Eye in a Disk: eyeIntegration human pan-eye and body transcriptome database version 1.0

Vinay Swamy[1] and David McGaughey[1,]

June 10, 2019

Abstract

PURPOSE: To develop an accessible and reliable RNA-seq transcriptome database of healthy human eye tissues and a matching reactive web application to query gene expression in eye and body tissues. METHODS: We downloaded the raw sequnce data for 1375 RNA-seq samples across 54 tissues in the GTEx project as a non-eye reference set. We then queried several public repositories to find all healthy, non-perturbed, human eye-related tissue RNA-seq samples. The 916 eye and 1375 GTEx samples were sent into a Snakemake-based reproducible pipeline we wrote to quantify all known transcripts and genes, removes samples with poor sequence quality and mislabels, normalizes expression values across each tissue, performs 882 differential expression tests, calculates GO term enrichment, and outputs all as a single SQLite database file: the Eye in a Disk (EiaD) dataset. Furthermore, we rewrote the web application eyeIntegration (https://eyeIntegration.nei.nih.gov) to display EiaD. RESULTS: The new eyeIntegration portal provides quick visualization of human eye-related transcriptomes published to date by database version, gene/transcript, 19 eye tissues, and 54 body tissues. As a test of the value of this unified pan-eye dataset, we showed that fetal and organoid retina are highly similar at a pan-transcriptome level but display distinct differences in certain pathways and gene families like protocadherin and HOXB family members. CONCLUSION: The eyeIntegration v1.0 web app serves the pan-human eye and body transcriptome dataset, EiaD. This offers the eye community a powerful and quick means to test hypotheses on human gene and transcript expression across 54 body and 19 eye tissues.

[1] Ophthalmic Genetics and Visual Function Branch, National Eye Institute, National Institutes of Health

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

Introduction

RNA-seq is the predominant technology for deciphering transcriptomes

From anterior to posterior along the light trajectory, the human eye is composed of the cornea, lens, retina, retinal pigment epithelium (RPE), and choroid. The differentiation, maturation, and function of these tissues is mediated through spatial and temporal specific transcript and gene expression patterns, also known as the transcriptome. Today, RNA-sequencing (RNA-seq) is the predominant technology for quantifying the transcriptome. Analysis of the transcripts' expression across tissue, time, and perturbation allows researchers to decipher the genetic controls of eye development and function. To this end, a wide variety of human tissue sources have been

used to assess gene function, including primary tissue (fetal and post-mortem), differentiated stem cells, immortalized cell lines, and most recently, organoids. These tissue types have been deeply sequenced across the cornea[1–7], lens[8], retina[9–17], and RPE (choroid)[14,17–34].

The GTEx gene expression web app lacks eye-specific tissues

The Genotype-Tissue Expression (GTEx) Project has generated RNA-seq data across dozens of post-mortem human tissues from hundreds of unique donors and presents the gene and transcript level data in a comprehensive and user-friendly web app (https://gtexportal.org/); however eye tissues have not been included[35,36]. Recently Ratnapriya et al. have published a huge set of post-mortem retina, both normal and with varying degrees of age-related macular degeneration (AMD) and the GTEx project is providing the data as a download link. This data, as of June 2019, is not available in the interactive GTEx visualizations[37]. The Sequence Read Archive (SRA) and European Nucleotide Archive (ENA) are the primary repositories for all raw sequence data and two groups have quantified large portions of the RNA-seq data, including some human eye tissues, from the SRA: recount2 and ARCHS4[38,39]. To date, no curation of the sample level metadata has been done, therefore it is challenging to parse out which eye tissues are present and even more difficult to determine whether any samples were chemically or genetically perturbed. More targeted web resources that allow researchers to quickly assess gene expression in eye tissues include iSYTE, EXPRESS, and retina.Tigem.it[16,40,41]. However iSYTE only includes lens samples, EXPRESS is limited to a subset of mouse lens and retina samples, and retina.Tigem.it is retina only. We thus aimed our efforts at developing an easily accessible and reliable RNA-seq based transcriptome database of healthy human eye tissues and a matching reactive web application to query gene expression in eye and body tissues.

The eyeIntegration app interactively serves huge GTEx and human eye tissue datasets (EiaD)

The eyeIntegration web resource (https://eyeIntegration.nei.nih), originally released in 2017 at version 0.6, provides the largest set of transcriptomes from hand-curated human eye tissues along with hundreds of GTEx tissue samples[42]. This interactive web app allows for quick transcript and gene comparisons across many eye tissues and dozens of other body tissues. The dataset that the original eyeIntegration web app served was created with a series of scripts, several of which were run interactively to manually assess quality control for the samples. The interactive nature of some of the steps precluded efficient and regular data updates for the data.

To better meet the needs of the eye research community we have re-written the bioinformatic pipeline that creates the eye and body RNA-seq dataset to allow for regular, versioned updates for eyeIntegration. We call this reproducible and versioned transcriptome dataset "Eye in a Disk" (EiaD). The pipeline automates the EiaD creation, ensures full reproducibility of the results, allow for external data comparison, provides consistent sample quality control, and improves efficiency for future sample updates. The 2019 EiaD dataset contains several new tissue types, full gene product quantification, along with hundreds of new samples and improved sample labeling. The eyeIntegration web app has also been re-written to provide many new features, including versioned EiaD datasets, custom URL shortcut creation, new visualizations, improved data table searching, easy download of core datasets, and local install of the entire interactive resource with three commands. Additionally, we are prototyping new tools to display single cell RNA-seq (scRNA-seq) data to provide researchers access to cell type specific information about gene expression across murine retinal development.

The EiaD dataset can be used to identify potential avenues to improve retina organoid maturation

Retina organoids are an increasingly popular means to model human retina development. We used our pan-study EiaD dataset to show that, at a pan-transcriptome level, organoids are highly similar to early fetal retina tissue. We also show that important temporal gene expression patterns in the fetal retina tissue are recapitulated in the organoids. As the organoid differentiation methods do not yet produce fully mature retina, we focused on identifying differentially expressed processes between organoid retina and embryonic retina and detected, for example, identifying protocadherin and HOXB family gene expression differences which suggest targetable pathways to improve and benchmark organoid differentiation methods.

