A short-read *de novo* transcriptome construction pipeline optimized by long reads reveals novel developmentally regulated gene isoforms and disease targets in hundreds of eye samples

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 previously annotated and novel mRNA transcripts within a given sample. However, this process lacks a way to be evaluated 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— but it is significantly more expensive than short-read sequencing. We generated long-read Pacbio and short-read Illumina RNA sequencing data from an induced pluripotent stem cell- derived retinal pigmented epithelium (iPSC-RPE) cell line. We use the long-read data to identify simple but powerful metrics for assesing *de novo* transcriptome construction and to optimize a short-read based *de novo* transcriptome construction pipeline. We then apply this this pipeline to construct transcriptomes for 340 short-read RNA-seq samples originating from healthy adult and fetal retina, cornea, and RPE to generate the first pan-eye transcriptome annotation. 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 (1), (2), (3). Alternative usage of isoforms has also been implicated in multiple diseases including cancer, cardiovascular disease, Alzheimer's disease and diabetic retinopathy (4), (5), (6), (7).

Accurate annotation of gene isoforms is fundamental for understanding their biological impact. For example, 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 (8), (9). Therefore, identifying novel gene isoforms is a key step in the study of gene isoforms. 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 (10). 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 (11),(2), (12).

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 (13), (14). However, since the GTEx dataset does not include samples from any ocular tissues, the CHESS database remains an incomplete annotation of the human transcriptome.

Despite the increasing number of tools developed, there is no gold standard to evaluate the precision and sensitivity of *de novo* transcriptome construction on real (not simulated) biological data. Long-read sequencing technologies provide a potential solution to this problem as long-read sequencing can capture full length transcripts and thus, can be used to identify a more comprehensive range of gene isoforms. While previous iterations of long-read sequencing technologies typically had higher error rates, the new PacBio Sequel II system sequences long-reads as accurately as short-read based sequencing (15).

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

Finally, we applied this optimized pipeline to a data set containing 340 ocular tissue samples compiled from mining previously published, publicly available short-read RNA-seq data (16). We built transcriptomes for three major ocular tissues: cornea, retina, and RPE, using RNA-seq data from both adult and fetal tissues to create a high-quality pan-eye transcriptome. In addition to ocular samples, we used a subset of the GTEx data set to construct transcriptomes for tissues in 44 other locations across the body.

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

Methods

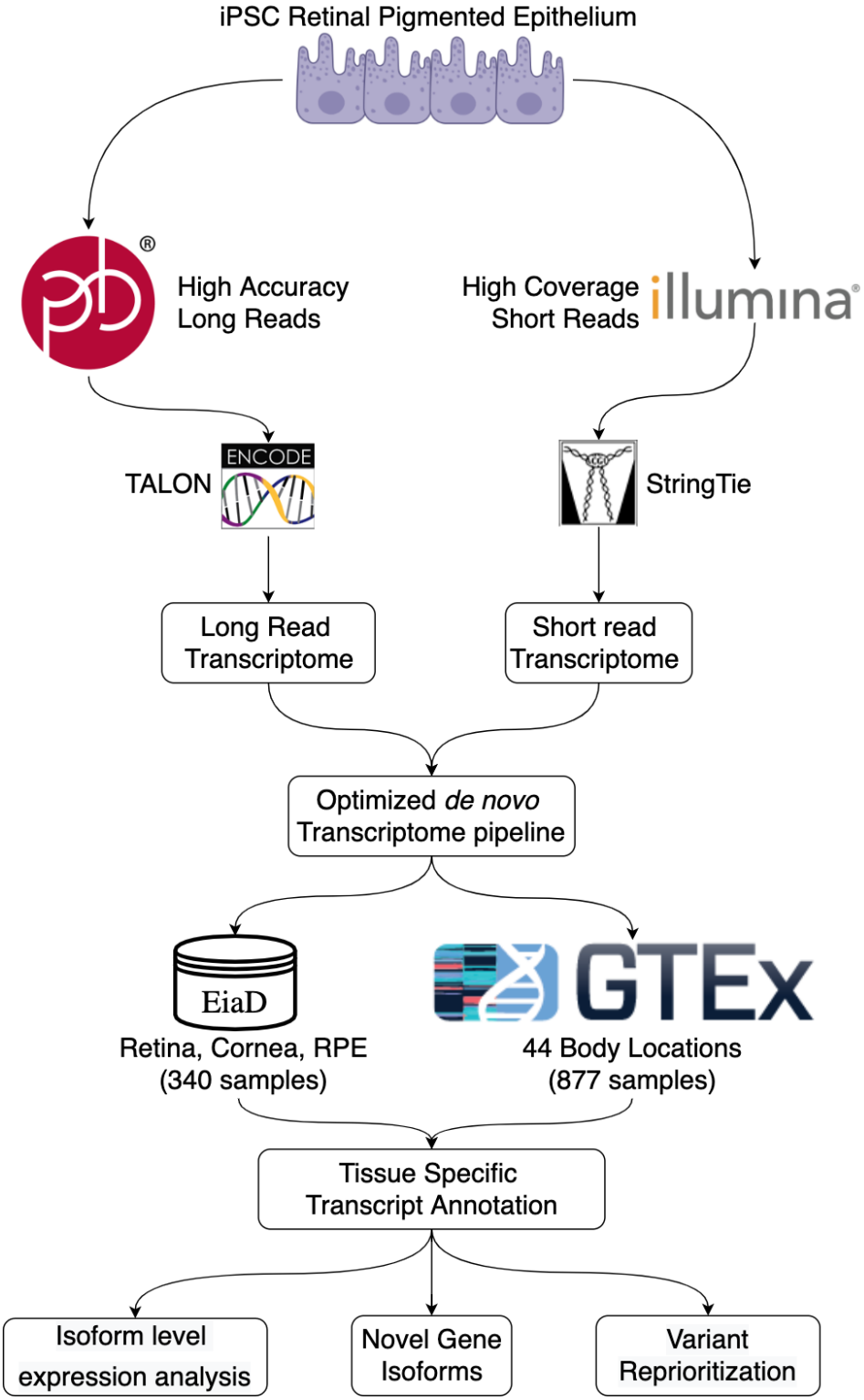


Figure 1. Workflow for long-read informed *de novo* transcriptome construction and analysis.

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

Human iPSCs were differentiated into RPE using previously described protocols in (17) and (18). iPSC-derived RPE (iPSC-RPE) cells at 42 days post differentiation were lysed with TRIzol reagent (Thermo Fisher Scientific; cat # 15596026) and total RNA was isolated using the Direct-zol RNA MiniPrep Kit (Zymo Research, Irvine, CA). 5-6ug total RNA that passed quality control metric (RIN >.9) were used for PacBio library preparation. For PacBio HiFi circular consensus sequencing(CCS), libraries were prepared following the "Procedure-Checklist-Iso-Seq-Express-Template-Preparation-for-Sequel-and-Sequel-II-Systems" protocol. Two libraries were generated: one to capture transcripts 2 kilobases(kb) or smaller, and one to capture transcripts between 2-5kb. Sequencing was done on the PacBio Sequel II system for a movie time of 24 hours.

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

Code availability and software versions.

To improve reproducibility, all code used for both the anaylzing the data and generating the figures for this paper was written as multiple Snakemake pipelines. Each Snakefile contains the exact parameters for all tools and scripts used in each analysis. (19) All code (and versions) used for this project is publicly available in the following github repositories: https://github.com/vinay-swamy/ocular_transcriptomes_pipeline (main pipeline), https://github.com/vinay-swamy/ocular_transcriptomes_longread_analysis (long-read analysis pipeline), https://github.com/vinay-

swamy/ocular_transcriptomes_paper (figures and tables for this paper), https://github.com/vinay-swamy/ocular_transcriptomes_shiny (webapp). Additionally, all Snakefiles are included as supplementary data.(supplementary data files 1-3)

Analysis of long-read data

PacBio sequencing movies were processed into full length, non-chimeric (FLNC) reads using the IsoSeq3 3.1.2 pipeline in the Pacbio SMRT link v7.0 software. The existing ENCODE long-read RNA-seq pipeline (https://github.com/ENCODE-DCC/long-read-rna-pipeline) was rewritten as a Snakemake workflow as follows. Transcripts were aligned to the human genome using minimap2(18), using an alignment index built on the gencode v28 primary human genome. Sequencing errors in aligned long-reads were corrected using TranscriptClean (19) with default parameters. Splice junctions for TranscriptClean were obtained using the TranscriptClean accessory script "get_SJs_from_gtf.py" using the gencode v28 comprehensive transcript annotation as the input. A list of common variants to avoid correcting were obtained from the ENCODE portal (https://www.encodeproject.org/files/ENCFF911UGW/). The long-read transcriptome annotation was generated with TALON (20). A TALON database was generated using the talon_initialize_database command, with all default parameters, except for the "–5P" and "–3p" parameters. These parameters represent the maximum distance between close 5' start and 3' ends of similar transcript to merge and were both set to 100 to match parameters used in later tools. Annotation in GTF format was generated using the talon_create_GTF command, and transcript abundance values were generated using the talon_abundance command.

Analysis of short-read RPE data

Each sample was aligned to the Gencode release 28 hg38 human genome assembly using the genomic aligner STAR and the resulting BAM files were sorted using samtools sort. (8),(20),(21). For each sorted BAM file, a per-sample base transcriptome was constructed using StringTie with the Gencode V28 comprehensive annotation as a guiding annotation (8),(12). All sample transcriptomes were merged with the long-read transcriptome using gffcompare(22) with default parameters. We note that the default

values for the distance to merge similar 5' starts and 3 ends of transcripts in gffcompare is the same to what we chose for TALON. We defined the metric construction accuracy, used to evaluate short-read transcriptome construction as the following:

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

Construction of subtissue-specific transcriptomes.

We used studies with healthy, unperturbed RNA-seq samples from 50 distinct locations of the body. We downloaded and performed quality control of the pertinent sequencing data from the sequence read archive (SRA) using methods from our previous work (16). We constructed a transcriptome for each sample, and merged samples together to create 50 subtissue-specific transcriptomes. We define subtissue as a unique body location and are either temporally different versions of the same tissue(adult vs fetal tissue), or different regions of a larger tissue (cortex vs cerebellum in brain). Tissue refers to complete whole tissues (retina, brain, liver). For each subtissue-specific transcriptome, we removed transcripts that had an average expression less than 1 Transcripts Per Million (TPM) across all samples of the same subtissue type. All subtissue-specific transcriptomes were merged to form a single unified annotation file in general transfer format(GTF) to ensure transcript identifiers were the same across subtissues. We merged all ocular subtissue transcriptomes to generate a separate pan-eye transcriptome.

Subtissue specific transcriptome quantification

For each resulting subtissue specific transcriptome, we extracted transcript sequences using the tool gffread and used these sequences to build a subtissue-specific quantification index using the index mode of the alignment-free quantification tool Salmon (22), (23). For each sample, we quantified transcript expression using the quant mode of Salmon, using a sample's respective subtissue specific quantification index. We similarly quantified all ocular samples using the pan-eye transcriptome and the Gencode v28 reference transcriptome.

Annotation of novel exons

Analysis of novel transcripts was done using a custom Rscript "annotate_and_make_tissue_gtfs.R" . First, a comprehensive set of distinct, annotated exons was generated by merging exon annotation from gencode, ensembl, UCSC, and refseq. We then defined a novel exon as any exon within our transcriptomes that does not exactly match the chromosome, start, end and strand of an annotated exon. Novels exons were classified by splitting exons into 3 categories: first, last, and middle exons. We then extracted all annotated exon start and stop sites from our set of previously annotated exons. Novel middle exons that have an annotated start but an unannotated end were categorized as a novel alternative 3' end exons and similarly novel middle exons with an unannotated start but annotated end were categorized as a novel alternative 5' start exons. Novel middle exons whose start and end match annotated exon start and ends were considered retained introns. Novel middle exons whose start and end do not match annotated starts and ends were considered fully novel exons. We then classified novel first and last exons. Novel first exons were first exons whose start is not in the set of annotated exon starts, and novel last exons were terminal exons whose end is not in the set of annotated exon ends.

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

PhyloP scores for the phylop 20-way multi species alignment were downloaded from UCSC's FTP server on October 16th, 2019 and converted from bigWig format to bed format using the wig2bed tool in BEDOPs (24), (25). The average score per exon in both the gencode and DNTX annotation was calculated by intersecting exon locations with phylop scores and then averaging the per base score for each exon, using the intersect and groupby tools from the bedtools suite, respectively. Significant difference in mean phylop score was tested with a Mann Whitney U test.

CAGE peaks were download from the FANTOM FTP server (https://fantom.gsc.riken.jp/5/datafiles/reprocessed/hg38_latest/extra/CAGE_peaks/hg38_fair+new_CAGE_peaks_phase1and2.bed.gz) on June 15th 2020 (26). Transcriptional start sites (TSS) were extracted from gencode and DNTX annotations; TSS is defined as the start

of the first exon of a transcript. Distance to CAGE peaks was calculated using the closest tool in the bedtools suite. Significant difference in mean distance to CAGE peak between DNTX and gencode annotation was tested with a Mann Whitney U test.

Polyadenylation signal annotations were downloaded from the polyA site atlas (https://polyasite.unibas.ch/download/atlas/2.0/GRCh38.96/atlas.clusters.2.0.GRCh38.96.bed.gz) on June 15th 2020 (27). Transcriptional end sites(TES) were extracted from gencode and DNTX annotations; TES is defined as the end of the terminal exon of a transcript. Distance to polyA signal was calculated using the closest tool in the bedtools suite (28). Significant difference in mean distance to polyA signal was tested with a Mann Whitney U test.

