, we were able to identify biologically
346 relevant transcriptomes. We found that concordance between long and short-read-based
347 transcriptomes is directly related to transcript length and transcript expression. We saw a
348 clear inability within the PacBio data set to accurately detect transcripts shorter than
349 2000bp for both previously annotated and novel transcripts. As many of the transcripts
350 constructed using short-reads are below this threshold, long-read sequencing data
351 enriched for smaller transcript sizes would provide greater insight in future studies.

352 We used a large dataset compiled from published RNA-seq data to build the pan-eye
353 transcriptomes, an approach that has several key advantages. First, the large sample size
354 overcomes the noisy nature of RNA-seq data. Second, as the cohort is constructed from
355 many independent studies, we are more confident that the transcriptomes accurately
356 reflect the biology of their originating subtissue and are not a technical artifact due to
357 preparation of the samples. As another line of evidence, the *de novo* transcriptomes match
358 existing large scale data sets and are more conserved than existing annotations
359 (Supplemental Figure 2).

360 In each ocular subtissue we examined, we found hundreds of novel gene isoforms,
361 many of which were novel due to novel exons. Within ocular subtissues, these novel
362 isoforms are most commonly specific to single subtissue. This makes sense as a majority of
363 the exons in our *de novo* transcriptomes are first and last exons, which have been
364 previously shown to significantly contribute to the tissue specificity of gene isoforms
365 (Reyes and Huber, 2018). We also found that on average novel isoforms represent about
366 20.584 % of their parent gene's expression. Future studies are needed to identify the
367 function of these isoforms. One possibility is that some of these isoforms are only
368 expressed in rare cell types, as transcript annotation was previously shown to be
369 incomplete in rare cell types (Zhang et al., 2020). This especially makes sense in the retina
370 which contain over a dozen distinct cell types, several of which contribute to 5% or less of
371 the total cell population (Yan et al., 2020). As we imposed a strict expression filter as part
372 of our transcriptome pipeline, we may have removed transcripts specific to rare cell types.

373 In conclusion, we created the first pan-eye transcriptome annotation and showed
374 that it is useful in understanding the role of gene isoforms in ocular biology and improving
375 the ability to diagnose inherited eye diseases. We hope this work is useful as a starting
376 point for other researchers; [delete] to make the transcriptomes easily accessible to other
377 researchers we designed a webapp both for visualization and to quickly access tissue-
378 specific annotation files. We believe this project will enable other researchers to explore
379 new research directions and answer long pending questions.

380 # Methods

381 ## Generation of PacBio long-read RNA sequencing data and Illumina short-read
382 RNA sequencing data

383 Human iPSCs were differentiated into RPE using previously described protocols in
384 (Bryan et al., 2018) and (May-Simera et al., 2018). iPSC-derived RPE (iPSC-RPE) cells at 42
385 days post differentiation were lysed with TRIzol reagent (Thermo Fisher Scientific; cat #
386 15596026) and total RNA was isolated using the Direct-zol RNA MiniPrep Kit (Zymo
387 Research, Irvine, CA). 5-6 μg total RNA that passed quality control metric (RIN >.9) were
388 used for PacBio library preparation. For PacBio HiFi circular consensus sequencing(CCS),

389 libraries were prepared following the "Procedure-Checklist-Iso-Seq-Express-Template-
390 Preparation-for-Sequel-and-Sequel-II-Systems" protocol. Two libraries were generated:
391 one to capture transcripts 2 kilobases(kb) or smaller, and one to capture transcripts
392 between 2-5kb. Sequencing was done on the PacBio Sequel II system for a movie time of 24
393 hours.

394 For Illumina sequencing, Poly-A selected stranded mRNA libraries were constructed
395 from 0.5-1 μg total RNA using the Illumina TruSeq Stranded mRNA Sample Prep Kits
396 according to manufacturer's instructions. Amplification was performed using 10-12 cycles
397 to minimize the risk of over-amplification. Unique dual-indexed barcode adapters were
398 applied to each library. Libraries were pooled in equimolar ratio and sequenced together
399 on a HiSeq 4000. At least 57 million 75-base read pairs were generated for each individual
400 library. Data was processed using illumina Real Time Analysis (RTA) version 2.7.7. All
401 library preparation and sequencing was performed at the National Institutes of Health
402 Intramural Sequencing Center (NISC).

403 Code availability and software versions.

404 To improve reproducibility, all code used for both the analyzing the data and
405 generating the figures for this paper was written as multiple Snakemake pipelines. Each
406 Snakefile contains the exact parameters for all tools and scripts used in each analysis.
407 (Köster and Rahmann, 2012) All code (and versions) used for this project is publicly
408 available in the following github repositories: https://github.com/vinay-
409 swamy/ocular_transcriptomes_pipeline (main pipeline), https://github.com/vinay-
410 swamy/ocular_transcriptomes_longread_analysis (long-read analysis pipeline),
411 https://github.com/vinay-swamy/ocular_transcriptomes_paper (figures and tables for this
412 paper), https://github.com/vinay-swamy/ocular_transcriptomes_shiny (webapp).
413 Additionally, all Snakefiles are included as supplementary data.(supplementary data files 1-
414 3)

415 Analysis of long-read data

416 PacBio sequencing movies were processed into full length, non-chimeric (FLNC)
417 reads using the IsoSeq3 3.1.2 pipeline in the PacBio SMRT link v7.0 software. The existing

418 ENCODE long-read RNA-seq pipeline (https://github.com/ENCODE-DCC/long-read-rna-
419 pipeline) was rewritten as a Snakemake workflow as follows. Transcripts were aligned to
420 the human genome using minimap2(18), using an alignment index built on the gencode
421 v28 primary human genome. Sequencing errors in aligned long-reads were corrected using
422 TranscriptClean (19) with default parameters. Splice junctions for TranscriptClean were
423 obtained using the TranscriptClean accessory script "get_SJs_from_gtf.py" using the
424 gencode v28 comprehensive transcript annotation as the input. A list of common variants
425 to avoid correcting were obtained from the ENCODE portal
426 (https://www.encodeproject.org/files/ENCFF911UGW/). The long-read transcriptome
427 annotation was generated with TALON (20). A TALON database was generated using the
428 talon_initialize_database command, with all default parameters, except for the "–5P" and "–
429 3p" parameters. These parameters represent the maximum distance between close 5' start
430 and 3' ends of similar transcript to merge and were both set to 100 to match parameters
431 used in later tools. Annotation in GTF format was generated using the talon_create_GTF
432 command, and transcript abundance values were generated using the talon_abundance
433 command.