Methods

Identification of potential eye samples

We exhaustively searched the SRA with the SRAdb R package for eye related tissues using the query 'cornea|retina|RPE|macula|fovea|choroid|sclera|iris|lens|eye' across all columns and rows in the 'SRA' table[43,44]. As the SRAdb is being deprecated we also ran searches on the SRA and Gene Expression Omnibus (GEO) web pages with as follows: (("Homo sapiens"[orgn:__txid9606]) AND (transcriptomic[Source]) AND

("2019/01/01"[Publication Date] : "3000"[Publication Date]) AND (retina[Text Word] OR RPE[Text Word] OR macula[Text Word] OR fovea[Text Word] OR choroid[Text Word] OR sclera[Text Word] OR iris[Text Word] OR lens[Text Word] OR cornea[Text Word] OR 'trabecular meshwork'[Text Word] OR 'canals of schlemm'[Text Word] OR 'cillary body'[Text Word] OR 'optic nerve'[Text Word] OR 'laminar cribosa'[Text Word] OR retina[Title] OR RPE[Title] OR macula[Title] OR fovea[Title] OR choroid[Title] OR sclera[Title] OR iris[Title] OR lens[Title] OR cornea[Title] OR 'trabecular meshwork'[Title] OR 'canals of schlemm'[Title] OR 'cillary body'[Title] OR 'optic nerve'[Title] OR 'laminar cribosa'[Title]))". We hand selected relevant studies and selected healthy, control or unmodified samples spanning primary adult tissue, primary fetal tissue, induced pluripotent stem cell (iPSC)-derived tissue, stem cell derived organoids, and immortalized cell lines. In order to compare gene expression in the eye against expression in other body tissues, we obtained samples from 54 different body tissues from the GTEx project. Using SRA metadata from each study we extracted sample and run accessions, library type, tissue of origin, and sub-tissue of origin. Any of the preceding information missing from the SRA metadata was added by hand, when available. Stem cell-derived tissues and cell lines are marked as sub-tissues of the tissue they model.

Raw data download and quantification

We downloaded the relevant SRA files for each sample directly from the NCBI ftp server using the file transfer software Aspera. SRA files were converted to FASTQ format using the tool fastq-dump from the SRAtoolkit software package[43]. Samples only available in the BAM format were converted to FASTQ format using SAMTools[45]. Sample transcriptomes were quantified using the alignment free quantification software Salmon, using transcriptomic index built from gencode v28 protein coding transcript sequences using the transcriptomic aligner Salmon[46,47]. Using the resulting expression quantification, we identified lowly or unused transcripts within the gencode annotation, and removed transcripts that accounted for 5% or less of the total expression for its parent gene as per Sonneson et al[48]. Samples were re-quantified against a transcriptomic index built on the filtered transcript sequences. The Salmon count values were quantified as (transcript) length scaled Transcripts Per Million (TPM) to the transcript and gene level using tximport[49].

Quality control

We first removed samples with a Salmon calculated mapping rate less than 40%. This value was selected as being the far left tail of the distribution of mapping rates across samples (Supplemental Figure 2). We removed

lowly expressed genes by calculating the median expression across all samples for each gene and kept genes that had a median count >200 across all samples. To reduce the noise from experimental variability between each study, we normalized samples by sequence library size using the calcNormFactors function from the edgeR R package, and then quantile smoothed expression data using the R package qsmooth at the tissue level[50,51]. In a change from our previous eyeIntegration work[42], we now correct our counts for mapping rate and tissue type with the limma batchEffects function[52]. The transformed values are used for the box plot and t-SNE visualizations.

To identify outliers we followed an approach similar to a method in Wright et al[53]. Briefly, we first selected the 3000 genes with the highest variance across all samples and then for each sub-tissue type T and each sample i in T, we first calculated r_i, the average correlation between i and all other samples in T. Next, we calculated D_i, where $D_i = \frac{(r_i - r)}{median(r_i - r)}$ and r is the grand mean of all r_i for i in T. We removed samples with $D_i <$ -17.5; we determined this threshold by generating a tSNE plot of our samples, and visually identifying outliers in adult retina tissue. The $(max(D_i))$ amongst these outliers was -17.58 and from this we chose -17.5 as our outlier threshold.

To calculate pearson correlation (R^2) between GTEx-calculated TPM gene values and our GTEx TPM gene values, we downloaded "GTEx_Analysis_2016-01-15_v7_RNASeQCv1.1.8_gene_tpm.gct.gz" and matched against our GTEx TPM values, running the pearson correlation with $log2(TPM + 0.01)$ values as per Zhang et al. with the cor function in R[54].

Differential Gene Expression Analysis and GO term enrichment

We used the non-transformed length scaled TPM values to determine differential gene expression between different sub-tissue types. First, we generated a synthetic body set to serve as single representative sub-tissue type for pan-body gene expression by randomly sampling GTEx tissues. We used the voom function from the limma R package to convert gene expression to precision weights, and then performed pairwise differential expression tests for all combinations of eye sub-tissues (using mapping rate as a covariate), the synthetic body tissue, and human body tissues using an empirical Bayes test[52,55]. We extracted significant genes (FDR $p < 0.01$) for all 882 comparisons and used these to calculate GO enrichment. The significant gene list for each eye sub-tissue was split into upregulated and down regulated sets and each set was tested for enrichment using the enrichGO function from the clusterProfiler R package (q-value < 0.01)[56].

eyeIntegration web app and R package

The data generated in the above steps is consolidated into a SQLite database, with the original dataset for eyeIntegration and the new 2019 EiaD dataset each getting a separate database file. The code that creates the eyeIntegration web app is written in Shiny and R and has been wrapped into an R package (https://github.com/davemcg/eyeIntegration_app/) that can be deployed on a local computer or a web server (https://eyeIntegration.nei.nih.gov). The app can be deployed on a local computer with 50GB of free disk space by running three commands in R: "devtools::install_github('davemcg/eyeIntegration_app')", "eyeIntegrationApp::get_eyeIntegration_datasets()", and "eyeIntegrationApp::run_eyeIntegration()".

Snakemake reproducible pipeline

While the sample search and metadata parsing in a semi-curated process, the processing from the raw data to the creation of the SQLite EiaD database underlying eyeIntegration is wrapped in a Snakemake pipeline, which ensures full reproducibility of the results[57]. We make the code for the pipeline available at https://github.com/davemcg/EiaD_build.

scRNA-seq processing

The eyeIntegration site, as of June 2019, hosts two large scRNA-seq datasets from Macosko et al. and Clark et al[58,59]. We use the processed gene count data directly from each group, as well as their cluster assignments which specify what cell type each individual cell is. The count data is mean averaged to the cell type, age, and gene level for the single cell expression section of eyeIntegration. We also display t-SNE and UMAP-based two-dimensional visualizations of the Macosko and Clark data, respectively, in the web app. For detail so the t-SNE processing we did on the Macosko dataset, see the methods of Bryan et al[42].

Power Calculation

We use the ssizeRNA R package to calculate power (p) across samples (n) at an FDR of 0.05[60]. Important parameters for ssize RNA include the variability (dispersions for the samples and genes), which were calculated directly from our EiaD length scaled TPM values by the edgeR packages estimateCommonDisp and estimateTagwiseDisp. The code to calculate the power is given as 'power_calc.R'.

Manuscript as code and reproducibility

This manuscript's figures, tables, and most numbers, are all created and laid out in a R markdown document that interweaves code and text. The knitr and pandoc program is used to lay out the figures and tables and output a docx file. The code that generates this manuscript can be found at https://github.com/davemcg/eyeIntegration_v1_app_manuscript.