Identification of novel protein coding transcripts

Protein-coding transcripts in the unified transcriptome were identified using the TransDecoder suite (11). Transcript sequences in fasta format were extracted from the final pan-body transcriptome using the TransDecoder util script "gtf_genome_to_cdna_fasta.pl". Potential open reading frames(ORFs) were generated from transcript sequences using the LongestORF module within TransDecoder, and the single best ORF for each transcript was extracted with the Predict module within Transdecoder. The resulting ORFs were mapped to genomic locations with the TransDecoder util script "gtf_to_alignment_gff3.pl". For each ORF start and stop codons were extracted with the script "agat_sp_add_start_stop.pl" scripts from the AGAT toolkit (https://github.com/NBISweden/AGAT/). Transcripts with no detectable ORF or missing a start or stop codon were labelled as noncoding.

Analysis of novel isoforms in eye tissues

An Upset plot was generated using the ComplexUpset package (https://github.com/krassowski/complex-upset) (29). Fraction Isoform Usage(FIU) was calculated for each transcript t associated with a parent gene g using the following formula: $FIU_t = \frac{TPM_t}{TPM_g}$. Raincloud plots of FIU were generated using the R_Rainclouds package (30).

Analysis of fetal retina RNA-seq data.

RNA-seq samples from Mellough et al. were downloaded from the SRA using methods from a previous study (16). Samples were quantified using Salmon with a quantification index generated using our fetal retina *de novo* transcriptome. Outliers within the dataset were identified by first performing principal component analysis of transcript level expression data, calculating the center of all data using the first two principal components, and subsequently removing five samples furthest away from the center of all data. The remaining samples were normalized using calcNormFactors from the R package edgeR and converted to weights using the voom function from the R package limma (31), (32). Differential expression was modeled using the lmFit function using developmental time point as the model design and tested for significant change in expression using the Ebayes function from limma. Gene Set enrichment was tested using the R package clusterprofileR (33). Heatmaps were generated using the ComplexHeatmap package (34).

Prediction of variant impact using *de novo* transcriptomes.

Noncoding variants previously associated with retinal disease from the Blueprint Genetics Retinal dystrophy panel were obtained from the Blueprint Genetics website (https://blueprintgenetics.com/tests/panels/ophthalmology/retinal-dystrophy-panel/). The variants were converted from HGVS to VCF format using a custom python script "HGVS_to_VCF.py". This VCF was then remapped to the hg38 human genome build using the tool crossmap (35). The VCF of variants was used as the input variants for the Variant Effect Predictor(VEP) tool from Ensembl, with each subtissue specific transcriptome as the input annotation (36). VEP was additionally run using the gencode V28 comprehensive annotation as the input annotation to identify variants whose predicted impact increased in severity.

Figures, Tables, and Computing Resources

All statistical analyses, figures and tables in this paper were generated using the R programming language. (37) A full list of packages and versions can be found in the supplementary file session_info.txt. All computation was performed on the National Institutes of Health high performance computer system Biowulf (hpc.nih.gov).

Results

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. These cells were differentiated using an optimized protocol, and thus minimal biological variation is expected (38), (39). 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

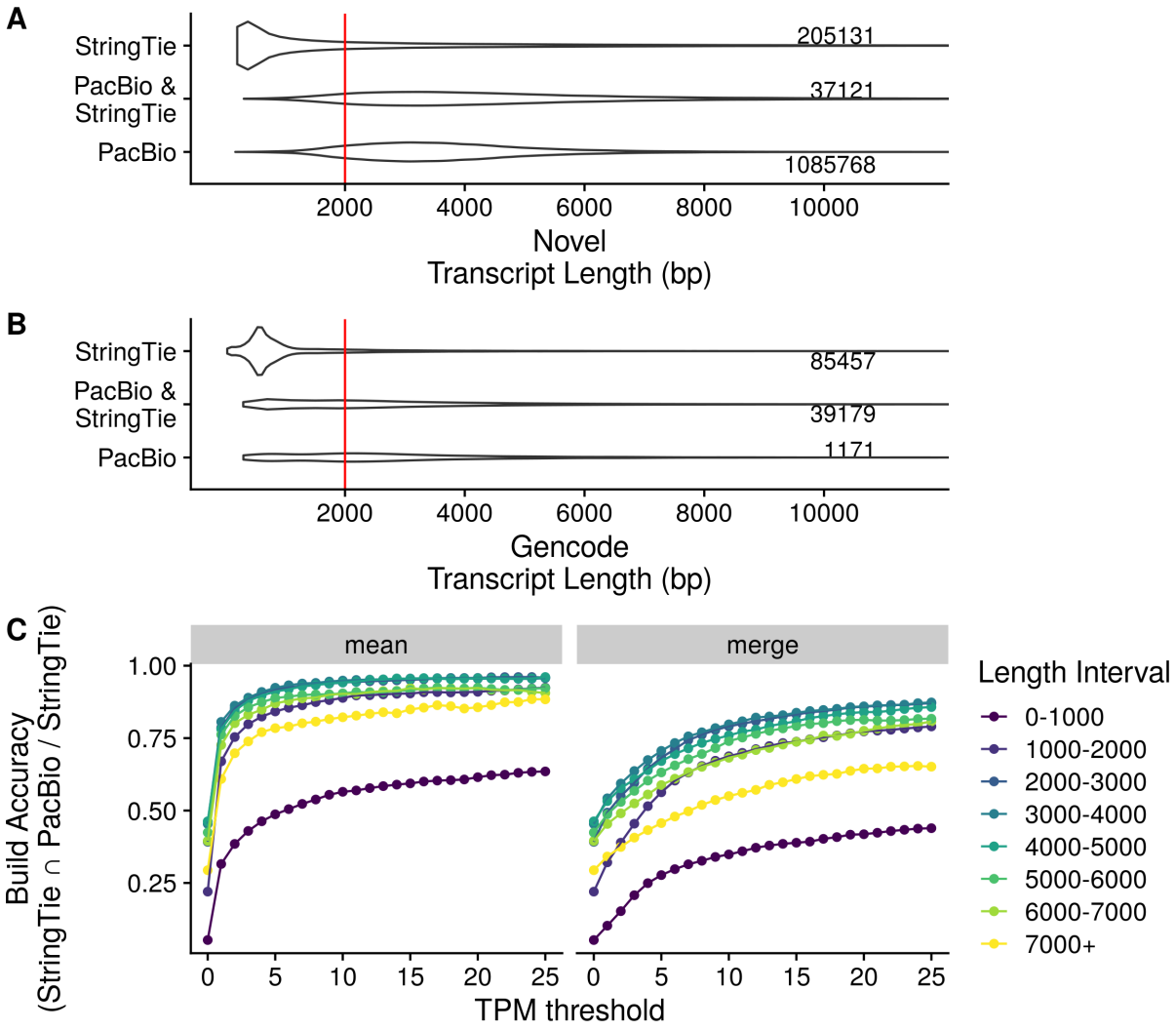


Figure 2. Transcript length and expression dictate transcriptome construction accuracy. A,B) Distributions of novel(A) and previously annotated(B) transcript lengths between Pacbio (long-read) and Stringtie (short-read) transcriptomes. Each distribution is labeled with the total number of transcripts in the distribution C) short-read construction accuracy stratified by transcript length at different Transcripts Per Million (TPM)-based transcript exclusion thresholds.

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 (40). 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
Retina	Adult	105	8	49714
RPE	Fetal	49	7	49967
Cornea	Adult	43	6	51469
Retina	Fetal	89	6	66255
RPE	Adult	48	4	32012
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 published, publicly available ocular tissue RNA-seq samples curated in EiaD using an efficient Snakemake pipeline (19). We included both 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 (41). To more accurately determine the tissue specificity of novel ocular transcripts, we supplemented our publicly collated normal (non-disease, non-perturbed) 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 (13). 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 novel as any region of the human genome that has not been previously annotated within the Gencode, Ensembl, UCSC, and Refseq annotation databases (8) , (42) , (43). 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 megabases of previously unannotated genomic sequence (Table 1). We refer to the final pan-body transcriptome as the DNTX annotation hereafter.

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

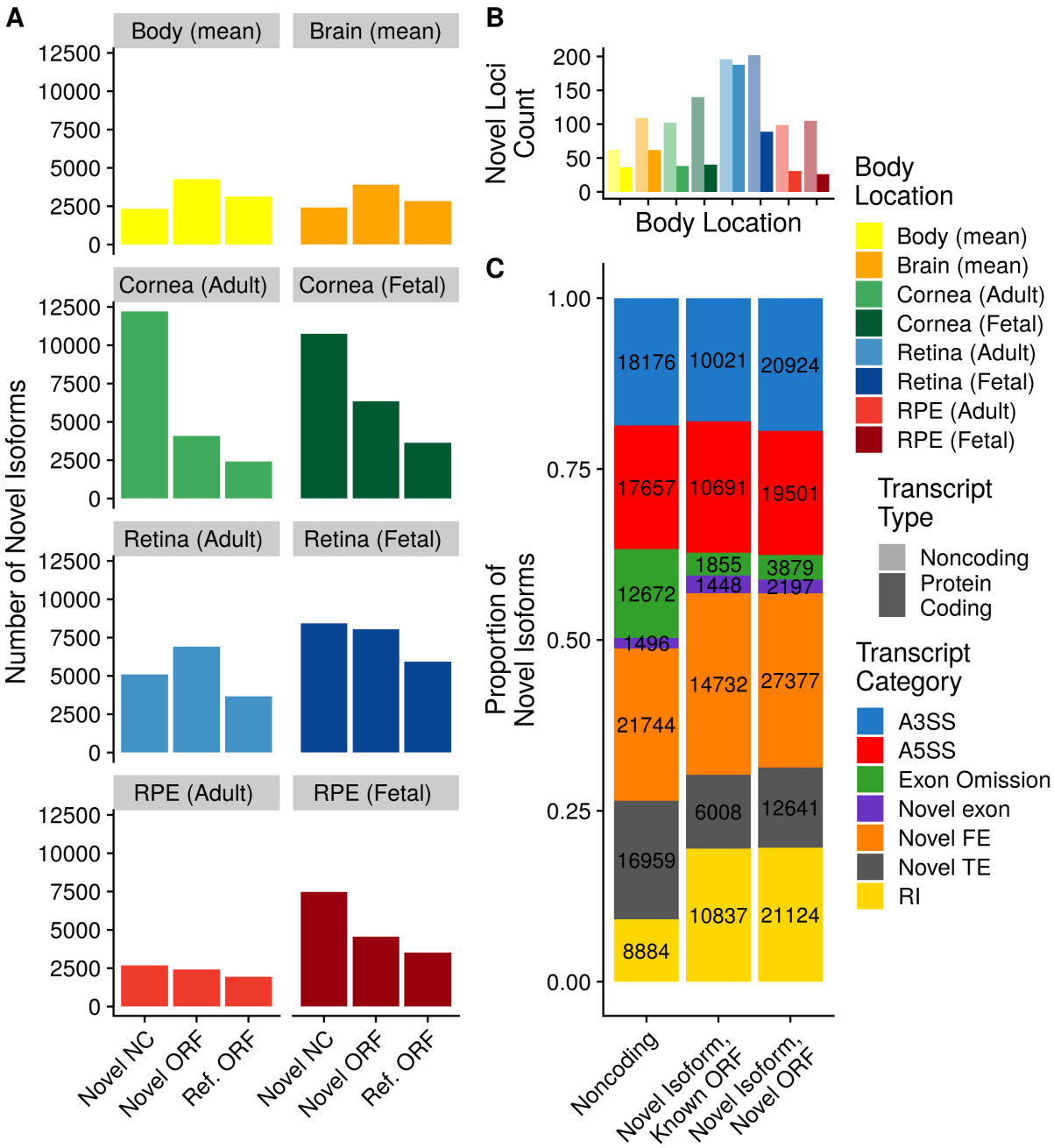


Figure 3. Overview of novel isoforms. A) Number of novel gene isoforms, grouped by transcript type. Brain and body represent an average of 13 and 34 distinct subtissues, respectively. B) Novel protein coding and noncoding loci. Novel exon composition of novel isoforms, by isoform type. Labels indicate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

Novel isoforms can occur due to an omission of a previously annotated exon, commonly referred as exon skipping or the addition of an unannotated exon which we refer to as a novel exon. We further classified novel exons by the biological process that may be driving their formation: alternative promoter usage driving the addition of novel first exons (FE), alternative polyadenylation driving the addition of novel terminal exons (TE), and alternative splicing driving the formation of all novel exons that are not the first or last exon (44), (45), (46). We then split alternatively spliced exons into their commonly seen patterns, alternative 5' splice site (A5SS), alternative 3' splice site (A3SS), and retained introns (RI). Exons whose entire sequence was unannotated and is not a retained intron are fully novel exons. We note that all three of these mechanisms can lead to exon skipping, so for simplicity we grouped all novel isoforms resulting from exon skipping together. We found that the majority of novel exons within our dataset are novel FEs. We noticed that the majority of RI exons lead to novel ORFs, whereas novel isoforms with omitted exons more often lead to noncoding isoforms. (Fig 3C)

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

We validated *de novo* transcriptomes using three independent approaches. We first looked for evolutionary conservation since it is commonly accepted as a proxy for functional significance. We used the PhyloP 20 way species alignment, a measure of conservation between species, to calculate the average conservation score for each exon in the DNTX annotation and compared that to the average conservations score for each exon in the GENCODE annotation (24). We found that, on average, exons in the DNTX annotation are more conserved than exons in the GENCODE annotation (pvalue <2.2e-16) (Supplemental Figure 2A).

Next, since we observed an enrichment in novel first and last exons within our data set, we decided to compare the TSS and TES within the DNTX annotation to two well-established annotation databases from FANTOM and the polyA Atlas (26), (27). We compared DNTX and GENCODE TSS's to CAGE-seq data from the FANTOM consortium; as CAGE-seq is optimized to detect the 5' end of transcripts, we reasoned that it can serve as a valid ground truth set to evaluate TSS detection (47). We calculated the absolute distance

of DNTX TSS's to CAGE peaks, and compared them to the absolute distance of GENCODE TSS's to CAGE peaks. We found that, on average, DNTX TSS's were closer to CAGE peaks than GENCODE TSS's (pvalue <2.2e-16)(Supplemental Figure 2B).