434 **Analysis of short-read RPE data**

435 Each sample was aligned to the Gencode release 28 hg38 human genome assembly
436 using the genomic aligner STAR and the resulting BAM files were sorted using samtools
437 sort (Frankish et al., 2019),(Dobin et al., 2013),(Li et al., 2009). For each sorted BAM file, a
438 per-sample base transcriptome was constructed using StringTie with the Gencode v28
439 comprehensive annotation as a guiding annotation (Frankish et al., 2019),(Pertea et al.,
440 2015). All sample transcriptomes were merged with the long-read transcriptome using
441 gffcompare(Pertea and Pertea, 2020) with default parameters. We note that the default
442 values for the distance to merge similar 5' starts and 3 ends of transcripts in gffcompare is
443 the same to what we chose for TALON. We defined the metric construction accuracy, used
444 to evaluate short-read transcriptome construction as the following:

445 $$Construction\ Accuracy = \frac{short\ read\ transcriptome\ \cap\ long\ read\ transcriptome}{short\ read\ transcriptome}$$

446 **Construction of subtissue-specific transcriptomes.**

447 We constructed transcriptomes for 1217 samples in the Eye in a Disk(EiaD), a
448 dataset generated from aggregating publically available healthy, unperturbed RNA-seq
449 samples from 50 distinct locations of the body across 29 different studies. Specific
450 information on how this dataset was generated is detailed in the methods from our
451 previous work (Swamy and McGaughey, 2019). We constructed a transcriptome for each
452 sample, and merged samples together to create 50 subtissue-specific transcriptomes. We
453 define subtissue as a unique body location and are either temporally different versions of
454 the same tissue(adult vs fetal tissue), or different regions of a larger tissue (cortex vs
455 cerebellum in brain). Tissue refers to complete whole tissue (retina, brain, liver). For each
456 subtissue-specific transcriptome, we removed transcripts that had an average expression
457 less than 1 Transcripts Per Million (TPM) across all samples of the same subtissue type. All
458 subtissue-specific transcriptomes were merged to form a single unified annotation file in
459 general transfer format(GTF) to ensure transcript identifiers were the same across
460 subtissues. We merged all ocular subtissue transcriptomes to generate a separate pan-eye
461 transcriptome.

462 **Subtissue specific transcriptome quantification**

463 For each resulting subtissue specific transcriptome, we extracted transcript
464 sequences using the tool gffread and used these sequences to build a subtissue-specific
465 quantification index using the index mode of the alignment-free quantification tool Salmon
466 (Pertea and Pertea, 2020), (Patro et al., 2017). For each sample, we quantified transcript
467 expression using the quant mode of Salmon, using a sample's respective subtissue specific
468 quantification index. We similarly quantified all ocular samples using the pan-eye
469 transcriptome and the Gencode v28 reference transcriptome.

470 **Annotation of novel exons**

471 First, a comprehensive set of distinct, annotated exons was generated by merging
472 exon annotation from gencode, ensembl, UCSC, and refseq. We then defined a novel exon as
473 any exon within our transcriptomes that does not exactly match the chromosome, start,
474 end and strand of an annotated exon. Novels exons were classified by splitting exons into 3

475 categories: first, last, and middle exons. We then extracted all annotated exon start and stop
476 sites from our set of previously annotated exons. Novel middle exons that have an
477 annotated start but an unannotated end were categorized as a novel alternative 3' end
478 exons and similarly novel middle exons with an unannotated start but annotated end were
479 categorized as a novel alternative 5' start exons. Novel middle exons whose start and end
480 match annotated exon start and ends were considered retained introns. Novel middle
481 exons whose start and end do not match annotated starts and ends were considered fully
482 novel exons. We then classified novel first and last exons. Novel first exons were first exons
483 whose start is not in the set of annotated exon starts, and novel last exons were terminal
484 exons whose end is not in the set of annotated exon ends. This analysis of novel transcripts
485 is implemented in our Rscript "annotate_and_make_tissue_gtfs.R".

486 Validation of DNTX with phylop, CAGE data, and polyA signals

487 PhyloP scores for the phylop 20-way multi species alignment were downloaded
488 from UCSC's FTP server on October 16th, 2019 and converted from bigWig format to bed
489 format using the wig2bed tool in BEDOPs (Pollard et al., 2010), (Neph et al., 2012). The
490 average score per exon in both the gencode and DNTX annotation was calculated by
491 intersecting exon locations with phylop scores and then averaging the per base score for
492 each exon, using the intersect and groupby tools from the bedtools suite, respectively.
493 Significant difference in mean phylop score was tested with a Mann Whitney U test.

494 CAGE peaks were download from the FANTOM FTP server
495 (https://fantom.gsc.riken.jp/5/datafiles/reprocessed/hg38_latest/extra/CAGE_peaks/hg3
496 8_fair+new_CAGE_peaks_phase1and2.bed.gz) on June 15th 2020 (Noguchi et al., 2017).
497 Transcriptional start sites (TSS) were extracted from gencode and DNTX annotations; TSS
498 is defined as the start of the first exon of a transcript. Distance to CAGE peaks was
499 calculated using the closest tool in the bedtools suite. Significant difference in mean
500 distance to CAGE peak between DNTX and gencode annotation was tested with a Mann
501 Whitney U test.

502 Polyadenylation signal annotations were downloaded from the polyA site atlas
503 (https://polyasite.unibas.ch/download/atlas/2.0/GRCh38.96/atlas.clusters.2.0.GRCh38.96

504 .bed.gz) on June 15th 2020 (Herrmann et al., 2020). Transcriptional end sites(TES) were
505 extracted from gencode and DNTX annotations; TES is defined as the end of the terminal
506 exon of a transcript. Distance to polyA signal was calculated using the closest tool in the
507 bedtools suite (Quinlan and Hall, 2010). Significant difference in mean distance to polyA
508 signal was tested with a Mann Whitney U test.

509 Identification of novel protein coding transcripts

510 Protein-coding transcripts in the unified transcriptome were identified using the
511 TransDecoder suite (Haas et al., 2013). Transcript sequences in fasta format were extracted
512 from the final pan-body transcriptome using the TransDecoder util script
513 "gtf_genome_to_cdna_fasta.pl". Potential open reading frames(ORFs) were generated from
514 transcript sequences using the LongestORF module within TransDecoder, and the single
515 best ORF for each transcript was extracted with the Predict module within Transdecoder.
516 The resulting ORFs were mapped to genomic locations with the TransDecoder util script
517 "gtf_to_alignment_gff3.pl". For each ORF start and stop codons were extracted with the
518 script "agat_sp_add_start_stop.pl" scripts from the AGAT toolkit
519 (https://github.com/NBISweden/AGAT/). Transcripts with no detectable ORF or missing a
520 start or stop codon were labelled as noncoding.