The relevant code-bases (https://github.com/davemcg/eyeIntegration_v1_app_manuscript, https://github.com/davemcg/EiaD_build) and the EiaD dataset itself has been deposited into Zenodo with accession 10.5281/zenodo.3238677 to ensure the data can be accessed in the future, even should eyeIntegration and GitHub become inaccessible in the future.

Results

EiaD 2019 contains 24 new human eye RNA-seq studies, 448 new Retina AMD samples, 207 new eye samples, and 16 total eye sub-tissue types

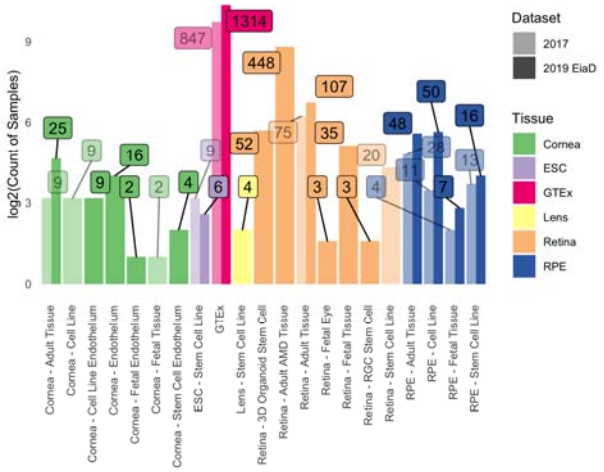


Figure 1: Substantial increase in eye tissue count and type from 2017 (180, lighter color) to 2019 (835, darker color) EiaD. We also improved the metadata labelling, the cornea samples (green) now delineates endothelial and epithelial tissues and the retina samples (orange) distinguish retina organoid and retinal ganglion cell (RGC) from stem cells. Counts for each bar plot given in the boxes. The y-axis is a log2 transformed count of samples passing our QC filters.

Our query on May 8th 2019 to the SRA found 107 potentially relevant studies. We removed non-pertinent studies and selected healthy or unmodified tissue from each relevant study for a total, including of 46 studies, 30 of which are new to the 2019 EiaD dataset. The 2019 EiaD dataset contains 835 human eye tissue samples and also

includes 1314 GTEx samples across 54 tissues for easy comparison (Table 1, Supplemental Table 1). The 2019 EiaD contains 6 undifferentiated iPSC, 56 cornea, 4 lens, 648 retina, and 121 RPE (choroid) samples; in total we have added 655 new samples to the 2019 EiaD (Figure 1). We refer to native-tissue extracted RPE as RPE (choroid) because it is not possible to remove the choroid from the RPE without culturing.

Stem cell-derived cornea, stem cell-derived lens, and fetal retina are three new types of sub-tissues that are now available in EiaD. We have also substantially improved the granularity of the cornea tissue metadata, now delineating whether the tissue is from the endothelium or epithelium (Figure 1); previously these had been grouped together as adult tissue. Another substantial addition to the 2019 EiaD are non-protein coding genes; while protein-coding is the most common gene and transcript typse, there are dozens of different non-coding classes. The 2017 version of eyeIntegration only quantified protein coding genes and transcripts. We now quantify expression across 41 gene and 45 transcript types, including protein coding, retained intron, lincRNA, antisense, and pseuodogenes (Supplemental Table 2).

We have also added the large retina AMD post-morten Ratnapriya et al. cohort to EiaD 2019[37]. This cohort contains hundreds of samples ranging from non-AMD (Minnesota Grading System (MGS) 1) to severe AMD (MGS 4). While eyeIntegration is intended to be a source for normal tissues, we have made an exception for this study, as this is a large cohort and AMD is a common disease. We found our corrections methods did not group the non-AMD Ratnapriya et al. samples with our other collected retina samples (see Retina MGS in Supplemental Figure 3). This may be related to the lower mapping rate of the Ratnapriya et al. data (see Retina MGS in Supplemental Figure 2).

467 more GTEx samples and 9 new GTEx body sub-tissue types added to 2019 EiaD

Our previous dataset for eyeIntegration version 0.6 held about 20 samples per GTEx tissue type. We ran power calculations to assess our ability to detected >= 1 log2(Fold Change) in gene expression between two conditions to determine whether this is a sufficient number of samples (Supplemental Figure 4). Our calculations suggest, for example, that we have 83% power to detect a 1 log2(Fold Change) difference in gene expression with two groups of twenty samples. To increase our power to make significant eye to body comparisons, we added about 10 more samples per GTEx tissue types (which at 30 samples, would give about 90% power). We also took this opportunity to add bladder, bone marrow, cervix uteri, fallopian tube, ovary, prostrate, testis, uterus, and vagina GTEx tissue samples (Supplemental Table 1).

Rigorous quality control and reproducible workflow system ensures high quality transcriptomes that consistently cluster together by tissue type

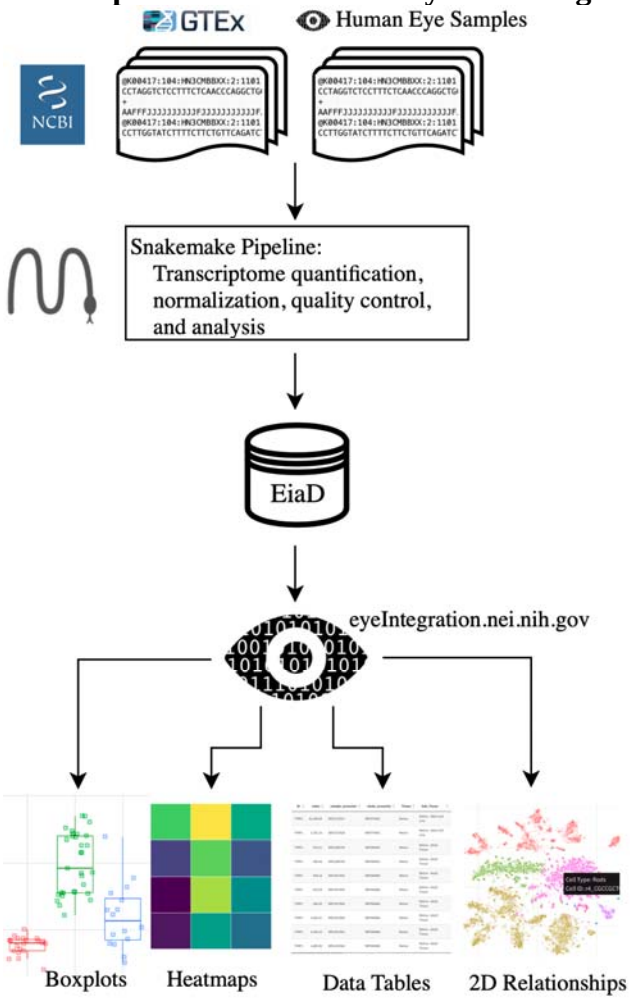


Figure 2: Raw RNA-seq data from the SRA is run through our pipeline to create the EiaD, which is used by eyeIntegration app to serve interactive gene expression visualizations across 73 tissues

We built an automated pipeline for processing and analyzing all data for the web app using the program Snakemake, a python-based workflow management system that allows for efficient parallel execution of the analysis, facilitates reproduction by others, and simplifies long term maintenance of the EiaD data (Figure 2, Supplemental Figure 5)[57]. To create a high quality final dataset across the 2291 initial samples (Supplemental Table 3) and 67,315,523,736 reads we developed a rigorous quality control procedure as part of our analysis, considering a sample's read mapping rate and median count level as well as behavior relative to samples of the same sub-tissue type (see Methods). To identify tube mislabeling or sample extraction issues, we used sample-level gene correlation metrics (see methods) to identify variability within samples of the same sub-tissue and ensure overall consistency in data processing (Figure 3). After these steps 81 eye samples and 61 GTEx samples were removed.