Finally, we evaluated TES's using the polyA Atlas, which is comprised of polyadenylation signal annotation generated from aggregating 3' seq data from multiple studies. As 3'-seq data is designed to accurately capture the 3' ends of transcripts, it can similarly serve as a ground truth set to evaluate the accuracy of TES's (48). We calculated the absolute distance of DNTX TES's to annotated polyA signals and compared them to the absolute distance of GENCODE TES's to polyA signals. We found that on average DNTX TES's are closer to annotated polyadenylation signals than gencode TSS's (pvalue <2.2e-16) (Supplemental Figure 2C)

De novo transcriptomes reduce overall transcriptome sizes

De novo transcriptomes removed on average 76.141 % of a subtissue's base transcriptome. We defined base transcriptome for a subtissue as any transcript in the GENCODE annotation with non-zero TPM in at least one sample of a given subtissue. This was a large reduction in transcriptome size and we wanted to ensure that we were not unduly discarding data. We quantified transcript expression of each sample using Salmon with two methods: once using the full gencode V28 human transcript annotation, and once using its associated subtissue specific transcriptome. We found that despite the 76.141 % reduction in number of transcripts between the base gencode and *de novo* transcriptomes (Supplemental Figure 3A), the average Salmon mapping rate increased by 2.041 % indicating that the vast majority of gene expression data is retained within our transcriptome (Supplemental Figure 3B).

Novel Isoforms are identified in ocular tissues

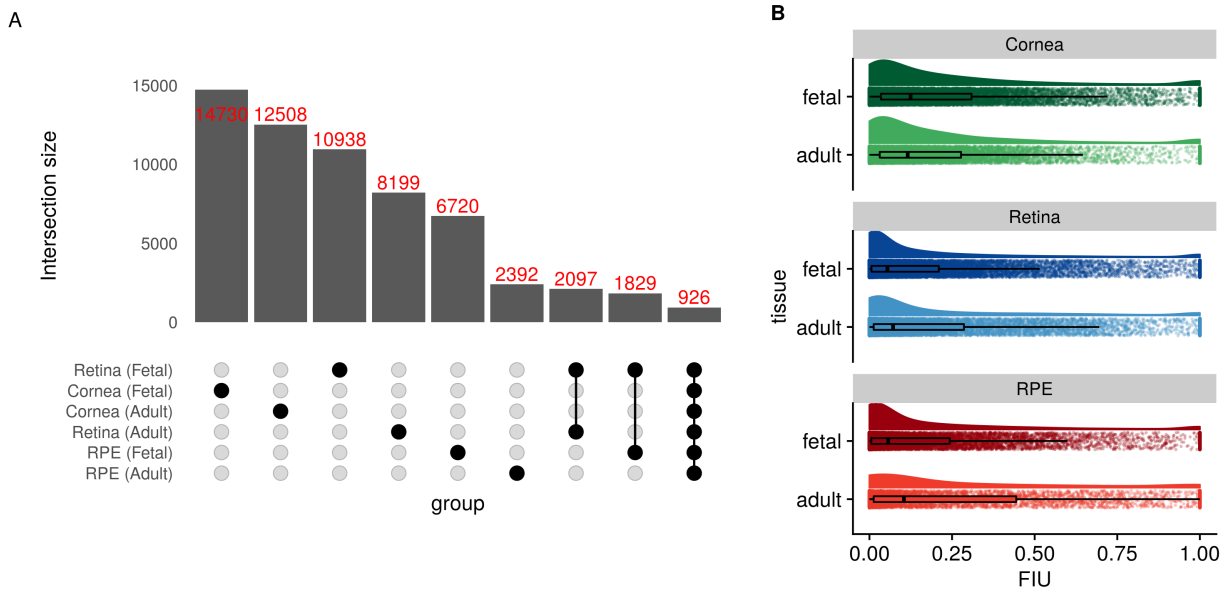


Figure 4. Overview of novel gene isoforms in the eye. A) Set intersection of novel isoforms in ocular transcriptomes. B) Boxplots of fraction isoform usage (FIU) overlaid over FIU data points with estimated distribution of data set above each boxplot.

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

Differential usage of gene isoforms occurs during retinal development

Multiple studies have shown that gene isoforms play a significant role in eye development (49), (50). We hypothesized that the DNTX annotation provides additional insight into alternative isoform usage and identifies novel gene isoforms potentially involved in eye development. We used RNA-seq data of the developing retina from Mellough et al, an independent data set that we did not include for transcriptome construction, and used a subset of the DNTX annotation corresponding to fetal retina to

quantify transcript expression and identify transcripts with significant changes in expression across retinal development. Transcripts that are differentially expressed (qvalue <.01) and have a mean FIU difference of .25 in at least one comparison of time points are indicative of differential transcript usage (DTU).

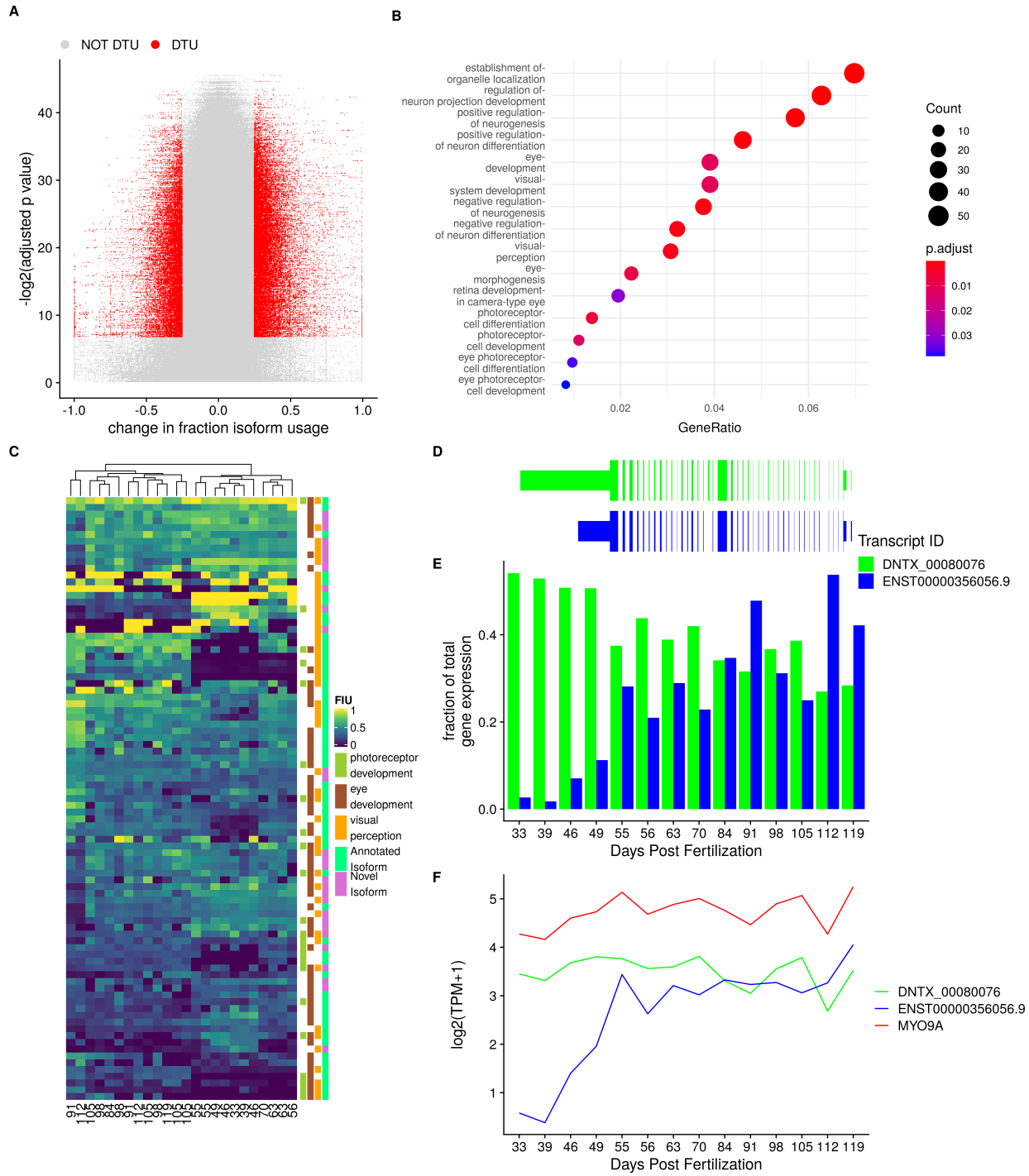


Figure 5 Differential Transcript usage during Retinal Development. A) Volcano Plot of tested transcripts B) Dot plot for gene set enrichment analysis C) Heatmap of hiearchical clustering of transcripts with DTU associated with eye development D) Transcript models for *MYO9A*, a gene undergoing DTU F)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

We analyzed 24 samples across 14 developmental days post fertilization and found 1717 transcripts across 812 genes displaying DTU (Fig 5A). We found that genes involved in DTU are enriched(qvalue <.05) for genes related to eye and neurological development (Fig 5B), and that hierarchical clustering of DTU transcripts generates an early stage and late stage cluster (Fig 5C). One of these genes,*MYO9A*, is a classical example of DTU. *MYO9A* is associated with the visual perception GO term, plays a role in ocular development, and has been associated with ocular disease (51). While expression of *MYO9A* remains relatively unchanged across development, expression of two of its associated isoforms in fetal retina (Fig 5D) changes dramatically during development: a novel isoform is highly expressed early during development, but switched to the canonical isoform later in development (Fig 5E,F). This novel isoform contains a novel exon within the protein coding region of the isoform as well as novel last exon extending the 3' UTR (Fig 5d).A full list of genes and transcripts displaying DTU is available in Supplemental data (supplemental data 4).

De novo transcriptomes allow for a more precise variant prioritization.

The identification of a disease-causing variant through genome sequencing is a common step in diagnosing genetic disease, when disease causing variants cannot be determined from exonic sequencing. Prediction of a variant's biological impact and subsequent variant prioritization is a fundamental step in this process. Many methods for predicting variant effects on protein function or gene expression are based on location within the body of a transcript; for example variants that disrupt splice sites and start/stop codons are considered to be the most damaging, while variants within intronic and intergenic regions have unknown impact or are not classified, and, thus, are not included for further consideration. However, multiple studies have identified pathogenic deep intronic variants for retinal dystrophies (52), (53), (54), (55), (56), (57), (58). Pathogenic intronic variants are thought to function by introducing a novel splice site, disrupting

regulatory motifs, or altering a tissue-specific transcript. To explore this third possibility, we mapped known pathogenic intronic variants 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.

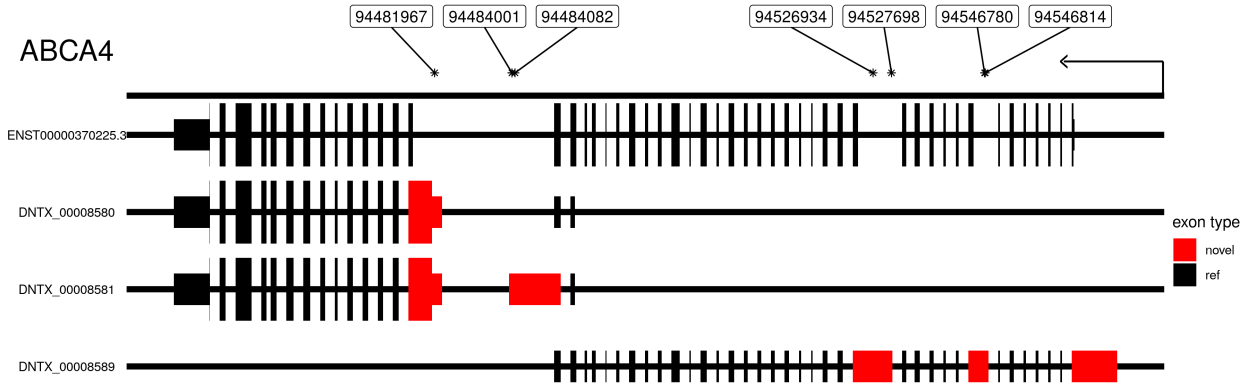


Figure 6. Transcript models for selected Isoforms of *ABCA4* along with location of pathogenic intronic variants. Location is on the hg19 human genome build. Thick lines indicate protein coding regions. Arrow indicates direction of transcription. Introns not drawn to scale

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 (59). 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