521 Analysis of novel isoforms in eye tissues

522 An Upset plot was generated using the ComplexUpset package
523 (https://github.com/krassowski/complex-upset) (Lex et al., 2014). Fraction Isoform Usage
524 (FIU) was calculated for each transcript t associated with a parent gene g using the
525 following formula: $FIU_t = \frac{TPM_t}{TPM_g}$. Raincloud plots of FIU were generated using the
526 "R_Rainclouds" R package (Allen et al., 2019).

527 Analysis of fetal retina RNA-seq data.

528 RNA-seq samples from Mellough et al. were obtained from EiaD, and were not
529 included in the main dataset used for building transcriptomes. Outliers within the dataset
530 were identified by first performing principal component analysis of transcript level
531 expression data, calculating the center of all data using the first two principal components,

532 and subsequently removing five samples furthest away from the center of all data. The
533 remaining samples were normalized using calcNormFactors from the R package edgeR and
534 converted to weights using the voom function from the R package limma (Robinson et al.,
535 2010), (Ritchie et al., 2015). Differential expression was modeled using the lmFit function
536 using developmental time point as the model design and tested for significant change in
537 expression using the Ebayes function from limma. Gene Set enrichment was tested using
538 the R package clusterprofileR (Yu et al., 2012). Heatmaps were generated using the
539 ComplexHeatmap package (Gu et al., 2016).

540 Prediction of variant impact using *de novo* transcriptomes.

541 Noncoding variants previously associated with retinal disease from the Blueprint
542 Genetics Retinal dystrophy panel were obtained from the Blueprint Genetics website
543 (https://blueprintgenetics.com/tests/panels/ophthalmology/retinal-dystrophy-panel/).
544 The variants were converted from HGVS to VCF format using a custom python script
545 "HGVS_to_VCF.py". This VCF was then remapped to the hg38 human genome build using
546 the tool crossmap (Zhao et al., 2014). The VCF of variants was used as the input variants for
547 the Variant Effect Predictor(VEP) tool from Ensembl, with each subtissue specific
548 transcriptome as the input annotation (McLaren et al., 2016). VEP was additionally run
549 using the gencode v28 comprehensive annotation as the input annotation to identify
550 variants whose predicted impact increased in severity.

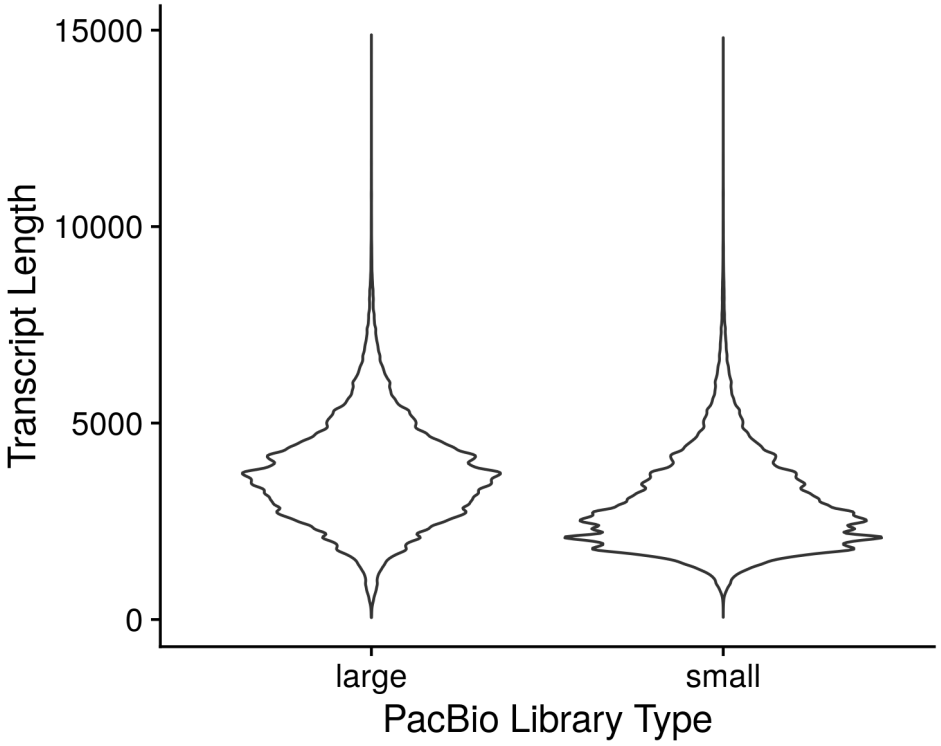
551 Figures, Tables, and Computing Resources

552 All statistical analyses, figures and tables in this paper were generated using the R
553 programming language. (R Core Team, 2019) A full list of packages and versions can be
554 found in the supplementary file session_info.txt. All computation was performed on the
555 National Institutes of Health high performance computer system Biowulf (hpc.nih.gov).