To ensure there are no substantial differences in quantification of gene TPM values, we calculated the R^2 between GTEX and EiaD generated TPM values for our shared GTEx samples (see methods); we computed an R^2 of 0.89. Zhang et al. report that RNA-seq quantifications done between alignment-free methods (used in EiaD) and alignment-based methods (used by GTEx) get a R^2 ranging from 0.89 to 0.93[54]. As Zhang et al. compared quantification methods with identical gene references (we use Gencode GRCh38 gene models and GTEx uses hg19) and did not scale TPM score differently, our result falls in line with expectations.

After our quality control and processing workflow, we found that samples of the same tissue type and origin cluster well together (Figure 3 and Supplemental Table 4). For example, in the retina group, primary adult tissue clusters tightly and distinctly from other cell types, and retinal organoids and fetal retina samples cluster together. Our ability to uniformly cluster data by known biological source independent of study origin demonstrates that our workflow can effectively account for technical variation between studies.

While t-SNE is a powerful algorithm for grouping samples, it is not consistent for determining the relationships between clusters[61]; PCA is more useful in this regard. We ran a PCA dimensionality reduction (Supplemental Figure 6) on all samples, finding that the eye tissues still generally group together and apart from all other human body tissues. Adult retina is most similar to the brain tissue. RPE and cornea are most similar to blood, bone marrow, and skin.

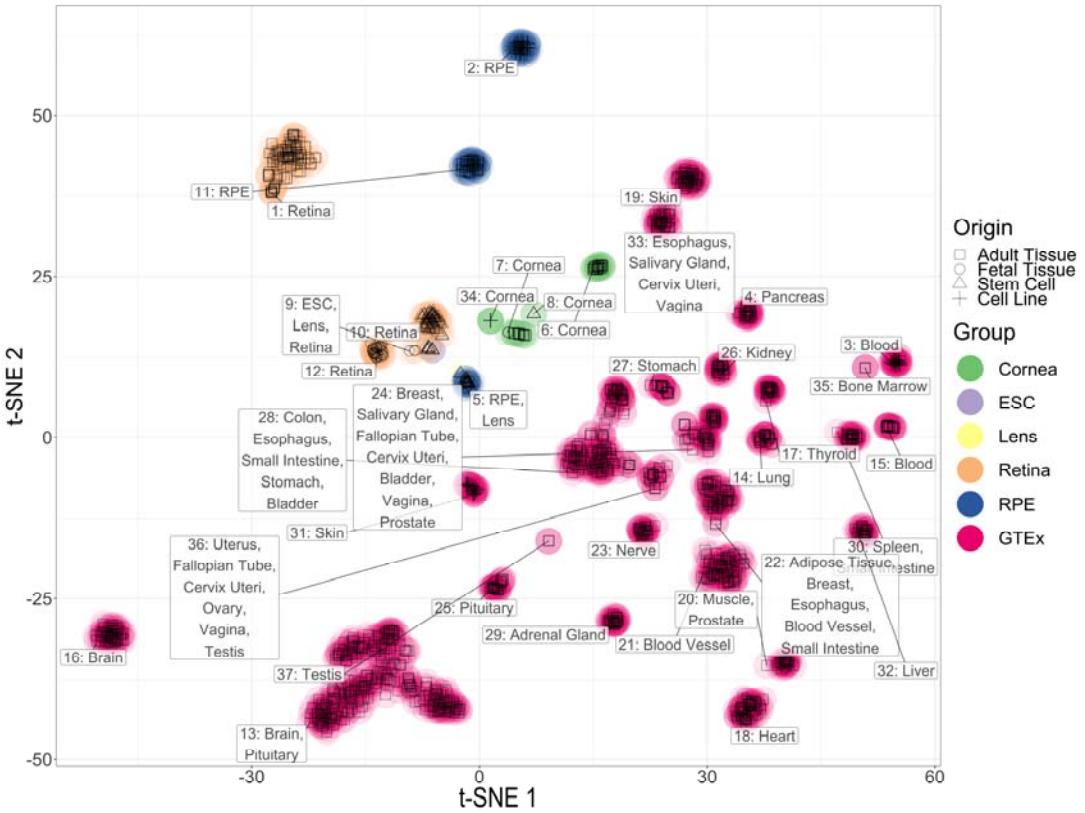


Figure 3: t-SNE two-dimensional transcriptome profiles by sample demonstrate effective quality control and transcriptome processing. Colors match different tissue types and shapes of points define the origin of the tissues.

Tissue	Pre QC Count	Count	Sub-Tissue Types (Count)
Cornea	62	56	Adult Tissue (25), Cell Line Endothelium (9), Endothelium (16), Fetal Endothelium (2), Stem Cell Endothelium (4)
ESC	12	6	Stem Cell Line (6)
Eye lid	4	0	
Lens	9	4	Stem Cell Line (4)
Retina	681	648	3D Organoid Stem Cell (52), Adult Tissue (107), Adult Tissue AMD MGS 2 (172), Adult Tissue AMD MGS 3 (112), Adult Tissue AMD MGS 4 (61), Adult Tissue MGS 1 (103), Fetal Eye (3), Fetal Tissue (35), RGC Stem Cell (3)
Retinal Endothelium	4	0	
RPE	144	121	Adult Tissue (48), Cell Line (50), Fetal Tissue (7), Stem Cell Line (16)

Table 1: EiaD contains a large set of diverse eye tissues, including embryonic stem cells (ESC). Eye lid and retina endothelium samples were included, but all failed to pass our QC filters.

The eyeIntegration web app provides interactive visual portal to all data

The EiaD 2019 dataset is used directly by the eyeIntegration web app (https://eyeIntegration.nei.nih.gov). The web app was designed to provide a simple interface that has the same general concept – select specific genes and tissue and view relevant information. The web-app is divided into four general categories: expression, two-dimension sample relationships, gene networks, and data tables.