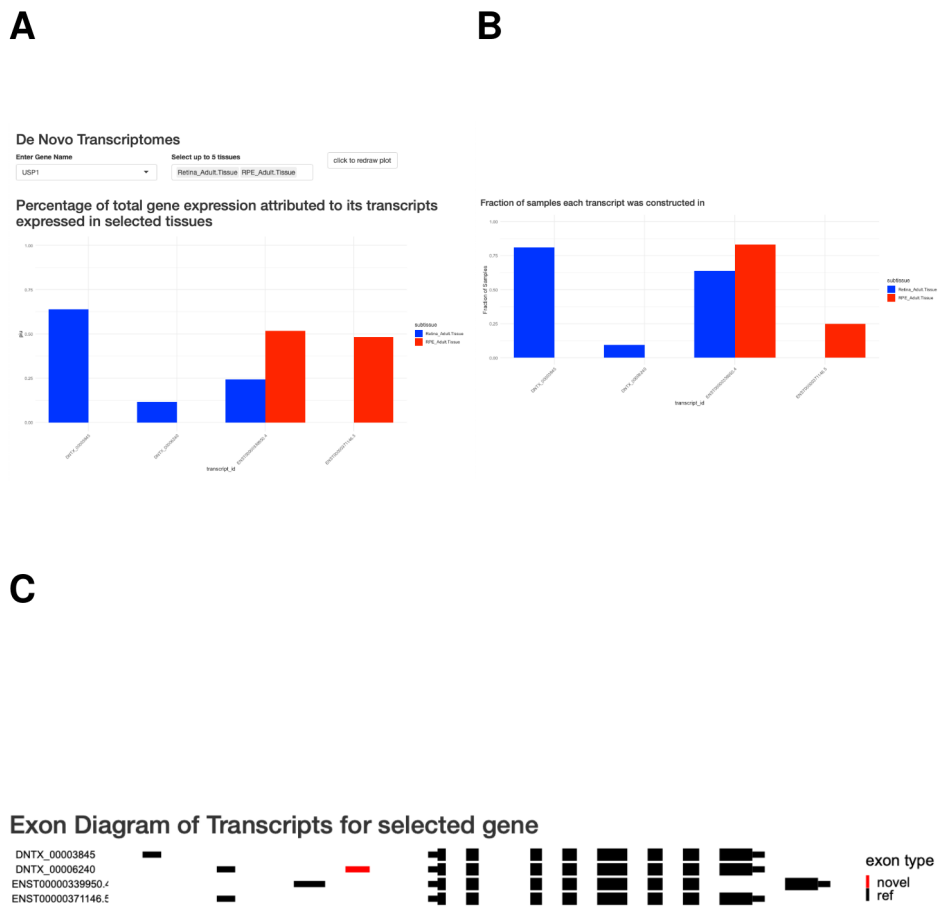


Figure 7. Screenshots from dynamic *de novo* transcriptome visualization tool. A). FIU bar plot for selected gene and subtissue. B). Exon level diagram of transcript body Thicklines represent coding region of transcript. novel exons colored in red. Tooltip contains genomic location and phylop score C)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

phylogenetic conservation score (Fig 7B). We additionally show a barplot of the fraction of samples each transcript was constructed in (Fig 7C). Users can also download the *de novo* transcriptomes for selected subtissues in GTF and fasta format. Instructions to download and run the app are available at https://github.com/vinay-swamy/ocular_transcriptomes_shiny. While visualization of direct transcript expresion is not a part of this app, it can be viewed in the eyeIntegration app (16) by selected 'DNTX' as the transcript annotation. Finally, we package all tools used for our transcriptome pipeline within a portable docker container with a stand-alone run script. This pipeline allows other researchers to run their own samples, and generate figures and annotations similar to what is shown here, available at https://github.com/vinay-swamy/ocular_transcriptomes_pipeline

Discussion

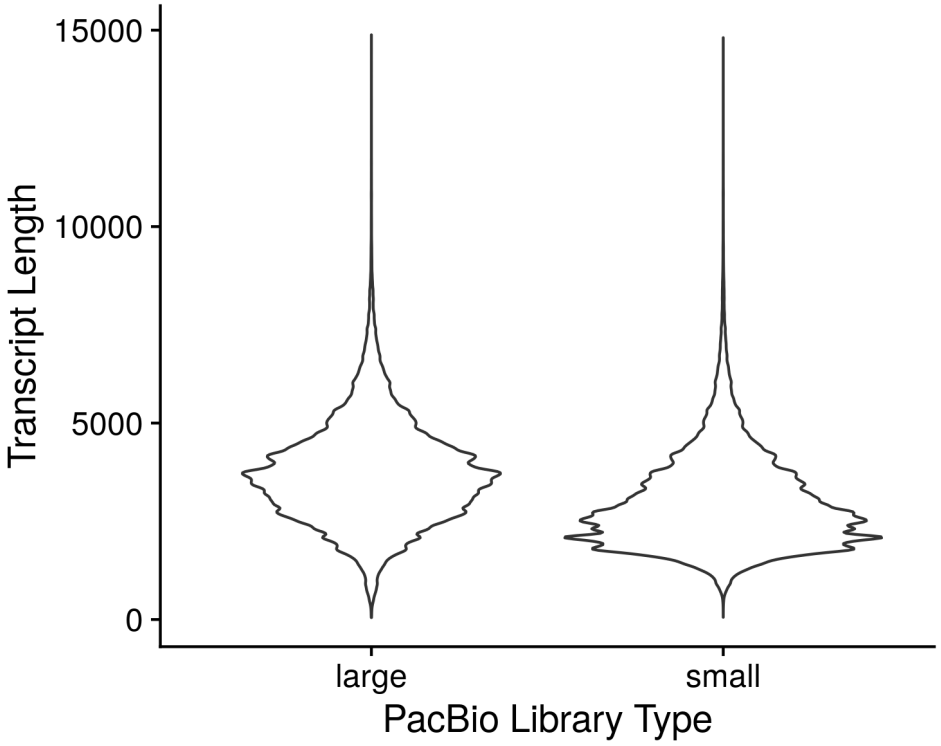
Motivated by the lack of a comprehensive transcriptome for the eye, we constructed transcriptomes for adult and fetal retina, RPE and cornea. By using long-read RNA-seq data to calibrate our short-read construction pipeline, we were able to identify biologically relevant transcriptomes. We found that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

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

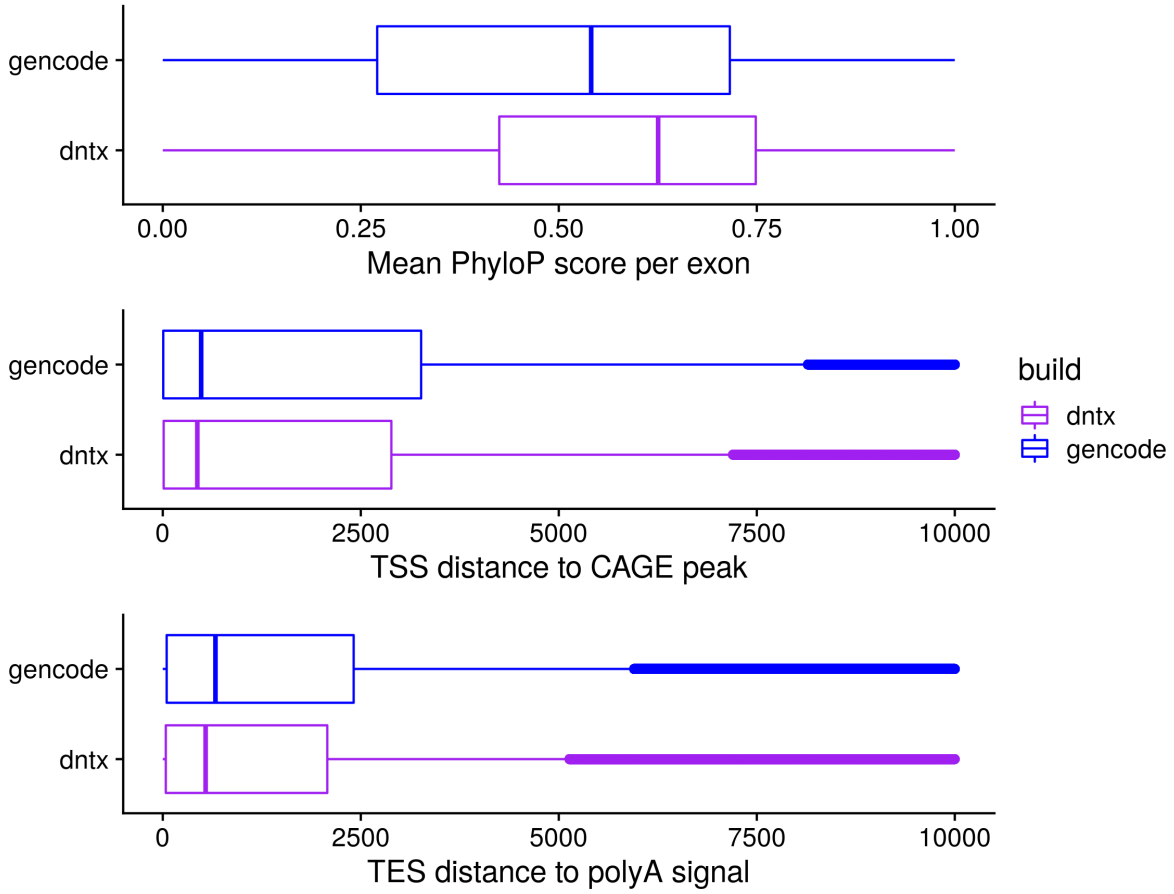
In each ocular subtissue we examined, we found hundreds of novel gene isoforms, many of which were novel due to novel exons. Within ocular subtissues, these novel isoforms are most commonly specific to single subtissue. This makes sense as a majority of the exons in our *de novo* transcriptomes are first and last exons, which have been previously shown to significantly contribute to the tissue specificity of gene isoforms (60). We also found that on average novel isoforms represent about 20.584 % of their parent gene's expression. Future studies are needed to identify the function of these isoforms. One possibility is that some of these isoforms are only expressed in rare cell types, as transcript annotation was previously shown to be incomplete in rare cell types (9). This especially makes sense in the retina which contain over a dozen distinct cell types, several of which contribute to 5% or less of the total cell population (61). As we imposed a strict expression filter as part of our transcriptome pipeline, we may have removed transcripts specific to rare cell types.

In conclusion, we created the first pan-eye transcriptome annotation and showed that it is useful in understanding the role of gene isoforms in ocular biology and improving the ability to diagnose inherited eye diseases. This work is most useful as a starting point for other researchers; we want to make the transcriptomes easily accessible to other researchers, so we designed a webapp for visualization and to access tissue-specific annotation files. We believe this project will enable other researchers to explore new research directions and answer long pending questions

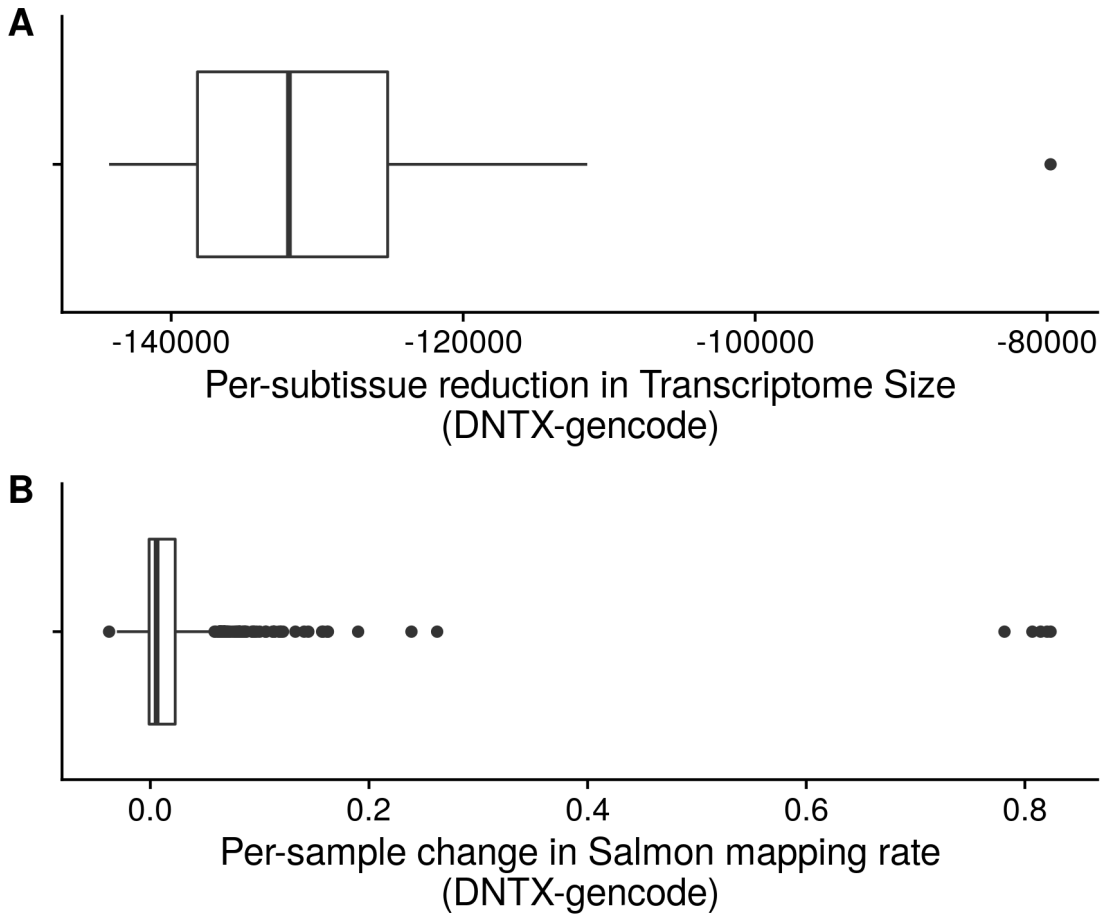
Supplemental Figures



Supplemental Figure 1. Distribution of PacBio long-read lengths for two 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

1. Dykes,I.M., Bueren,K.L. van and Scambler,P.J. (2018) HIC2 regulates isoform switching during maturation of the cardiovascular system. *Journal of Molecular and Cellular Cardiology*, **114**, 29–37.

2. Trapnell,C., Williams,B.A., Pertea,G., Mortazavi,A., Kwan,G., Baren,M.J. van, Salzberg,S.L., Wold,B.J. and Pachter,L. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, **28**, 511–515.

3. Mitra,M., Lee,H.N. and Coller,H.A. (2020) Splicing Busts a Move: Isoform Switching Regulates Migration. *Trends in Cell Biology*, **30**, 74–85.

4. Vitting-Seerup,K. and Sandelin,A. (2017) The Landscape of Isoform Switches in Human Cancers. *Molecular Cancer Research*, **15**, 1206–1220.

5. Neagoe Ciprian, Kulke Michael, del Monte Federica, Gwathmey Judith K., de Tombe Pieter P., Hajjar Roger J. and Linke Wolfgang A. (2002) Titin Isoform Switch in Ischemic Human Heart Disease. *Circulation*, **106**, 1333–1341.

6. Mills,J.D., Nalpathamkalam,T., Jacobs,H.I.L., Janitz,C., Merico,D., Hu,P. and Janitz,M. (2013) RNA-Seq analysis of the parietal cortex in Alzheimer's disease reveals alternatively spliced isoforms related to lipid metabolism. *Neuroscience Letters*, **536**, 90–95.

7. Perrin,R.M., Konopatskaya,O., Qiu,Y., Harper,S., Bates,D.O. and Churchill,A.J. (2005) Diabetic retinopathy is associated with a switch in splicing from anti- to pro-angiogenic isoforms of vascular endothelial growth factor. *Diabetologia*, **48**, 2422–2427.