556 Competing Interests

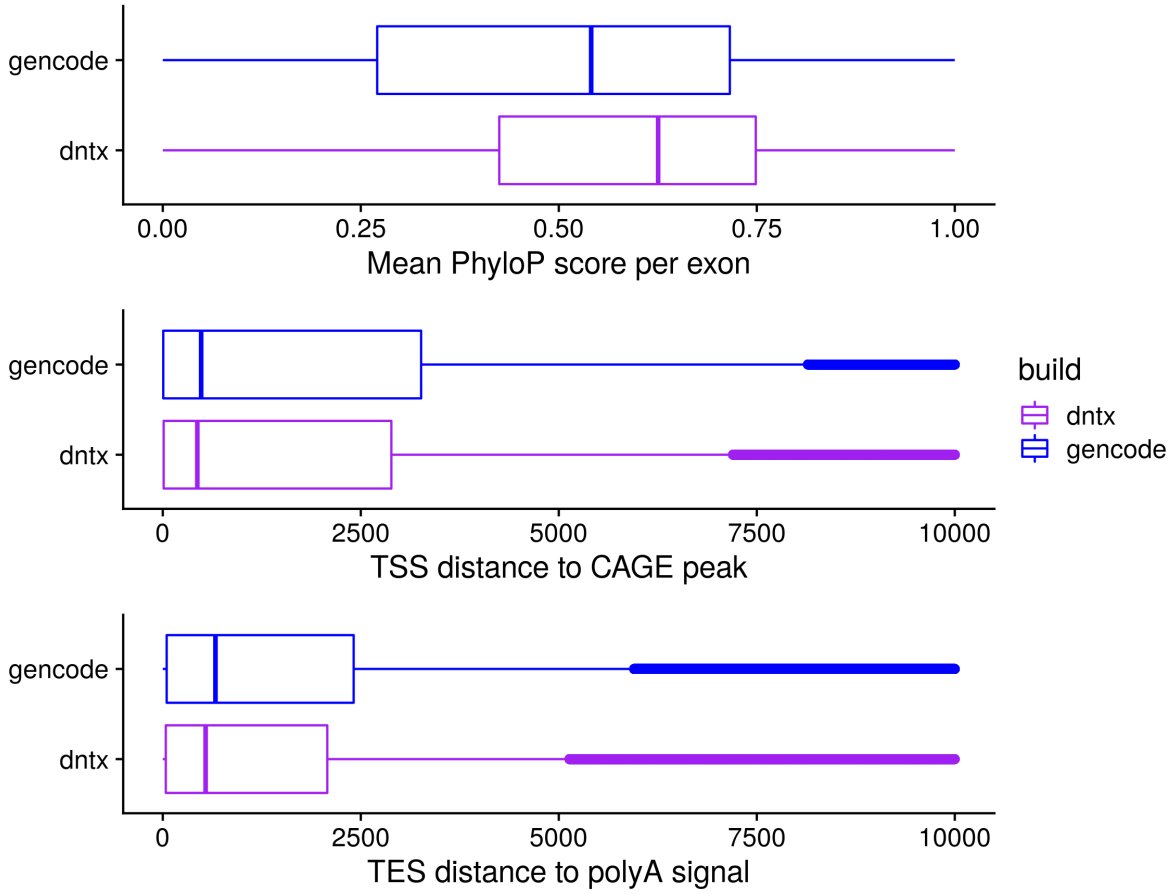
557 All authors declare no Competing interests.

558 **Supplemental Figures**

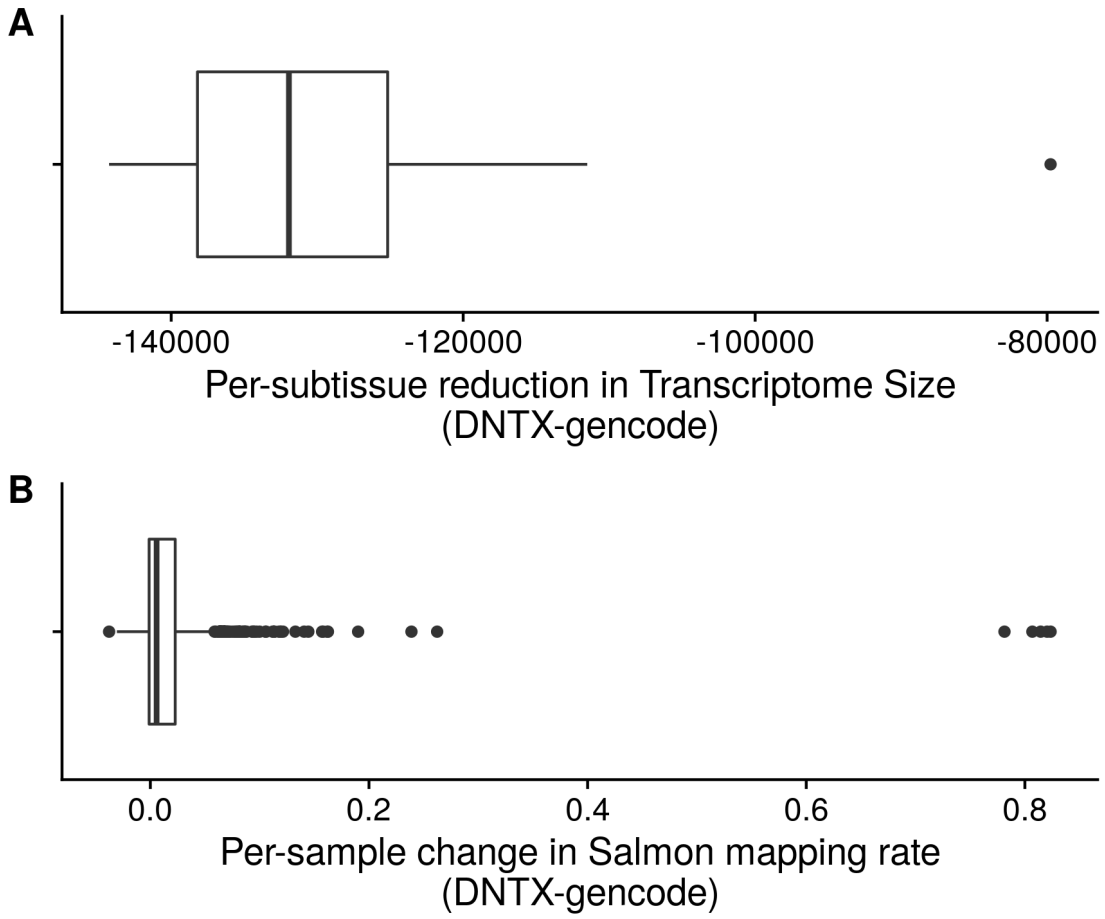


559

560 Supplemental Figure 1. Distribution of PacBio long-read lengths for two
561 library sizes.



Supplemental Figure 2. Comparison of DNTX annotation to Gencode annotation. A) Average per exon Phylop score for Gencode and DNTX transcripts. B) Average distance of DNTX transcriptional start sites (TSS) and Gencode TSS to CAGE-seq peaks from the FANTOM consortium. C) Average distance of DNTX transcriptional end sites (TES) and Gencode TES to polyadenylation signals in the PolyA site atlas.



Supplemental Figure 3. Comparison of Salmon mapping rate change vs transcriptome size decrease.