Custom gene and tissue expression boxplots

The 'Expression' tab of the webpage provides a wealth of information about both gene- and transcript-level expression for eye and body tissues, giving the user the ability to compare the expression of different genes within a single tissue, as well as the expression of genes across multiple tissues (Figure 4A). The user first selects either the 2017 or 2019 gene or transcript EiaD dataset, then Hugo Gene Nomeclature Committee (HGNC) genes names (or ENSEMBL transcripts), then tissues. A boxplot is then generated after hitting the "Re(Draw) Plot" button with overlaid individual data points. On mouse-over, the metadata for the individual sample is displayed. A tabular report is generated based on selected genes and tissues: a table with links to Ensembl, GeneCard, and OMIM for each gene for quick referencing, and a table containing expression levels for each selected gene in each selected tissue. The tables can be arranged or sorted to the user's preference and can be easily downloaded for local use.

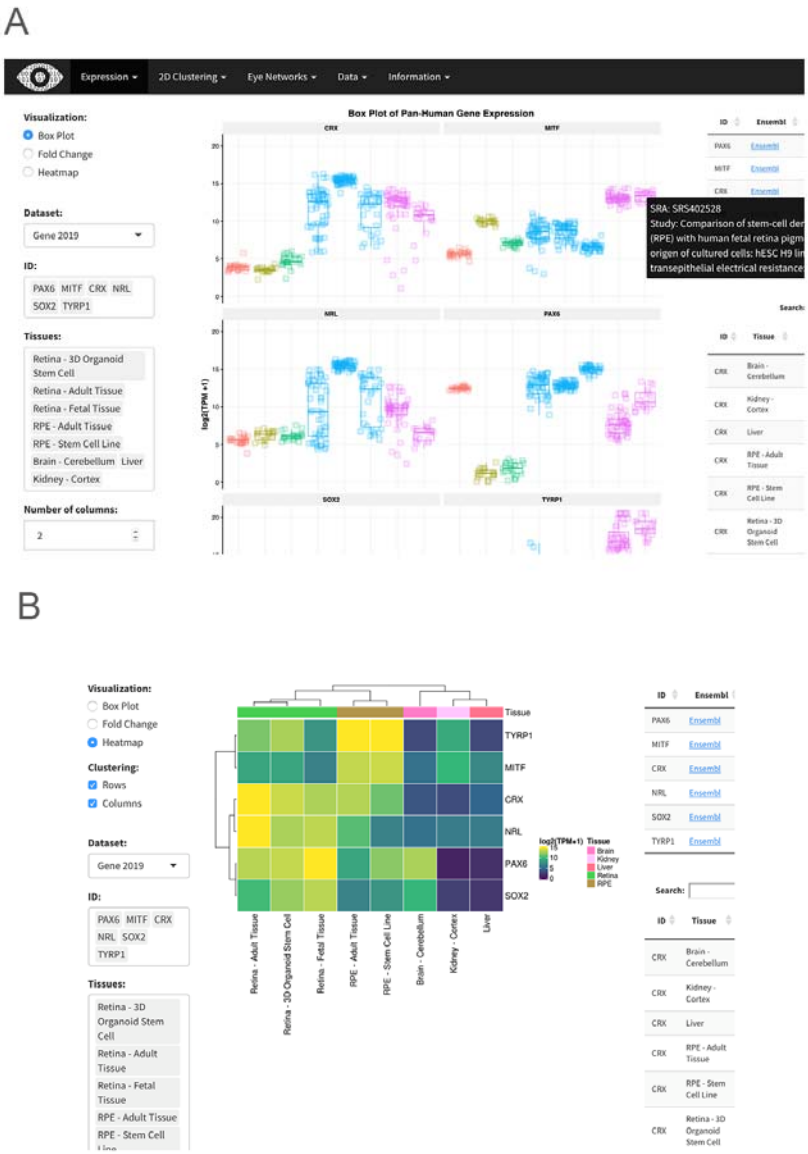


Figure 4: Screenshots from eyeIntegration web app. A. Pan-tissue gene expression box plots with accompanying data tables. The data tables display the rank (lower is more highly expressed) of each gene in each sub tissue, the decile of the rank (10 is the highest decile of expression), and the gene's mean log2(TPM + 1) score for each sub tissue. B. Heatmap visualization.

Heatmap built by the R package ComplexHeatmaps based on expression can be drawn for selected genes and tissues and gene expression can be compared across many genes and tissues (Figure 4B)[62]. Finally, a session can be saved or shared by building a custom link for the session with the "Build URL Shortcut" button.

Differential expression and gene ontology enrichment tests allow quick comparison of gene differences between groups

We performed multiple differential comparisons at the sub-tissue level within all eye tissues and against a pan-body synthetic set comprised of a stratified sample of all tissues present in our subset of the GTEx dataset, allowing quick identification of eye specific genes across 882 different comparisons. We have expanded the differential tests in the 2019 EiaD by adding the GTEx tissues as direct comparisons to our eye sub-tissues. The user can view the results selecting 'Differential' under the 'Expression' tab (Supplemental Figure 7D). As with 'Expression', the user can select which version of the web app to draw data from as well as select for gene- or transcript-level comparisons. The user additionally has the option to select different gene classes to examine, e.g. protein coding, lincRNA.

The results of differential expression are presented in a tabular format showing $\log_2$ fold change, average expression, and p-values. Depending on the comparison, there are 1 to 33380 differentially expressed genes (Supplementary Files). The table can be easily searched for any given gene, viewed and ordered to the user's preference, and downloaded in CSV format. Differential expression can be visualized through fold change bar graphs with the 'Pan-tissue plots' selection under 'Expression'. Additionally, we performed GO enrichment for all differential comparisons. Enriched GO terms are presented first as a word cloud, for quick comparison of GO enrichment. We provide tables, with similar viewing options as the differential expression table, for enriched GO terms in each class of a given differential comparison.

Murine scRNA-seq enables testing of retina cell type specific expression

We incorporated scRNA-seq data from murine retina across two studies[58,59]. This allows researchers to quickly examine gene expression across individual cell types in the retina. Single cell gene expression data is visualized through a heatmap showing the expression of a gene across multiple retinal cell types and different developmental time points, from embryonic day (E)11 to postnatal day (P)14 (when available), and a table of expression values is generated containing the expression data used to draw the heatmap (Figure 4C). We also provide t-SNE/UMAP based clustering using cell type specific labeling created by the publishing authors (Figure 4D, see Methods). The plots show all cell types present at a given developmental stage, and highlights cells expressing a gene above a user-selected given level.

EiaD 2019 suggests that iPSC-derived organoids and fetal retina have closely related transcriptomes

There are, currently, two major approaches to studying developing human retina: post-mortem fetal tissue and stem-cell derived organoids. We looked at how well these approaches to studying developing retina compare at a transcriptomic level, both for tissue - organoid relationships and how well they correlate across early development.

To evaluate how the tissues and organoids compare at a transcriptome level, we looked at the same t-SNE plot from Figure 3 and focused in on the three types of retina tissue (adult, fetal, and organoid) (Figure 5A). Here we saw three distinct groupings: adult retina (1), developing fetal retina and stem cell-derived organoid (2), and undifferentiated and early differentiating stem cells (3). We identified several organoid samples in cluster 3, but these share one important difference from the rest of the organoid samples in cluster 3: they have been differentiating for less than 30 days (shape 'X'). All of the organoid retina samples in cluster 2 are older than 50 days.