8. Frankish,A., Diekhans,M., Ferreira,A.-M., Johnson,R., Jungreis,I., Loveland,J., Mudge,J.M., Sisu,C., Wright,J. and Armstrong,J. *et al.* (2019) GENCODE reference annotation for the human and mouse genomes. *Nucleic acids research*, **47**, D766–D773.

9. Zhang,D., Guelfi,S., Garcia-Ruiz,S., Costa,B., Reynolds,R.H., D'Sa,K., Liu,W., Courtin,T., Peterson,A. and Jaffe,A.E. *et al.* (2020) Incomplete annotation has a disproportionate impact on our understanding of Mendelian and complex neurogenetic disorders. *Science Advances*, **6**, eaay8299.

10. Nagalakshmi,U., Wang,Z., Waern,K., Shou,C., Raha,D., Gerstein,M. and Snyder,M. (2008) The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing. *Science*, **320**, 1344–1349.

11. Haas,B.J., Papanicolaou,A., Yassour,M., Grabherr,M., Blood,P.D., Bowden,J., Couger,M.B., Eccles,D., Li,B. and Lieber,M. *et al.* (2013) De novo transcript sequence reconstruction from RNA-Seq: Reference generation and analysis with Trinity. *Nature protocols*, **8**.

12. Pertea,M., Pertea,G.M., Antonescu,C.M., Chang,T.-C., Mendell,J.T. and Salzberg,S.L. (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*, **33**, 290–295.

13. GTEx Consortium, Laboratory, Data Analysis &Coordinating Center (LDACC)—Analysis Working Group, Statistical Methods groups—Analysis Working Group, Enhancing GTEx (eGTEx) groups, NIH Common Fund, NIH/NCI, NIH/NHGRI, NIH/NIMH, NIH/NIDA and Biospecimen Collection Source Site—NDRI *et al.* (2017) Genetic effects on gene expression across human tissues. *Nature*, **550**, 204–213.

14. Pertea,M., Shumate,A., Pertea,G., Varabyou,A., Breitwieser,F.P., Chang,Y.-C., Madugundu,A.K., Pandey,A. and Salzberg,S.L. (2018) CHESS: A new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise. *Genome Biology*, **19**, 208.

15. Wenger,A.M., Peluso,P., Rowell,W.J., Chang,P.-C., Hall,R.J., Concepcion,G.T., Ebler,J., Fungtammasan,A., Kolesnikov,A. and Olson,N.D. *et al.* (2019) Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology*, **37**, 1155–1162.

16. Swamy,V. and McGaughey,D. (2019) Eye in a Disk: eyeIntegration Human Pan-Eye and Body Transcriptome Database Version 1.0. *Investigative Ophthalmology & Visual Science*, **60**, 3236–3246.

17. Bryan,J.M., Fufa,T.D., Bharti,K., Brooks,B.P., Hufnagel,R.B. and McGaughey,D.M. (2018) Identifying core biological processes distinguishing human eye tissues with precise systems-level gene expression analyses and weighted correlation networks. *Human Molecular Genetics*, **27**, 3325–3339.

18. May-Simera,H.L., Wan,Q., Jha,B.S., Hartford,J., Khristov,V., Dejene,R., Chang,J., Patnaik,S., Lu,Q. and Banerjee,P. *et al.* (2018) Primary Cilium-Mediated Retinal Pigment Epithelium Maturation Is Disrupted in Ciliopathy Patient Cells. *Cell reports*, **22**, 189–205.

19. Köster,J. and Rahmann,S. (2012) Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, **28**, 2520–2522.

20. Dobin,A., Davis,C.A., Schlesinger,F., Drenkow,J., Zaleski,C., Jha,S., Batut,P., Chaisson,M. and Gingeras,T.R. (2013) STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)*, **29**, 15–21.

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

22. Pertea,G. and Pertea,M. (2020) GFF Utilities: GffRead and GffCompare. *F1000Research*, **9**, 304.

23. Patro,R., Duggal,G., Love,M.I., Irizarry,R.A. and Kingsford,C. (2017) Salmon provides fast and bias-aware quantification of transcript expression. *Nature methods*, **14**, 417–419.

24. Pollard,K.S., Hubisz,M.J., Rosenbloom,K.R. and Siepel,A. (2010) Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Research*, **20**, 110–121.

25. Neph,S., Kuehn,M.S., Reynolds,A.P., Haugen,E., Thurman,R.E., Johnson,A.K., Rynes,E., Maurano,M.T., Vierstra,J. and Thomas,S. *et al.* (2012) BEDOPS: High-performance genomic feature operations. *Bioinformatics*, **28**, 1919–1920.

26. Noguchi,S., Arakawa,T., Fukuda,S., Furuno,M., Hasegawa,A., Hori,F., Ishikawa-Kato,S., Kaida,K., Kaiho,A. and Kanamori-Katayama,M. *et al.* (2017) FANTOM5 CAGE profiles of human and mouse samples. *Scientific Data*, **4**, 170112.

27. Herrmann,C.J., Schmidt,R., Kanitz,A., Artimo,P., Gruber,A.J. and Zavolan,M. (2020) PolyASite 2.0: A consolidated atlas of polyadenylation sites from 3′ end sequencing. *Nucleic Acids Research*, **48**, D174–D179.

28. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)*, **26**, 841–842.

29. Lex,A., Gehlenborg,N., Strobelt,H., Vuillemot,R. and Pfister,H. (2014) UpSet: Visualization of Intersecting Sets. *IEEE Transactions on Visualization and Computer Graphics*, **20**, 1983–1992.

30. Allen,M., Poggiali,D., Whitaker,K., Marshall,T.R. and Kievit,R.A. (2019) Raincloud plots: A multi-platform tool for robust data visualization. *Wellcome Open Research*, **4**, 63.

31. Robinson,M.D., McCarthy,D.J. and Smyth,G.K. (2010) edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.

32. Ritchie,M.E., Phipson,B., Wu,D., Hu,Y., Law,C.W., Shi,W. and Smyth,G.K. (2015) Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, **43**, e47–e47.

33. Yu,G., Wang,L.-G., Han,Y. and He,Q.-Y. (2012) clusterProfiler: An R Package for Comparing Biological Themes Among Gene Clusters. *OMICS : a Journal of Integrative Biology*, **16**, 284–287.

34. Gu,Z., Eils,R. and Schlesner,M. (2016) Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, **32**, 2847–2849.

35. Zhao,H., Sun,Z., Wang,J., Huang,H., Kocher,J.-P. and Wang,L. (2014) CrossMap: A versatile tool for coordinate conversion between genome assemblies. *Bioinformatics (Oxford, England)*, **30**, 1006–1007.

36. McLaren,W., Gil,L., Hunt,S.E., Riat,H.S., Ritchie,G.R.S., Thormann,A., Flicek,P. and Cunningham,F. (2016) The Ensembl Variant Effect Predictor. *Genome Biology*, **17**, 122.

37. R Core Team (2019) R: A Language and Environment for Statistical Computing R Foundation for Statistical Computing, Vienna, Austria.

38. Blenkinsop,T.A., Saini,J.S., Maminishkis,A., Bharti,K., Wan,Q., Banzon,T., Lotfi,M., Davis,J., Singh,D. and Rizzolo,L.J. *et al.* (2015) Human Adult Retinal Pigment Epithelial Stem Cell-Derived RPE Monolayers Exhibit Key Physiological Characteristics of Native Tissue. *Investigative Ophthalmology & Visual Science*, **56**, 7085–7099.

39. Maruotti,J., Sripathi,S.R., Bharti,K., Fuller,J., Wahlin,K.J., Ranganathan,V., Sluch,V.M., Berlinicke,C.A., Davis,J. and Kim,C. *et al.* (2015) Small-molecule–directed, efficient generation of retinal pigment epithelium from human pluripotent stem cells. *Proceedings of the National Academy of Sciences*, **112**, 10950–10955.

40. Pertea,M., Kim,D., Pertea,G.M., Leek,J.T. and Salzberg,S.L. (2016) Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nature Protocols*, **11**, 1650–1667.

41. Klimanskaya,I., Hipp,J., Rezai,K.A., West,M., Atala,A. and Lanza,R. (2004) Derivation and comparative assessment of retinal pigment epithelium from human embryonic stem cells using transcriptomics. *Cloning and Stem Cells*, **6**, 217–245.

42. Zerbino,D.R., Achuthan,P., Akanni,W., Amode,M.R., Barrell,D., Bhai,J., Billis,K., Cummins,C., Gall,A. and Girón,C.G. *et al.* (2018) Ensembl 2018. *Nucleic Acids Research*, **46**, D754–D761.

43. O'Leary,N.A., Wright,M.W., Brister,J.R., Ciufo,S., Haddad,D., McVeigh,R., Rajput,B., Robbertse,B., Smith-White,B. and Ako-Adjei,D. *et al.* (2016) Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, **44**, D733–745.

44. Landry,J.-R., Mager,D.L. and Wilhelm,B.T. (2003) Complex controls: The role of alternative promoters in mammalian genomes. *Trends in Genetics*, **19**, 640–648.

45. Tian,B. and Manley,J.L. (2017) Alternative polyadenylation of mRNA precursors. *Nature Reviews Molecular Cell Biology*, **18**, 18–30.

46. WANG,Y., LIU,J., HUANG,B., XU,Y.-M., LI,J., HUANG,L.-F., LIN,J., ZHANG,J., MIN,Q.-H. and YANG,W.-M. *et al.* (2015) Mechanism of alternative splicing and its regulation. *Biomedical Reports*, **3**, 152–158.

47. Takahashi,H., Kato,S., Murata,M. and Carninci,P. (2012) CAGE- Cap Analysis Gene Expression: A protocol for the detection of promoter and transcriptional networks. *Methods in molecular biology (Clifton, N.J.)*, **786**, 181–200.

48. Beck,A.H., Weng,Z., Witten,D.M., Zhu,S., Foley,J.W., Lacroute,P., Smith,C.L., Tibshirani,R., Rijn,M. van de and Sidow,A. *et al.* (2010) 3′-End Sequencing for Expression Quantification (3SEQ) from Archival Tumor Samples. *PLOS ONE*, **5**, e8768.

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

50. Mellough,C.B., Bauer,R., Collin,J., Dorgau,B., Zerti,D., Dolan,D.W.P., Jones,C.M., Izuogu,O.G., Yu,M. and Hallam,D. *et al.* (2019) An integrated transcriptional analysis of the developing human retina. *Development (Cambridge, England)*, **146**.

51. Gorman,S.W., Haider,N.B., Grieshammer,U., Swiderski,R.E., Kim,E., Welch,J.W., Searby,C., Leng,S., Carmi,R. and Sheffield,V.C. *et al.* (1999) The Cloning and Developmental Expression of Unconventional Myosin IXA (MYO9A) a Gene in the Bardet–Biedl Syndrome (BBS4) Region at Chromosome 15q22–q23. *Genomics*, **59**, 150–160.

52. Braun,T.A., Mullins,R.F., Wagner,A.H., Andorf,J.L., Johnston,R.M., Bakall,B.B., Deluca,A.P., Fishman,G.A., Lam,B.L. and Weleber,R.G. *et al.* (2013) Non-exomic and synonymous variants in ABCA4 are an important cause of Stargardt disease. *Human Molecular Genetics*, **22**, 5136–5145.

53. Bauwens,M., Garanto,A., Sangermano,R., Naessens,S., Weisschuh,N., De Zaeytijd,J., Khan,M., Sadler,F., Balikova,I. and Van Cauwenbergh,C. *et al.* (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.

54. Zernant,J., Xie,Y.(., Ayuso,C., Riveiro-Alvarez,R., Lopez-Martinez,M.-A., Simonelli,F., Testa,F., Gorin,M.B., Strom,S.P. and Bertelsen,M. *et al.* (2014) Analysis of the ABCA4 genomic locus in Stargardt disease. *Human Molecular Genetics*, **23**, 6797–6806.

55. Sangermano,R., Garanto,A., Khan,M., Runhart,E.H., Bauwens,M., Bax,N.M., Born,L.I. van den, Khan,M.I., Cornelis,S.S. and Verheij,J.B.G.M. *et al.* (2019) Deep-intronic ABCA4 variants explain missing heritability in Stargardt disease and allow correction of splice defects by antisense oligonucleotides. *Genetics in Medicine*, **21**, 1751–1760.

56. Jamshidi,F., Place,E.M., Mehrotra,S., Navarro-Gomez,D., Maher,M., Branham,K.E., Valkanas,E., Cherry,T.J., Lek,M. and MacArthur,D. *et al.* (2019) Contribution of non-coding mutations to RPGRIP1-mediated inherited retinal degeneration. *Genetics in medicine : official journal of the American College of Medical Genetics*, **21**, 694–704.