References

Allen M, Poggiali D, Whitaker K, Marshall TR, Kievit RA. 2019. Raincloud plots: A multi-platform tool for robust data visualization. *Wellcome Open Research* **4**:63. doi:10.12688/wellcomeopenres.15191.1

Bauwens M, Garanto A, Sangermano R, Naessens S, Weisschuh N, De Zaeytijd J, Khan M, Sadler F, Balikova I, Van Cauwenbergh C, Rosseel T, Bauwens J, De Leeneer K, De Jaegere S, Van Laethem T, De Vries M, Carss K, Arno G, Fakin A, Webster AR, Ravel de l'Argentière TJL de, Sznajer Y, Vuylsteke M, Kohl S, Wissinger B, Cherry T, Collin RWJ, Cremers FPM, Leroy BP, De Baere E. 2019. ABCA4-associated disease as a model for missing heritability in autosomal recessive disorders: Novel noncoding splice, cis-regulatory, structural, and recurrent hypomorphic variants. *Genetics in Medicine* **21**:1761–1771. doi:10.1038/s41436-018-0420-y

584 Beck AH, Weng Z, Witten DM, Zhu S, Foley JW, Lacroute P, Smith CL, Tibshirani R, Rijn M
585 van de, Sidow A, West RB. 2010. 3′-End Sequencing for Expression Quantification (3SEQ)
586 from Archival Tumor Samples. *PLOS ONE* **5**:e8768. doi:10.1371/journal.pone.0008768

587 Bharti K, Liu W, Csermely T, Bertuzzi S, Arnheiter H. 2008. Alternative promoter use in eye
588 development: Complex role and regulation of the transcription factor MITF. *Development*
589 *(Cambridge, England)* **135**:1169–1178. doi:10.1242/dev.014142

590 Blenkinsop TA, Saini JS, Maminishkis A, Bharti K, Wan Q, Banzon T, Lotfi M, Davis J, Singh D,
591 Rizzolo LJ, Miller S, Temple S, Stern JH. 2015. Human Adult Retinal Pigment Epithelial Stem
592 Cell-Derived RPE Monolayers Exhibit Key Physiological Characteristics of Native Tissue.
593 *Investigative Ophthalmology & Visual Science* **56**:7085–7099. doi:10.1167/iovs.14-16246

594 Braun TA, Mullins RF, Wagner AH, Andorf JL, Johnston RM, Bakall BB, Deluca AP, Fishman
595 GA, Lam BL, Weleber RG, Cideciyan AV, Jacobson SG, Sheffield VC, Tucker BA, Stone EM.
596 2013. Non-exomic and synonymous variants in ABCA4 are an important cause of Stargardt
597 disease. *Human Molecular Genetics* **22**:5136–5145. doi:10.1093/hmg/ddt367

598 Bryan JM, Fufa TD, Bharti K, Brooks BP, Hufnagel RB, McGaughey DM. 2018. Identifying
599 core biological processes distinguishing human eye tissues with precise systems-level gene
600 expression analyses and weighted correlation networks. *Human Molecular Genetics*
601 **27**:3325–3339. doi:10.1093/hmg/ddy239

602 Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras
603 TR. 2013. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)*
604 **29**:15–21. doi:10.1093/bioinformatics/bts635

605 Dykes IM, Bueren KL van, Scambler PJ. 2018. HIC2 regulates isoform switching during
606 maturation of the cardiovascular system. *Journal of Molecular and Cellular Cardiology*
607 **114**:29–37. doi:10.1016/j.yjmcc.2017.10.007

608 Frankish A, Diekhans M, Ferreira A-M, Johnson R, Jungreis I, Loveland J, Mudge JM, Sisu C,
609 Wright J, Armstrong J, Barnes I, Berry A, Bignell A, Carbonell Sala S, Chrast J, Cunningham F,
610 Di Domenico T, Donaldson S, Fiddes IT, Garcia Giron C, Gonzalez JM, Grego T, Hardy M,
611 Hourlier T, Hunt T, Izuogu OG, Lagarde J, Martin FJ, Martinez L, Mohanan S, Muir P, Navarro
612 FCP, Parker A, Pei B, Pozo F, Ruffier M, Schmitt BM, Stapleton E, Suner M-M, Sycheva I,
613 Uszczynska-Ratajczak B, Xu J, Yates A, Zerbino D, Zhang Y, Aken B, Choudhary JS, Gerstein
614 M, Guigo R, Hubbard TJP, Kellis M, Paten B, Reymond A, Tress ML, Flicek P. 2019. GENCODE
615 reference annotation for the human and mouse genomes. *Nucleic acids research* **47**:D766–
616 D773. doi:10.1093/nar/gky955

617 Geoffroy V, Stoetzel C, Scheidecker S, Schaefer E, Perrault I, Bär S, Kröll A, Delbarre M, Antin
618 M, Leuvrey A-S, Henry C, Blanché H, Decker E, Kloth K, Klaus G, Mache C, Martin-Coignard
619 D, McGinn S, Boland A, Deleuze J-F, Friant S, Saunier S, Rozet J-M, Bergmann C, Dollfus H,
620 Muller J. 2018. Whole-genome sequencing in patients with ciliopathies uncovers a novel
621 recurrent tandem duplication in IFT140. *Human Mutation* **39**:983–992.
622 doi:10.1002/humu.23539

623 Gorman SW, Haider NB, Grieshammer U, Swiderski RE, Kim E, Welch JW, Searby C, Leng S,
624 Carmi R, Sheffield VC, Duhl DM. 1999. The Cloning and Developmental Expression of
625 Unconventional Myosin IXA (MYO9A) a Gene in the Bardet–Biedl Syndrome (BBS4) Region
626 at Chromosome 15q22–q23. *Genomics* **59**:150–160. doi:10.1006/geno.1999.5867

627 GTEx Consortium, Laboratory, Data Analysis &Coordinating Center (LDACC)—Analysis
628 Working Group, Statistical Methods groups—Analysis Working Group, Enhancing GTEx
629 (eGTEx) groups, NIH Common Fund, NIH/NCI, NIH/NHGRI, NIH/NIMH, NIH/NIDA,
630 Biospecimen Collection Source Site—NDRI, Biospecimen Collection Source Site—RPCI,
631 Biospecimen Core Resource—VARI, Brain Bank Repository—University of Miami Brain
632 Endowment Bank, Leidos Biomedical—Project Management, ELSI Study, Genome Browser
633 Data Integration &Visualization—EBI, Genome Browser Data Integration &Visualization—
634 UCSC Genomics Institute, University of California Santa Cruz, Lead analysts: Laboratory,
635 Data Analysis &Coordinating Center (LDACC): NIH program management: Biospecimen
636 collection: Pathology: eQTL manuscript working group: Battle A, Brown CD, Engelhardt BE,
637 Montgomery SB. 2017. Genetic effects on gene expression across human tissues. *Nature*
638 **550**:204–213. doi:10.1038/nature24277

639 Gu Z, Eils R, Schlesner M. 2016. Complex heatmaps reveal patterns and correlations in
640 multidimensional genomic data. *Bioinformatics* **32**:2847–2849.
641 doi:10.1093/bioinformatics/btw313