To assess how similarly the fetal and organoid retina develop through time, we plotted expression of retinal progenitors, photoreceptors, and retinal ganglion markers by time in days (Figure 5B). Each row is a gene marker of either retinal progenitor, photoreceptor, or RGC. The rows are hierarchially clustered to put more similar expression patterns closer together, as denoted by the height of the dendrogram. We split the organoid tissues into three groups: Kaewkhaw et al. GFP+ and GFP- samples, and Eldred et al. samples[12,63]. The Kaewkhaw samples are flow sorted for a GFP marker (GFP+) under the control of the *CRX* promoter, an important regulator of photoreceptor development. GFP+ cells wouldbe enriched in photoreceptor populations. We saw that the retinal progenitor, photoreceptor, and RGC groups are largely clustered together, with patterns of expression consistent across the fetal retinal and organoid groups.

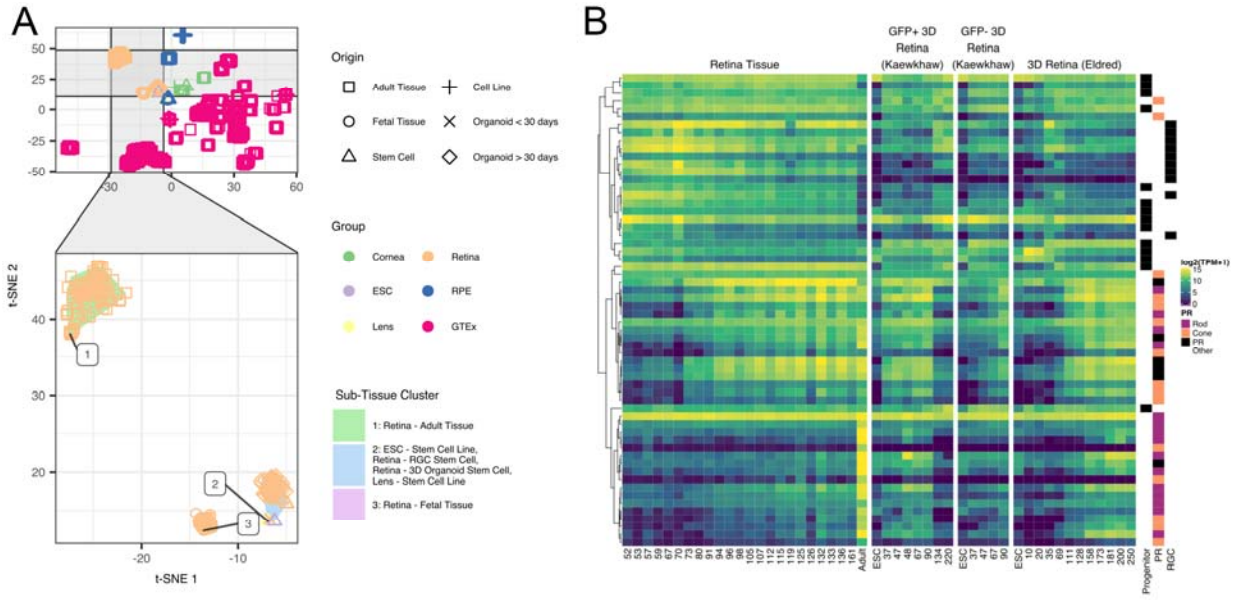


Figure 5: Organoid retina, stem cell retina, and fetal retina tissue have highly similar transcriptomes. The zoom inset (A) shows the retina samples. The "Sub-Tissue Cluster" shading shows the cluster membership of the three major groups. The shapes of the points show the different origin types - notable types include the square for adult, the 'X' for organoid under 30 days of differentiation, and the diamond for organoid over 30 days of differentiation. Major markers of retina progenitor, photoreceptors (cone and rod), and retinal ganglia cells (RGC) have similar gene expression patterns across development in retina fetal tissue and organoids.

Differential gene expression of organoid retina versus fetal tissue identifies sets of genes relating to patterning (HOXB family), cell adhesion (protocadherin family), and RGC identity (BRN3/POU4F, NEFL, GAP43, SNCG)

To identify specific changes between retinal organoid and fetal retina tissue, we performed differential gene expression and GO term enrichment analyses. The GO term enrichment identified cell adhesion (protocadherins) and patterning (HOXB family) as enriched gene sets in retinal organoids As there is some evidence suggesting that protocadherins influence RGC viability and we noticed that several RGC markers appeared to have lower expression in the organoids compared to the fetal tissue Figure 5B we looked more closely into RGC marker expression[64].

We plotted HOXB family, protocadherin family and RGC genes in a heatmap visualization, with columns as age in days of fetal or organoid retina. Rows are genes, split by the three different groups of genes and are internally clustered by how similar the expression patterns are. We observed that there are strong, consistent gene expression differences in these three groups of genes between fetal retina and the organoid samples (Supplemental Figure 8). We also plotted the differential expression values between all organoids and all fetal retina samples; all genes across all three groups are significantly differentially expressed with an FDR corrected p value < 0.01.

Limitations of the RNA-seq quantification in eyeIntegration

Salmon quantification, while highly performant and accurate, has a higher variance for lower read depth samples and shorter transcripts[54]. Extra care should be taken with comparisons with lower counts of samples (cornea, RGC) as smaller sample numbers decrease the confidence in differential expression. We do not recommend you directly compare our TPM values with your counts data as there are many important variables that will differ. Instead run our Snakemake pipeline (https://www.github.com/davemcg/EiaD_build), adding your samples. Finally, we would like to remind any users that RNA-seq methods measure mRNA levels, but the functional unit is the protein; westerns are still the gold standard with which to evaluate expression and localization.

Data accessibility

Individual data files for gene expression and sample metadata can be downloaded from the 'Data' tab on the web app. All data and code used to generate the web app can be installed from the R command line by running devtools::install_github('davidmcg/eyeIntegration_app'). The code for the EiaD data processing pipeline can be found at https://github.com/davemcg/EiaD_build.

Discussion

EiaD 2019 contains a large set of carefully curated, reproducibly processed human eye RNA-seq datasets alongside a human body tissue comparison set from the GTEx project. It is available for local install as an R package at https://www.github.com/davemcg/eyeIntegration_app and it is served via a web app, eyeIntegration at https://eyeIntegration.nei.nih.gov. The web app offers a wide range of user-driven visualizations to compare expression of genes across dozens of human body and eye tissues. Furthermore, murine scRNA-seq datasets have been incorporated, allowing for examination of retina cell type-specific gene expression. Several human and non-human primate studies have been posted in the past year on the pre-print server bioRxiv and as the raw data becomes publicly available, we will be updating this section of eyeIntegration[65–67].

If you wish to have your data added to EiaD in the future, we suggest you 1. deposit data into GEO/SRA, 2. use clear, descriptive, consistent, and detailed metadata for each sample, and 3. (optional) contact the corresponding

author. Contacting the corresponding author is only necessary if you feel your data should be included in EiaD and was deposited into the SRA before May 8th, 2019.