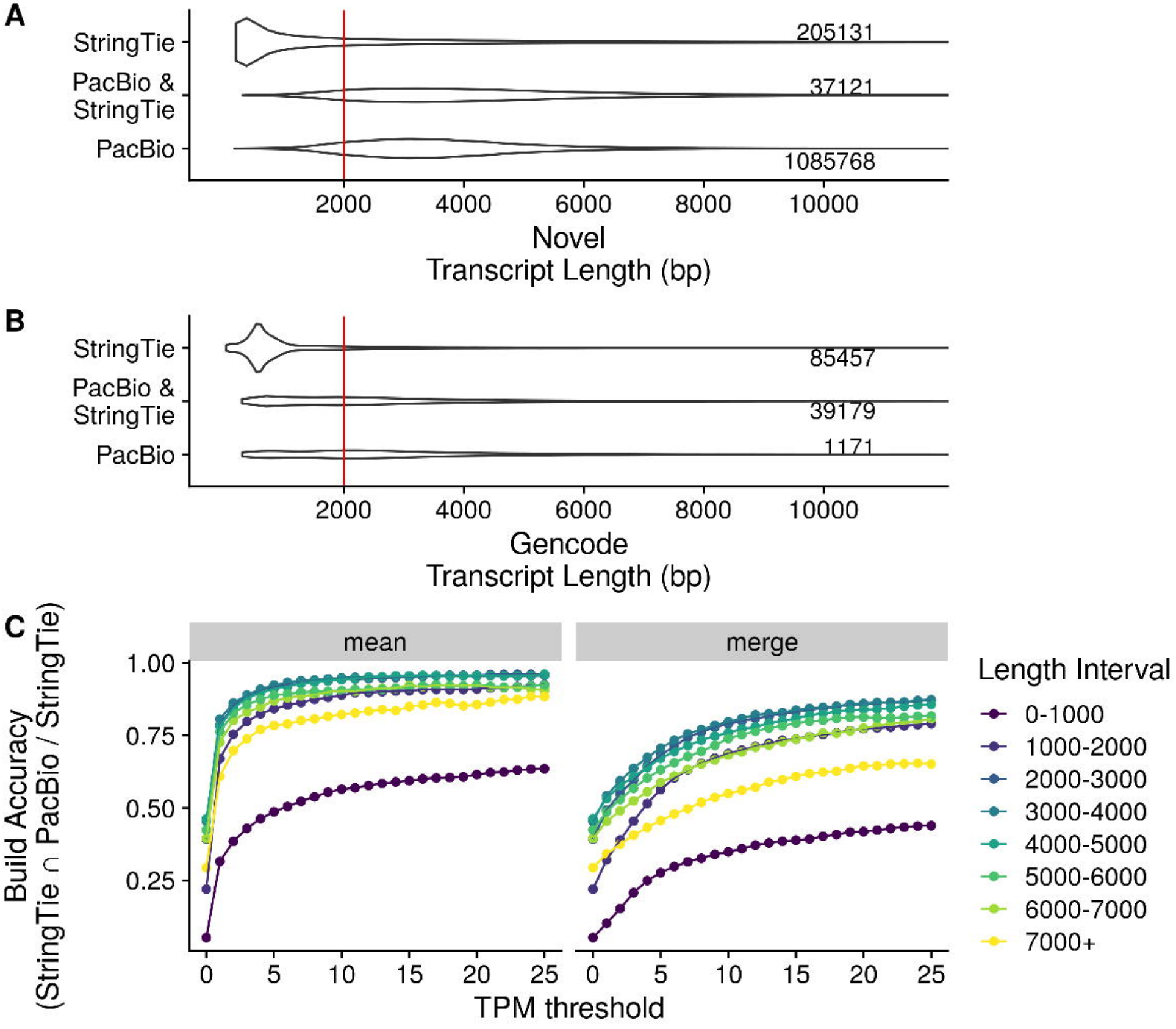
57. Mayer,A.K., Rohrschneider,K., Strom,T.M., Glöckle,N., Kohl,S., Wissinger,B. and Weisschuh,N. (2016) Homozygosity mapping and whole-genome sequencing reveals a deep intronic PROM1 mutation causing cone–rod dystrophy by pseudoexon activation. *European Journal of Human Genetics*, **24**, 459–462.

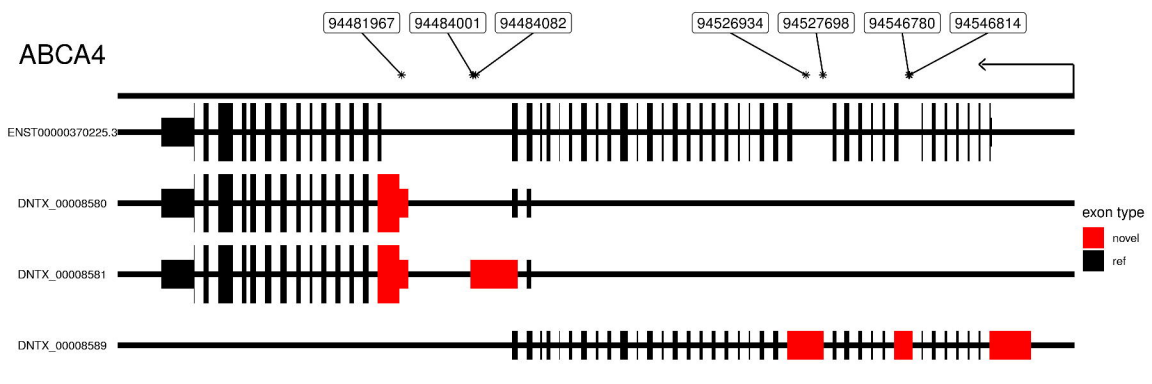
58. Geoffroy,V., Stoetzel,C., Scheidecker,S., Schaefer,E., Perrault,I., Bär,S., Kröll,A., Delbarre,M., Antin,M. and Leuvrey,A.-S. *et al.* (2018) Whole-genome sequencing in patients with ciliopathies uncovers a novel recurrent tandem duplication in IFT140. *Human Mutation*, **39**, 983–992.

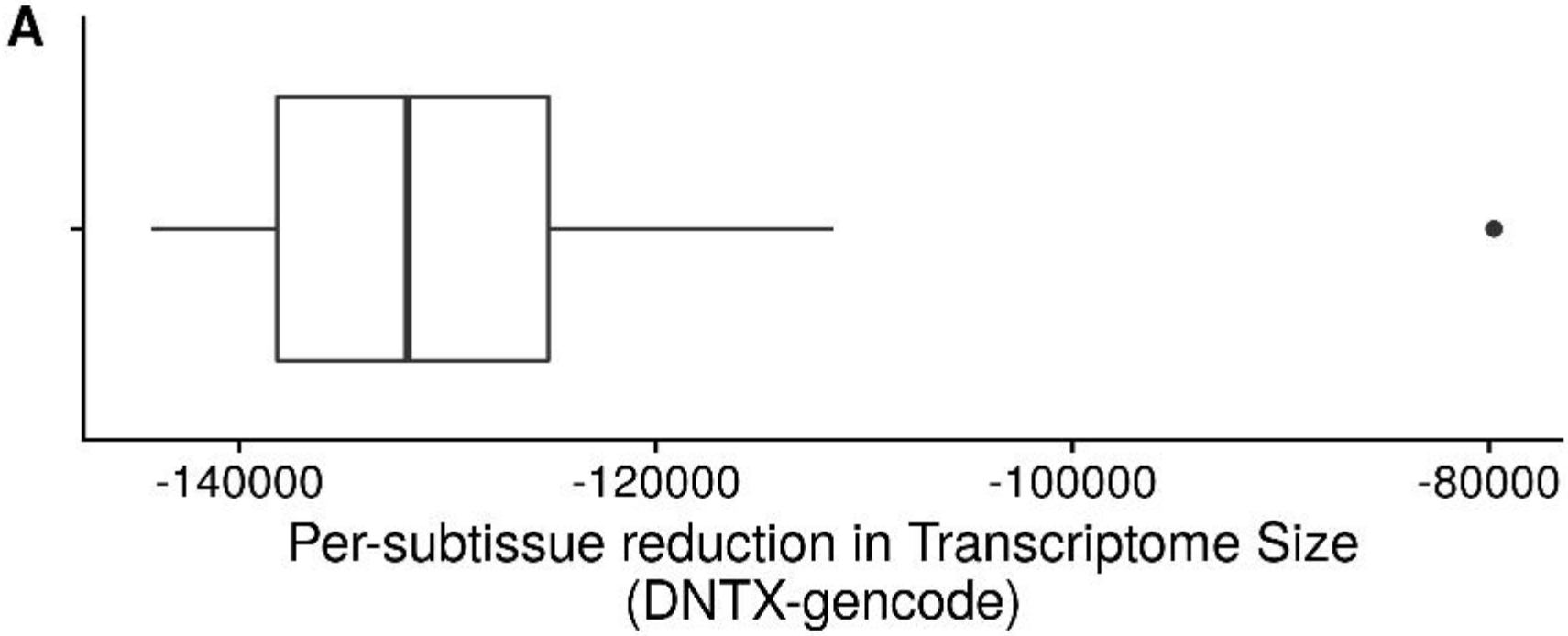
59. Lenis,T.L., Hu,J., Ng,S.Y., Jiang,Z., Sarfare,S., Lloyd,M.B., Esposito,N.J., Samuel,W., Jaworski,C. and Bok,D. *et al.* (2018) Expression of ABCA4 in the retinal pigment epithelium and its implications for Stargardt macular degeneration. *Proceedings of the National Academy of Sciences*, **115**, E11120–E11127.

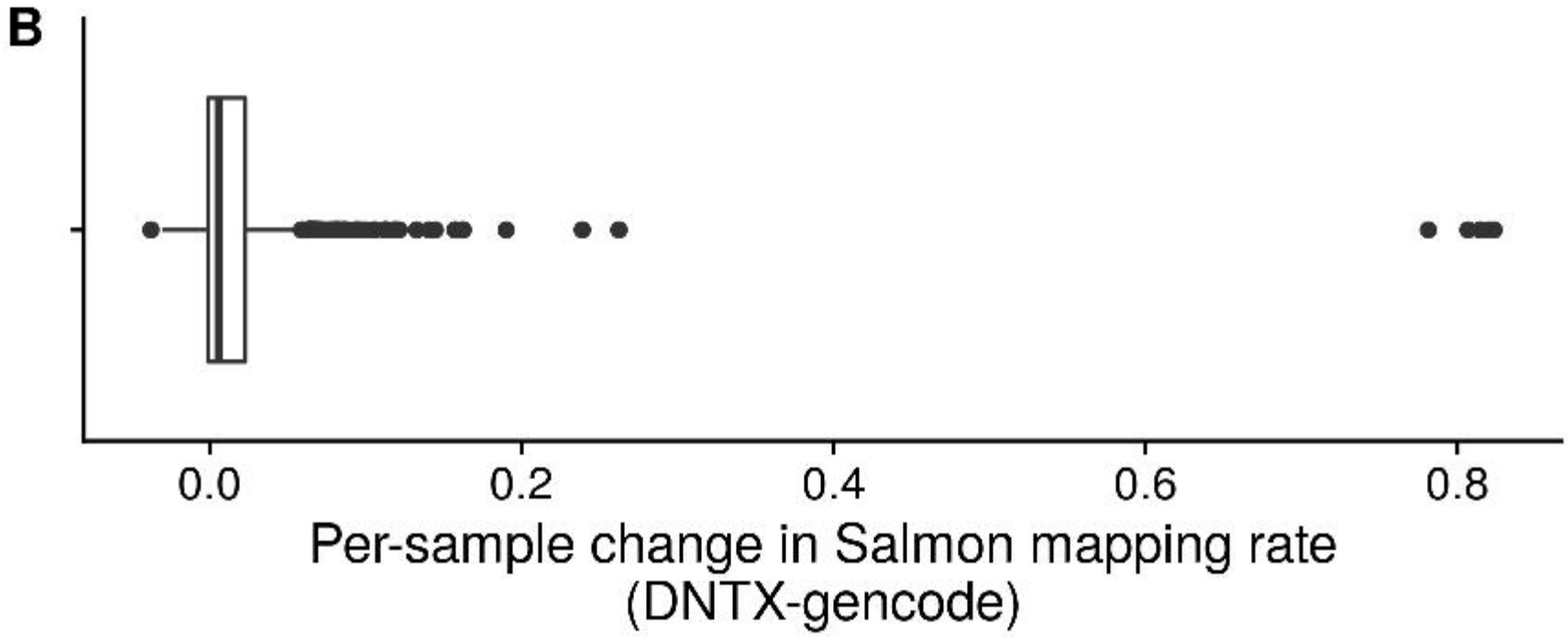
60. Reyes,A. and Huber,W. (2018) Alternative start and termination sites of transcription drive most transcript isoform differences across human tissues. *Nucleic Acids Research*, **46**, 582–592.

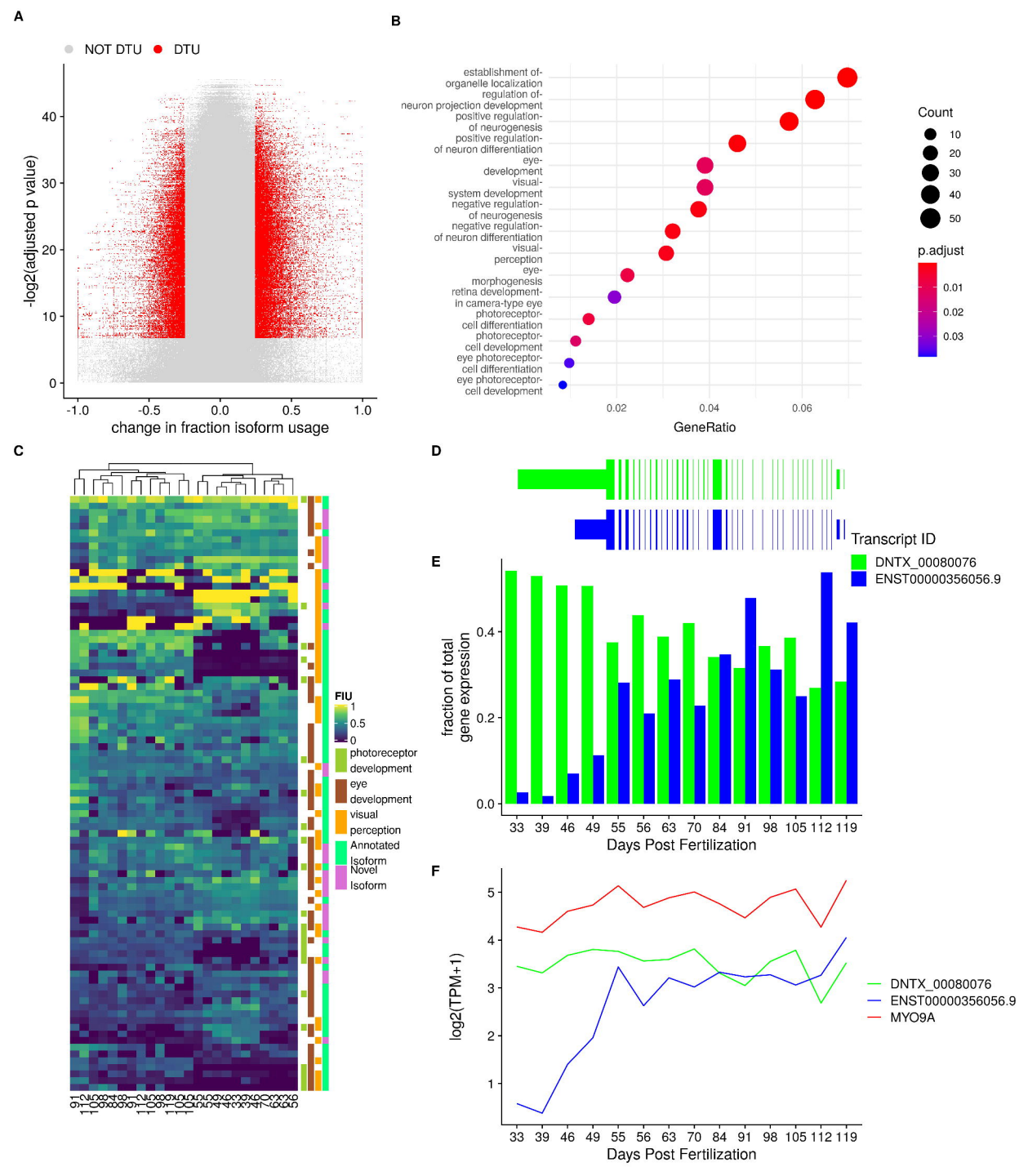
61. Yan,W., Peng,Y.-R., Zyl,T. van, Regev,A., Shekhar,K., Juric,D. and Sanes,J.R. (2020) Cell Atlas of The Human Fovea and Peripheral Retina. *Scientific Reports*, **10**, 9802.