642 Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D,
643 Li B, Lieber M, MacManes MD, Ott M, Orvis J, Pochet N, Strozzi F, Weeks N, Westerman R,
644 William T, Dewey CN, Henschel R, LeDuc RD, Friedman N, Regev A. 2013. De novo
645 transcript sequence reconstruction from RNA-Seq: Reference generation and analysis with
646 Trinity. *Nature protocols* **8**. doi:10.1038/nprot.2013.084

647 Herrmann CJ, Schmidt R, Kanitz A, Artimo P, Gruber AJ, Zavolan M. 2020. PolyASite 2.0: A
648 consolidated atlas of polyadenylation sites from 3′ end sequencing. *Nucleic Acids Research*
649 **48**:D174–D179. doi:10.1093/nar/gkz918

650 Jamshidi F, Place EM, Mehrotra S, Navarro-Gomez D, Maher M, Branham KE, Valkanas E,
651 Cherry TJ, Lek M, MacArthur D, Pierce EA, Bujakowska KM. 2019. Contribution of non-
652 coding mutations to RPGRIP1-mediated inherited retinal degeneration. *Genetics in medicine*
653 *: official journal of the American College of Medical Genetics* **21**:694–704.
654 doi:10.1038/s41436-018-0104-7

655 Klimanskaya I, Hipp J, Rezai KA, West M, Atala A, Lanza R. 2004. Derivation and
656 comparative assessment of retinal pigment epithelium from human embryonic stem cells
657 using transcriptomics. *Cloning and Stem Cells* **6**:217–245. doi:10.1089/clo.2004.6.217

658 Köster J, Rahmann S. 2012. Snakemake—a scalable bioinformatics workflow engine.
659 *Bioinformatics* **28**:2520–2522. doi:10.1093/bioinformatics/bts480

660 Landry J-R, Mager DL, Wilhelm BT. 2003. Complex controls: The role of alternative
661 promoters in mammalian genomes. *Trends in Genetics* **19**:640–648.
662 doi:10.1016/j.tig.2003.09.014

663 Lenis TL, Hu J, Ng SY, Jiang Z, Sarfare S, Lloyd MB, Esposito NJ, Samuel W, Jaworski C, Bok D,
664 Finnemann SC, Radeke MJ, Redmond TM, Travis GH, Radu RA. 2018. Expression of ABCA4
665 in the retinal pigment epithelium and its implications for Stargardt macular degeneration.
666 *Proceedings of the National Academy of Sciences* **115**:E11120–E11127.
667 doi:10.1073/pnas.1802519115

668 Lex A, Gehlenborg N, Strobelt H, Vuillemot R, Pfister H. 2014. UpSet: Visualization of
669 Intersecting Sets. *IEEE Transactions on Visualization and Computer Graphics* **20**:1983–1992.
670 doi:10.1109/TVCG.2014.2346248

671 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R,
672 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map
673 format and SAMtools. *Bioinformatics (Oxford, England)* **25**:2078–2079.
674 doi:10.1093/bioinformatics/btp352

675 Maruotti J, Sripathi SR, Bharti K, Fuller J, Wahlin KJ, Ranganathan V, Sluch VM, Berlinicke
676 CA, Davis J, Kim C, Zhao L, Wan J, Qian J, Corneo B, Temple S, Dubey R, Olenyuk BZ, Bhutto I,
677 Lutty GA, Zack DJ. 2015. Small-molecule–directed, efficient generation of retinal pigment
678 epithelium from human pluripotent stem cells. *Proceedings of the National Academy of*
679 *Sciences* **112**:10950–10955. doi:10.1073/pnas.1422818112

680 Mayer AK, Rohrschneider K, Strom TM, Glöckle N, Kohl S, Wissinger B, Weisschuh N. 2016.
681 Homozygosity mapping and whole-genome sequencing reveals a deep intronic PROM1
682 mutation causing cone–rod dystrophy by pseudoexon activation. *European Journal of*
683 *Human Genetics* **24**:459–462. doi:10.1038/ejhg.2015.144

684 May-Simera HL, Wan Q, Jha BS, Hartford J, Khristov V, Dejene R, Chang J, Patnaik S, Lu Q,
685 Banerjee P, Silver J, Insinna-Kettenhofen C, Patel D, Lotfi M, Malicdan M, Hotaling N,
686 Maminishkis A, Sridharan R, Brooks B, Miyagishima K, Gunay-Aygun M, Pal R, Westlake C,
687 Miller S, Sharma R, Bharti K. 2018. Primary Cilium-Mediated Retinal Pigment Epithelium
688 Maturation Is Disrupted in Ciliopathy Patient Cells. *Cell reports* **22**:189–205.
689 doi:10.1016/j.celrep.2017.12.038

690 McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, Flicek P, Cunningham F.
691 2016. The Ensembl Variant Effect Predictor. *Genome Biology* **17**:122. doi:10.1186/s13059-
692 016-0974-4

693 Mellough CB, Bauer R, Collin J, Dorgau B, Zerti D, Dolan DWP, Jones CM, Izuogu OG, Yu M,
694 Hallam D, Steyn JS, White K, Steel DH, Santibanez-Koref M, Elliott DJ, Jackson MS, Lindsay S,
695 Grellscheid S, Lako M. 2019. An integrated transcriptional analysis of the developing
696 human retina. *Development (Cambridge, England)* **146**. doi:10.1242/dev.169474

697 Mills JD, Nalpathamkalam T, Jacobs HIL, Janitz C, Merico D, Hu P, Janitz M. 2013. RNA-Seq
698 analysis of the parietal cortex in Alzheimer's disease reveals alternatively spliced isoforms
699 related to lipid metabolism. *Neuroscience Letters* **536**:90–95.
700 doi:10.1016/j.neulet.2012.12.042

701 Mitra M, Lee HN, Coller HA. 2020. Splicing Busts a Move: Isoform Switching Regulates
702 Migration. *Trends in Cell Biology* **30**:74–85. doi:10.1016/j.tcb.2019.10.007

703 Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M. 2008. The
704 Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing. *Science*
705 **320**:1344–1349. doi:10.1126/science.1158441

706 Neagoe Ciprian, Kulke Michael, del Monte Federica, Gwathmey Judith K., de Tombe Pieter
707 P., Hajjar Roger J., Linke Wolfgang A. 2002. Titin Isoform Switch in Ischemic Human Heart
708 Disease. *Circulation* **106**:1333–1341. doi:10.1161/01.CIR.0000029803.93022.93