As human fetal tissue is difficult to obtain and thus not very amenable for chemical or genetic modification, it is crucial for organoid-based models to be developed. Our merging of these datasets and analysis at the transcriptome level (as compared to cross-analyzing using a limited number of known marker genes) indicates that these two approaches successfully recapitulate fetal retina tissue, to a first approximation, at the whole transcriptome level. However as organoids do not develop to full function, it is important to look at how gene expression differs between retinal organoid and fetal tissue so as to suggest areas for improvement.

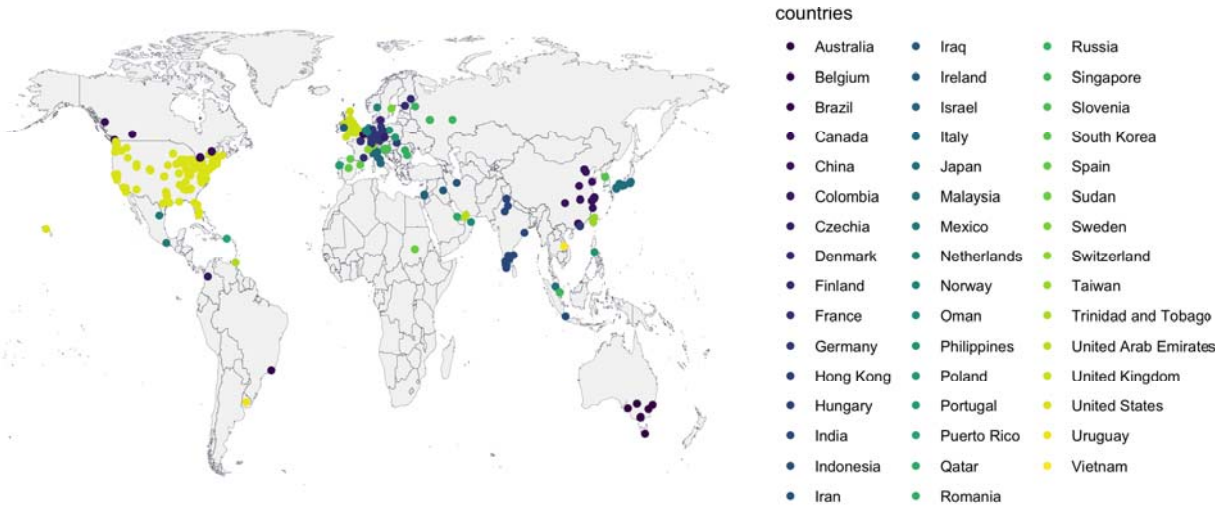
We used our large dataset to narrow in on three core processes which differ significantly and substantially between retinal organoids and fetal retina. First we showed that the HOXB family is overexpressed in the organoids. The homeobox family is well known to initiate polarity of the embryo during early development[68]. Retinoic acid is applied at about day 20 in culture to help differentiate stem cell to organoids and is also known to activate genes members of the HOXB family. The lack of HOXB expression at any age in fetal retina and the broad chromatin and gene expression changes HOXB family members can mediate suggests that HOXB activity may be unwanted for organoid maturation.

Next, we detected several protocadherins more highly expressed in the fetal tissue, relative to the organoids. Protocadherins mediate cell to cell connections and, in the developing mouse, are shown to be important for spinal internneurons and RGC survival[64,69]. We would predict that decreased protocadherin expression reduces the number and maturation of RGC. Indeed we observed that many canonical RGC markers, while present in detectable levels in the organoids, are signficantly underexpressed relative to fetal tissue. This result suggests that modifying culture conditions to promote protocadherin expression may result in higher RGC yield and survival.

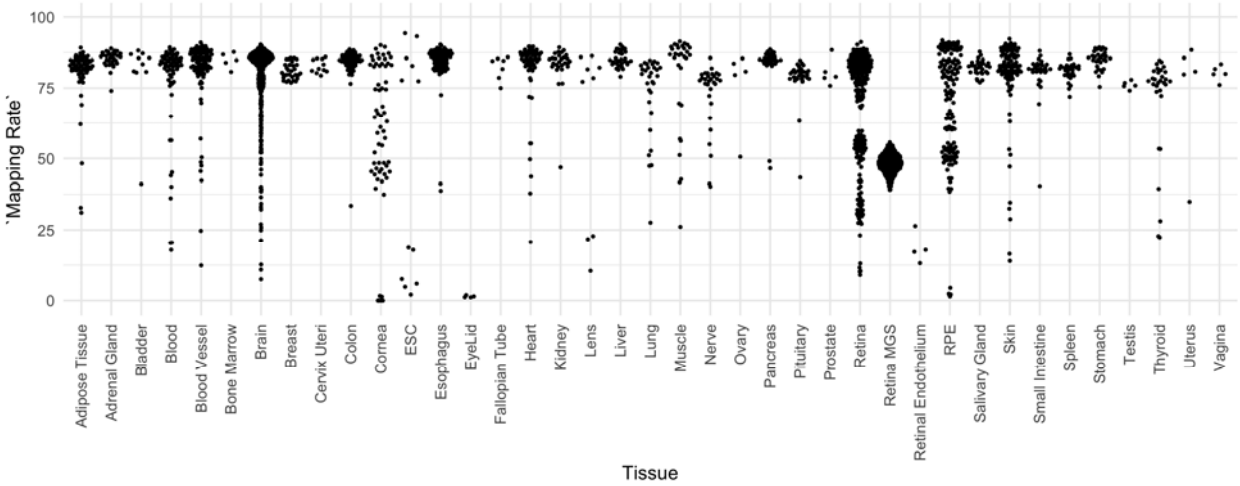
We built the Eye in a Disk dataset and the accompanying web app, eyeIntegration in the hopes that easily accessible gene expression across tissue space and time will be a useful tool for hypothesis generation and refinement in eye research. Wrapping all of the data processing steps in a Snakemake pipeline has several important advantages for the community: our code is publicly available for review, our analyses are reproducible, future sample updates can be streamlined in with less effort, and because all the processing is in modular pieces it is easier to add new analysis steps. In the future, we plan on regularly adding new samples to EiaD, offering *de novo* eye

tissue transcriptomes, expanding the single cell RNA-seq expression tooling, adding non-human eye samples, and epigenetic datasets.