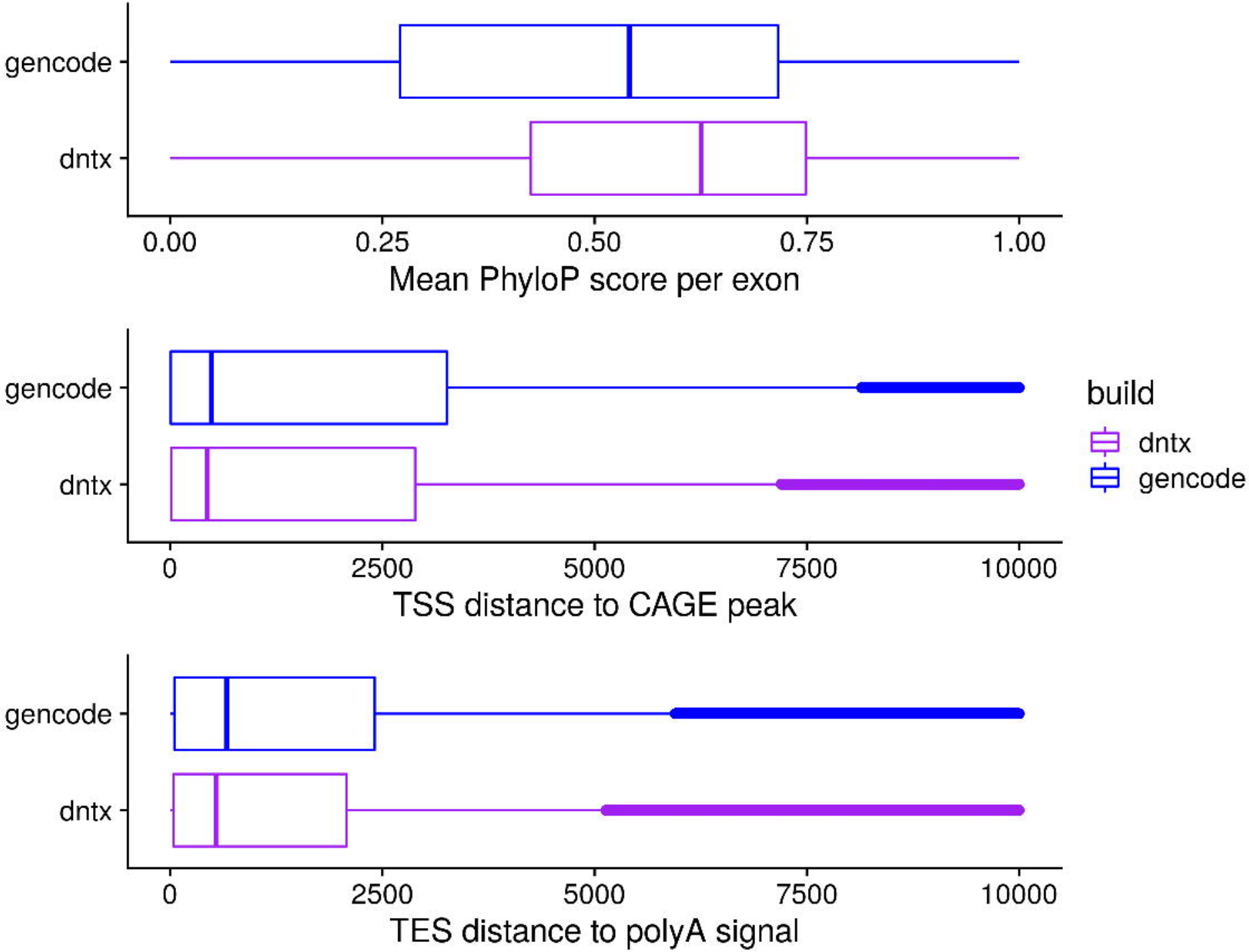

A
StringTie
PacBio &
StringTie
PacBio
205131
37121
1085768
2000
4000
6000
8000
10000
Novel
Transcript Length (bp)
B
StringTie
PacBio &
StringTie
PacBio
85457
39179
1171
2000
4000
6000
8000
10000
Gencode
Transcript Length (bp)
C
Build Accuracy
(StringTie ∩ PacBio / StringTie)
mean
merge
1.00
0.75
0.50
0.25
0
5
10
15
20
25
TPM threshold
Length Interval
0-1000
1000-2000
2000-3000
3000-4000
4000-5000
5000-6000
6000-7000
7000+

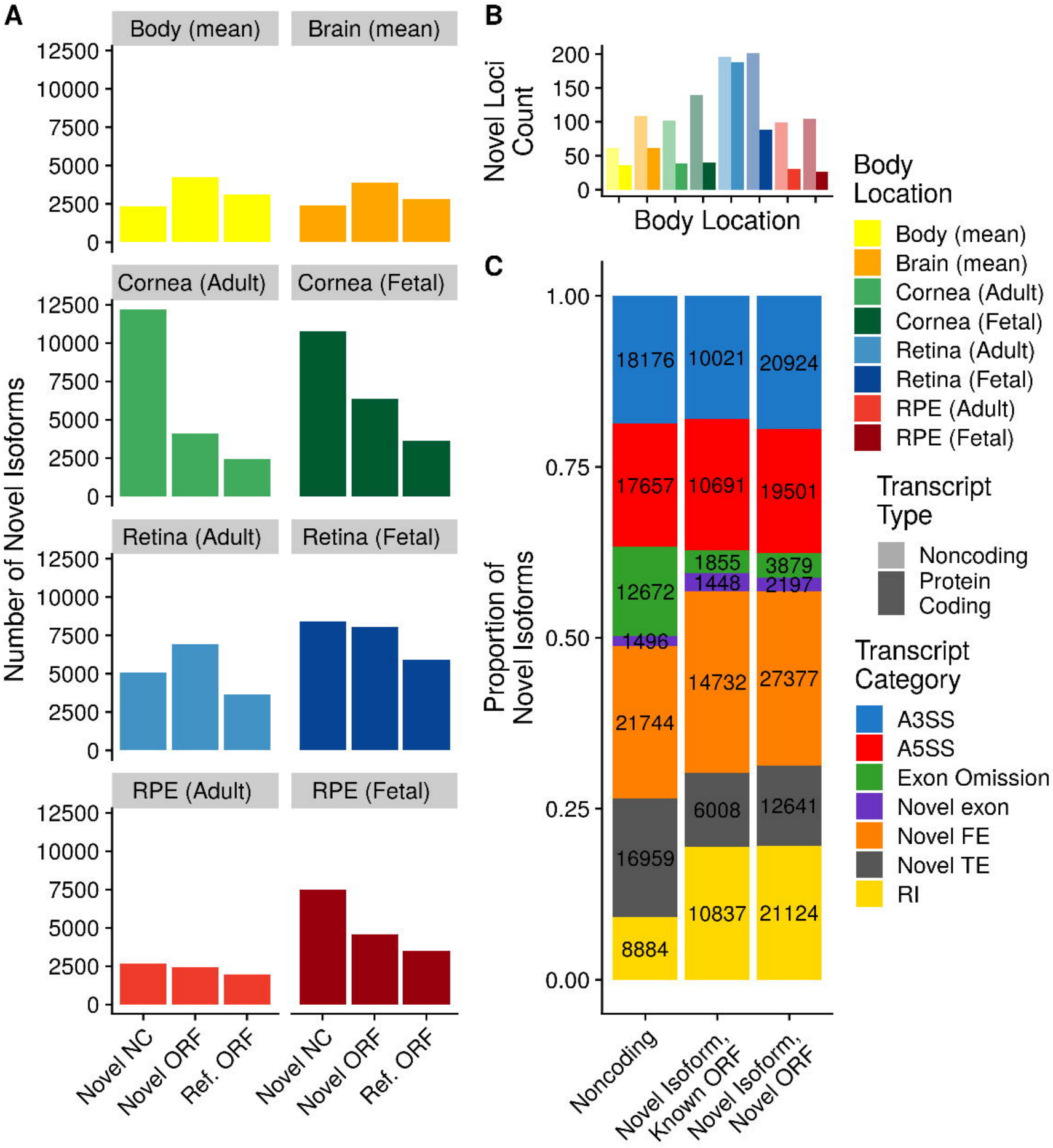

ABCA4
94481967
94484001
94484082
94526934
94527698
94546780
94546814
ENST00000370225.3
DNTX_00008580
DNTX_00008581
DNTX_00008589
exon type
novel
ref


A
-140000
-120000
-100000
-80000
Per-subtissue reduction in Transcriptome Size
(DNTX-gencode)

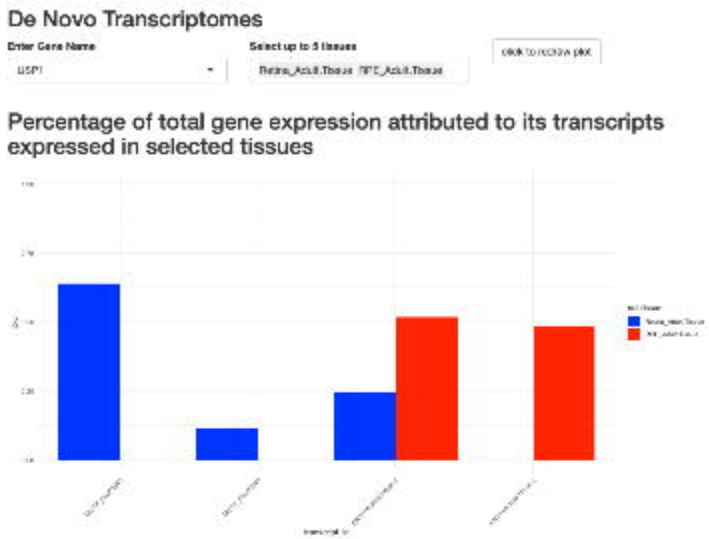

B
0.0
0.2
0.4
0.6
0.8
Per-sample change in Salmon mapping rate
(DNTX-gencode)


A
NOT DTU
DTU
-log2(adjusted p value)
change in fraction isoform usage
B
establishment of organelle localization
regulation of neuron projection development
positive regulation of neurogenesis
positive regulation of neuron differentiation
eye development
visual system development
negative regulation of neurogenesis
negative regulation of neuron differentiation
visual perception
eye morphogenesis
retina development in camera-type eye
photoreceptor cell differentiation
photoreceptor cell development
eye photoreceptor cell differentiation
eye photoreceptor cell development
GeneRatio
Count
p.adjust
C
FIU
photoreceptor development
eye development
visual perception
Annotated Isoform
Novel Isoform
D
Transcript ID
DNTX_00080076
ENST00000356056.9
E
fraction of total gene expression
Days Post Fertilization
F
log2(TPM+1)
DNTX_00080076
ENST00000356056.9
MYO9A
Days Post Fertilization

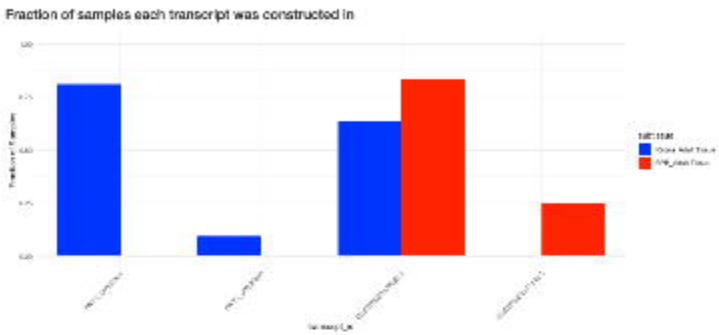

gencode
dntx
0.00
0.25
0.50
0.75
1.00
Mean PhyloP score per exon
gencode
dntx
0
2500
5000
7500
10000
TSS distance to CAGE peak
gencode
dntx
0
2500
5000
7500
10000
TES distance to polyA signal
build
dntx
gencode