709 Neph S, Kuehn MS, Reynolds AP, Haugen E, Thurman RE, Johnson AK, Rynes E, Maurano
710 MT, Vierstra J, Thomas S, Sandstrom R, Humbert R, Stamatoyannopoulos JA. 2012. BEDOPS:
711 High-performance genomic feature operations. *Bioinformatics* **28**:1919–1920.
712 doi:10.1093/bioinformatics/bts277

713 Noguchi S, Arakawa T, Fukuda S, Furuno M, Hasegawa A, Hori F, Ishikawa-Kato S, Kaida K,
714 Kaiho A, Kanamori-Katayama M, Kawashima T, Kojima M, Kubosaki A, Manabe R-i, Murata
715 M, Nagao-Sato S, Nakazato K, Ninomiya N, Nishiyori-Sueki H, Noma S, Saijyo E, Saka A, Sakai
716 M, Simon C, Suzuki N, Tagami M, Watanabe S, Yoshida S, Arner P, Axton RA, Babina M,
717 Baillie JK, Barnett TC, Beckhouse AG, Blumenthal A, Bodega B, Bonetti A, Briggs J,
718 Brombacher F, Carlisle AJ, Clevers HC, Davis CA, Detmar M, Dohi T, Edge ASB, Edinger M,
719 Ehrlund A, Ekwall K, Endoh M, Enomoto H, Eslami A, Fagiolini M, Fairbairn L, Farach-
720 Carson MC, Faulkner GJ, Ferrai C, Fisher ME, Forrester LM, Fujita R, Furusawa J-i,
721 Geijtenbeek TB, Gingeras T, Goldowitz D, Guhl S, Guler R, Gustincich S, Ha TJ, Hamaguchi M,
722 Hara M, Hasegawa Y, Herlyn M, Heutink P, Hitchens KJ, Hume DA, Ikawa T, Ishizu Y, Kai C,
723 Kawamoto H, Kawamura YI, Kempfle JS, Kenna TJ, Kere J, Khachigian LM, Kitamura T, Klein
724 S, Klinken SP, Knox AJ, Kojima S, Koseki H, Koyasu S, Lee W, Lennartsson A, Mackay-sim A,
725 Mejhert N, Mizuno Y, Morikawa H, Morimoto M, Moro K, Morris KJ, Motohashi H, Mummery
726 CL, Nakachi Y, Nakahara F, Nakamura T, Nakamura Y, Nozaki T, Ogishima S, Ohkura N,
727 Ohno H, Ohshima M, Okada-Hatakeyama M, Okazaki Y, Orlando V, Ovchinnikov DA, Passier
728 R, Patrikakis M, Pombo A, Pradhan-Bhatt S, Qin X-Y, Rehli M, Rizzu P, Roy S, Sajantila A,
729 Sakaguchi S, Sato H, Satoh H, Savvi S, Saxena A, Schmidl C, Schneider C, Schulze-Tanzil GG,
730 Schwegmann A, Sheng G, Shin JW, Sugiyama D, Sugiyama T, Summers KM, Takahashi N,
731 Takai J, Tanaka H, Tatsukawa H, Tomoiu A, Toyoda H, Wetering M van de, Berg LM van den,
732 Verardo R, Vijayan D, Wells CA, Winteringham LN, Wolvetang E, Yamaguchi Y, Yamamoto
733 M, Yanagi-Mizuochi C, Yoneda M, Yonekura Y, Zhang PG, Zucchelli S, Abugessaisa I, Arner E,
734 Harshbarger J, Kondo A, Lassmann T, Lizio M, Sahin S, Sengstag T, Severin J, Shimoji H,
735 Suzuki M, Suzuki H, Kawai J, Kondo N, Itoh M, Daub CO, Kasukawa T, Kawaji H, Carninci P,
736 Forrest ARR, Hayashizaki Y. 2017. FANTOM5 CAGE profiles of human and mouse samples.
737 *Scientific Data* **4**:170112. doi:10.1038/sdata.2017.112

738 O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B,
739 Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V,
740 Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Goldfarb T, Gupta T, Haft D, Hatcher E,
741 Hlavina W, Joardar VS, Kodali VK, Li W, Maglott D, Masterson P, McGarvey KM, Murphy MR,
742 O'Neill K, Pujar S, Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda A, Storz SS, Sun H,

743 Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W, Landrum MJ,
744 Kimchi A, Tatusova T, DiCuccio M, Kitts P, Murphy TD, Pruitt KD. 2016. Reference sequence
745 (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation.
746 *Nucleic Acids Research* **44**:D733–745. doi:10.1093/nar/gkv1189

747 Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. 2017. Salmon provides fast and bias-
748 aware quantification of transcript expression. *Nature methods* **14**:417–419.
749 doi:10.1038/nmeth.4197

750 Perrin RM, Konopatskaya O, Qiu Y, Harper S, Bates DO, Churchill AJ. 2005. Diabetic
751 retinopathy is associated with a switch in splicing from anti- to pro-angiogenic isoforms of
752 vascular endothelial growth factor. *Diabetologia* **48**:2422–2427. doi:10.1007/s00125-005-
753 1951-8

754 Pertea G, Pertea M. 2020. GFF Utilities: GffRead and GffCompare. *F1000Research* **9**:304.
755 doi:10.12688/f1000research.23297.1

756 Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. 2016. Transcript-level expression
757 analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nature Protocols*
758 **11**:1650–1667. doi:10.1038/nprot.2016.095

759 Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL. 2015. StringTie
760 enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature*
761 *Biotechnology* **33**:290–295. doi:10.1038/nbt.3122

762 Pertea M, Shumate A, Pertea G, Varabyou A, Breitwieser FP, Chang Y-C, Madugundu AK,
763 Pandey A, Salzberg SL. 2018. CHESS: A new human gene catalog curated from thousands of
764 large-scale RNA sequencing experiments reveals extensive transcriptional noise. *Genome*
765 *Biology* **19**:208. doi:10.1186/s13059-018-1590-2

766 Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. 2010. Detection of nonneutral substitution
767 rates on mammalian phylogenies. *Genome Research* **20**:110–121.
768 doi:10.1101/gr.097857.109

769 Quinlan AR, Hall IM. 2010. BEDTools: A flexible suite of utilities for comparing genomic
770 features. *Bioinformatics (Oxford, England)* **26**:841–842.
771 doi:10.1093/bioinformatics/btq033

772 R Core Team. 2019. R: A Language and Environment for Statistical Computing. Vienna,
773 Austria: R Foundation for Statistical Computing.