A
Body (mean)
Brain (mean)
Cornea (Adult)
Cornea (Fetal)
Retina (Adult)
Retina (Fetal)
RPE (Adult)
RPE (Fetal)
Number of Novel Isoforms
12500
10000
7500
5000
2500
0
Novel NC
Novel ORF
Ref. ORF
B
Novel Loci Count
200
150
100
50
0
Body Location
C
Proportion of Novel Isoforms
1.00
0.75
0.50
0.25
0.00
18176
10021
20924
17657
10691
19501
12672
1855
1448
3879
2197
1496
21744
14732
27377
16959
6008
12641
8884
10837
21124
Noncoding
Novel Isoform, Known ORF
Novel Isoform, Novel ORF
Body Location
Body (mean)
Brain (mean)
Cornea (Adult)
Cornea (Fetal)
Retina (Adult)
Retina (Fetal)
RPE (Adult)
RPE (Fetal)
Transcript Type
Noncoding
Protein Coding
Transcript Category
A3SS
A5SS
Exon Omission
Novel exon
Novel FE
Novel TE
RI

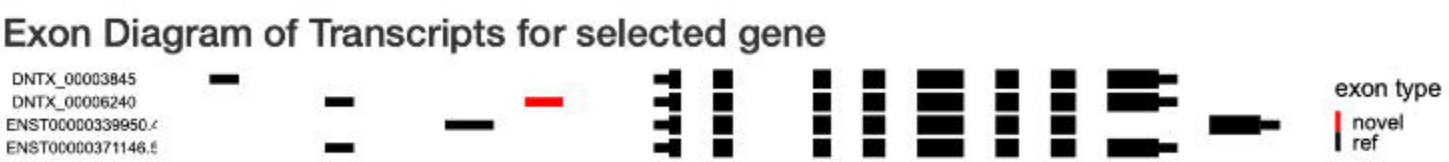
A

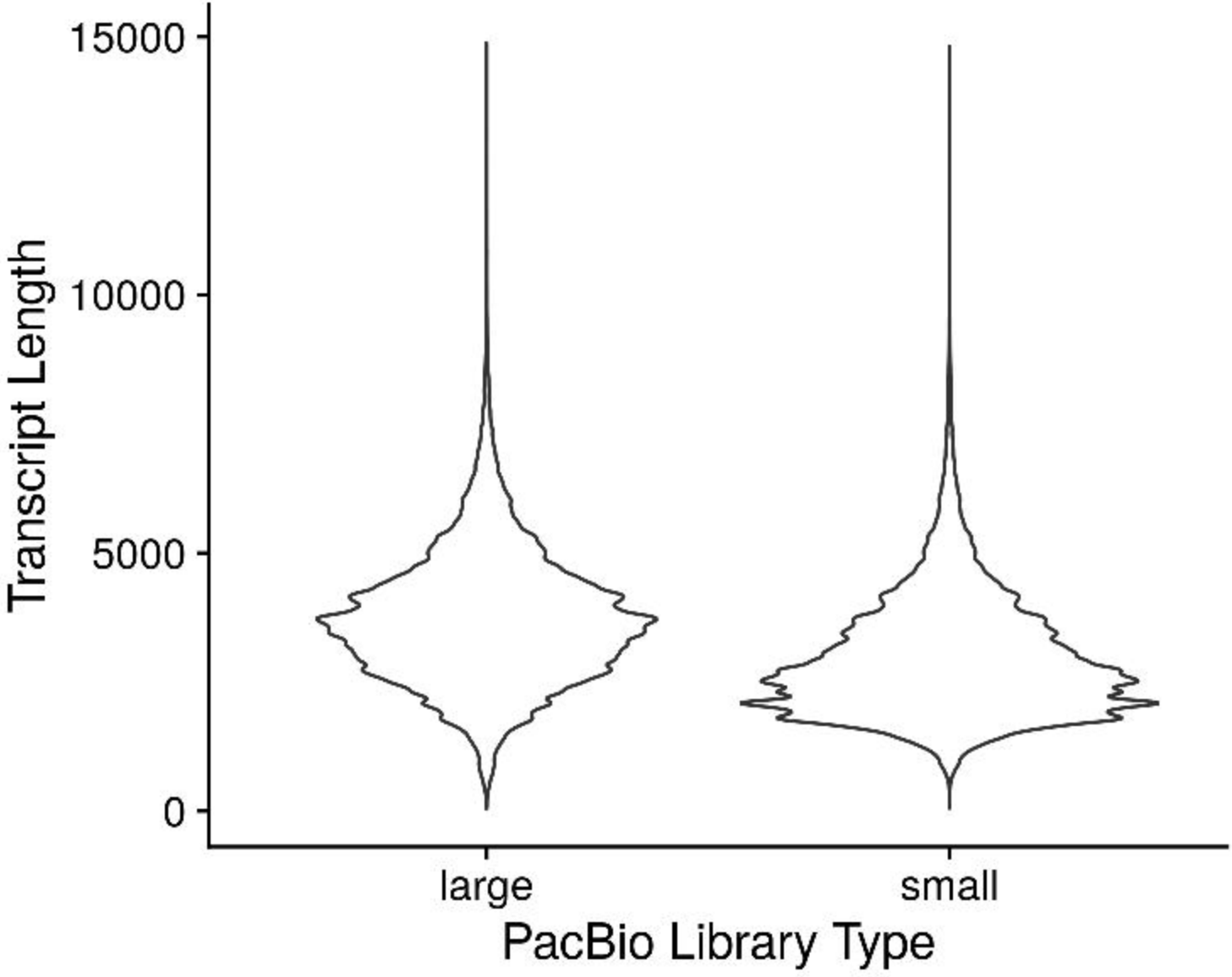

De Novo Transcriptomes
Enter Gene Name
USP1
Select up to 5 tissues
Percentage of total gene expression attributed to its transcripts expressed in selected tissues

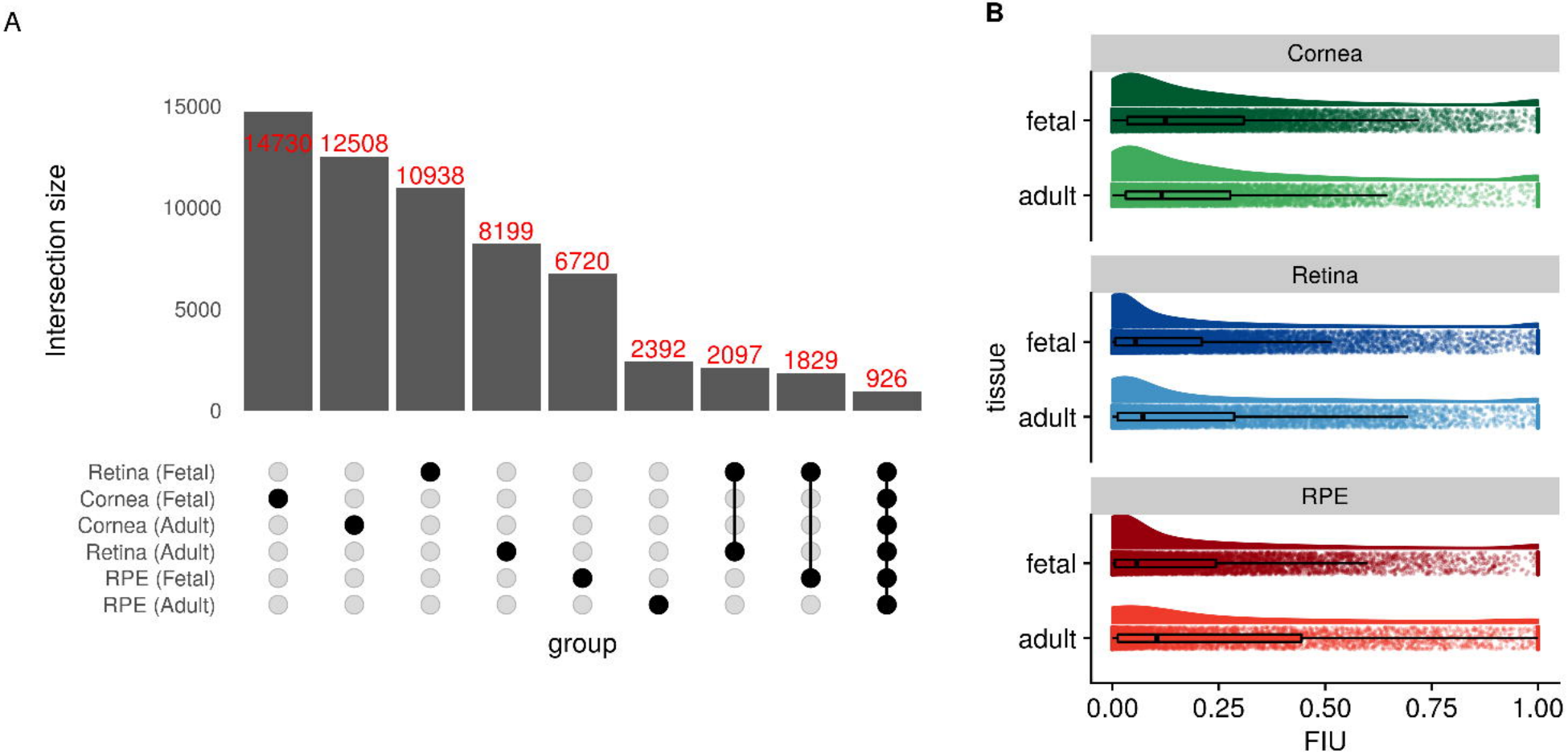
B

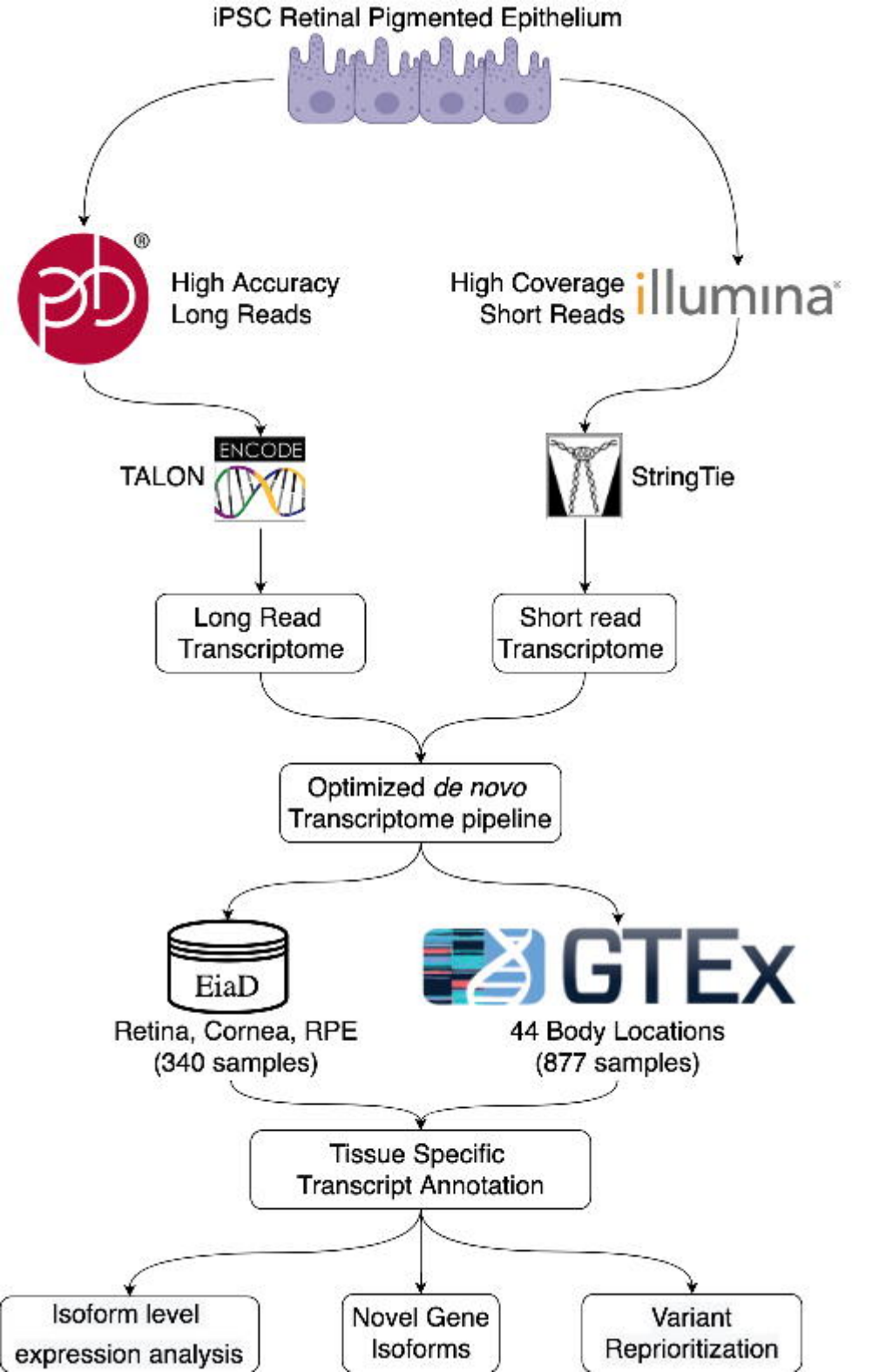

Fraction of samples each transcript was constructed in

C


Exon Diagram of Transcripts for selected gene
DNTX_00003845
DNTX_00006240
exon type
novel
ref


15000
10000
5000
0
Transcript Length
large
small
PacBio Library Type


A
Intersection size
15000
10000
5000
0
14730
12508
10938
8199
6720
2392
2097
1829
926
Retina (Fetal)
Cornea (Fetal)
Cornea (Adult)
Retina (Adult)
RPE (Fetal)
RPE (Adult)
group
B
Cornea
Retina
RPE
fetal
adult
tissue
0.00
0.25
0.50
0.75
1.00
FIU


iPSC Retinal Pigmented Epithelium
High Accuracy
Long Reads
High Coverage
Short Reads
illumina
TALON
ENCODE
StringTie
Long Read
Transcriptome
Short read
Transcriptome
Optimized de novo
Transcriptome pipeline
EiaD
GTEx
Retina, Cornea, RPE
(340 samples)
44 Body Locations
(877 samples)
Tissue Specific
Transcript Annotation
Isoform level
expression analysis
Novel Gene
Isoforms
Variant
Reprioritization