774 Reyes A, Huber W. 2018. Alternative start and termination sites of transcription drive most
775 transcript isoform differences across human tissues. *Nucleic Acids Research* **46**:582–592.
776 doi:10.1093/nar/gkx1165

777 Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. 2015. Limma powers
778 differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids*
779 *Research* **43**:e47–e47. doi:10.1093/nar/gkv007

780 Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: A Bioconductor package for
781 differential expression analysis of digital gene expression data. *Bioinformatics* **26**:139–140.
782 doi:10.1093/bioinformatics/btp616

783 Sangermano R, Garanto A, Khan M, Runhart EH, Bauwens M, Bax NM, Born LI van den, Khan
784 MI, Cornelis SS, Verheij JBGM, Pott J-WR, Thiadens AAHJ, Klaver CCW, Puech B, Meunier I,
785 Naessens S, Arno G, Fakin A, Carss KJ, Raymond FL, Webster AR, Dhaenens C-M, Stöhr H,
786 Grassmann F, Weber BHF, Hoyng CB, De Baere E, Albert S, Collin RWJ, Cremers FPM. 2019.
787 Deep-intronic ABCA4 variants explain missing heritability in Stargardt disease and allow
788 correction of splice defects by antisense oligonucleotides. *Genetics in Medicine* **21**:1751–
789 1760. doi:10.1038/s41436-018-0414-9

790 Swamy V, McGaughey D. 2019. Eye in a Disk: eyeIntegration Human Pan-Eye and Body
791 Transcriptome Database Version 1.0. *Investigative Ophthalmology & Visual Science*
792 **60**:3236–3246. doi:10.1167/iovs.19-27106

793 Takahashi H, Kato S, Murata M, Carninci P. 2012. CAGE- Cap Analysis Gene Expression: A
794 protocol for the detection of promoter and transcriptional networks. *Methods in molecular*
795 *biology (Clifton, NJ)* **786**:181–200. doi:10.1007/978-1-61779-292-2_11

796 Tian B, Manley JL. 2017. Alternative polyadenylation of mRNA precursors. *Nature Reviews*
797 *Molecular Cell Biology* **18**:18–30. doi:10.1038/nrm.2016.116

798 Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, Baren MJ van, Salzberg SL, Wold BJ,
799 Pachter L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated
800 transcripts and isoform switching during cell differentiation. *Nature Biotechnology* **28**:511–
801 515. doi:10.1038/nbt.1621

802 Vitting-Seerup K, Sandelin A. 2017. The Landscape of Isoform Switches in Human Cancers.
803 *Molecular Cancer Research* **15**:1206–1220. doi:10.1158/1541-7786.MCR-16-0459

804 Wang Y, Liu J, HUANG B, XU Y-M, LI J, HUANG L-F, LIN J, ZHANG J, MIN Q-H, YANG W-M,
805 WANG X-Z. 2015. Mechanism of alternative splicing and its regulation. *Biomedical Reports*
806 **3**:152–158. doi:10.3892/br.2014.407

807 Wenger AM, Peluso P, Rowell WJ, Chang P-C, Hall RJ, Concepcion GT, Ebler J,
808 Fungtammasan A, Kolesnikov A, Olson ND, Töpfer A, Alonge M, Mahmoud M, Qian Y, Chin C-
809 S, Phillippy AM, Schatz MC, Myers G, DePristo MA, Ruan J, Marschall T, Sedlazeck FJ, Zook
810 JM, Li H, Koren S, Carroll A, Rank DR, Hunkapiller MW. 2019. Accurate circular consensus
811 long-read sequencing improves variant detection and assembly of a human genome. *Nature*
812 *Biotechnology* **37**:1155–1162. doi:10.1038/s41587-019-0217-9

813 Yan W, Peng Y-R, Zyl T van, Regev A, Shekhar K, Juric D, Sanes JR. 2020. Cell Atlas of The
814 Human Fovea and Peripheral Retina. *Scientific Reports* **10**:9802. doi:10.1038/s41598-020-
815 66092-9

816 Yu G, Wang L-G, Han Y, He Q-Y. 2012. clusterProfiler: An R Package for Comparing
817 Biological Themes Among Gene Clusters. *OMICS : a Journal of Integrative Biology* **16**:284–
818 287. doi:10.1089/omi.2011.0118

819 Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, Billis K, Cummins C, Gall A,
820 Girón CG, Gil L, Gordon L, Haggerty L, Haskell E, Hourlier T, Izuogu OG, Janacek SH,
821 Juettemann T, To JK, Laird MR, Lavidas I, Liu Z, Loveland JE, Maurel T, McLaren W, Moore B,
822 Mudge J, Murphy DN, Newman V, Nuhn M, Ogeh D, Ong CK, Parker A, Patricio M, Riat HS,
823 Schuilenburg H, Sheppard D, Sparrow H, Taylor K, Thormann A, Vullo A, Walts B, Zadissa A,
824 Frankish A, Hunt SE, Kostadima M, Langridge N, Martin FJ, Muffato M, Perry E, Ruffier M,
825 Staines DM, Trevanion SJ, Aken BL, Cunningham F, Yates A, Flicek P. 2018. Ensembl 2018.
826 *Nucleic Acids Research* **46**:D754–D761. doi:10.1093/nar/gkx1098

827 Zernant J, Xie Y(, Ayuso C, Riveiro-Alvarez R, Lopez-Martinez M-A, Simonelli F, Testa F,
828 Gorin MB, Strom SP, Bertelsen M, Rosenberg T, Boone PM, Yuan B, Ayyagari R, Nagy PL,
829 Tsang SH, Gouras P, Collison FT, Lupski JR, Fishman GA, Allikmets R. 2014. Analysis of the
830 ABCA4 genomic locus in Stargardt disease. *Human Molecular Genetics* **23**:6797–6806.
831 doi:10.1093/hmg/ddu396

832 Zhang D, Guelfi S, Garcia-Ruiz S, Costa B, Reynolds RH, D'Sa K, Liu W, Courtin T, Peterson A,
833 Jaffe AE, Hardy J, Botía JA, Collado-Torres L, Ryten M. 2020. Incomplete annotation has a
834 disproportionate impact on our understanding of Mendelian and complex neurogenetic
835 disorders. *Science Advances* **6**:eaay8299. doi:10.1126/sciadv.aay8299

836 Zhao H, Sun Z, Wang J, Huang H, Kocher J-P, Wang L. 2014. CrossMap: A versatile tool for
837 coordinate conversion between genome assemblies. *Bioinformatics (Oxford, England)*
838 **30**:1006–1007. doi:10.1093/bioinformatics/btt730