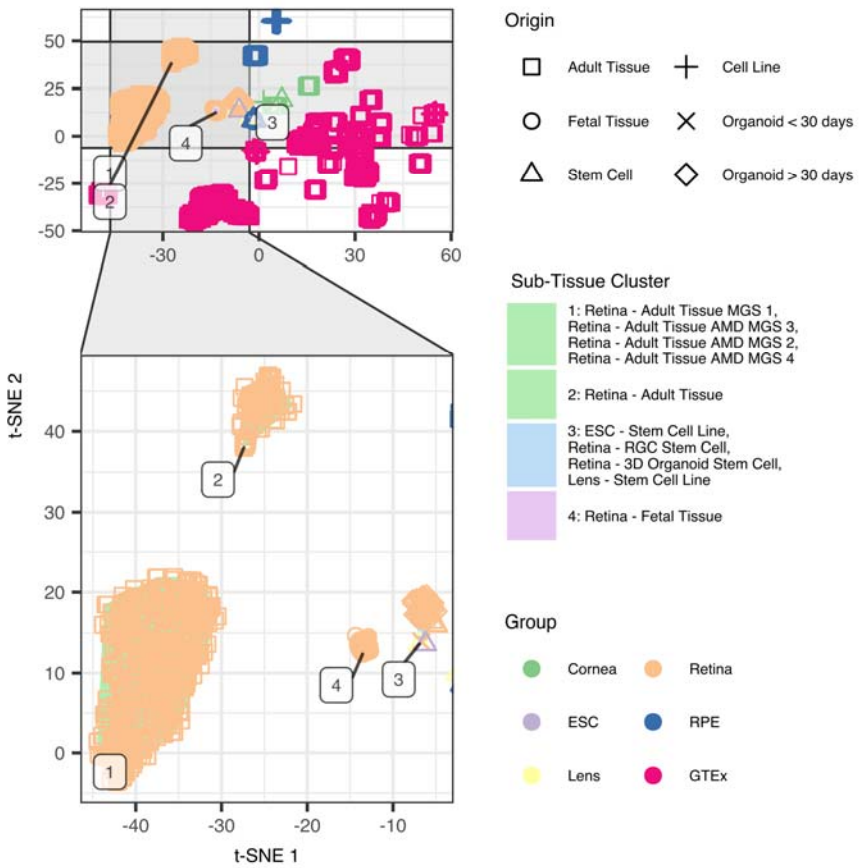
Supplementary Figures and Tables



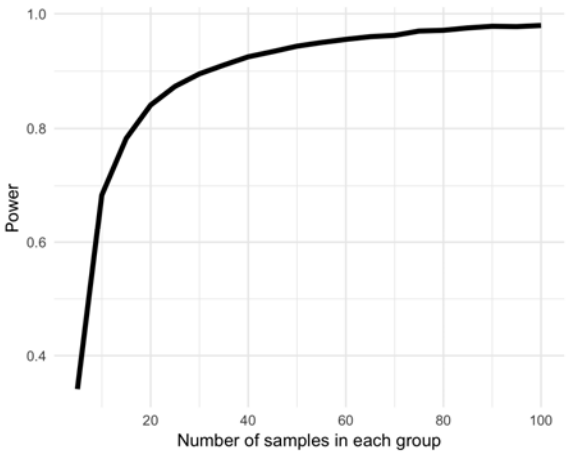
Supplemental Figure 1: As of June 2019, eyeIntegration has had usage across 367 cities and 47 countries.



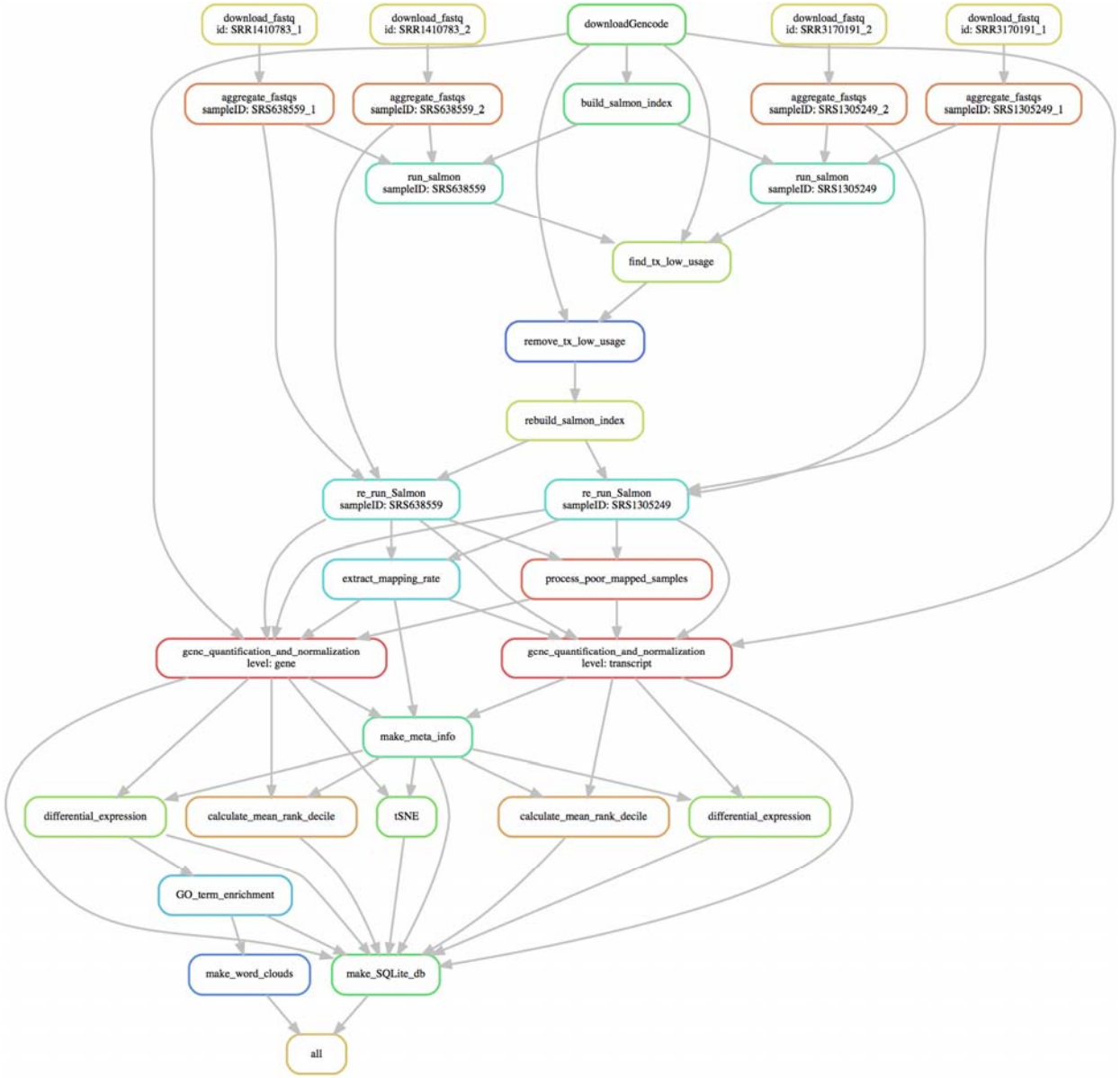
Supplemental Figure 2: Salmon mapping rate for each sample, grouped by tissue type. 1st quartile mapping rate is 51.5%, median is 80.5%, mean is 70.5%, and third quartile is 85.1%.



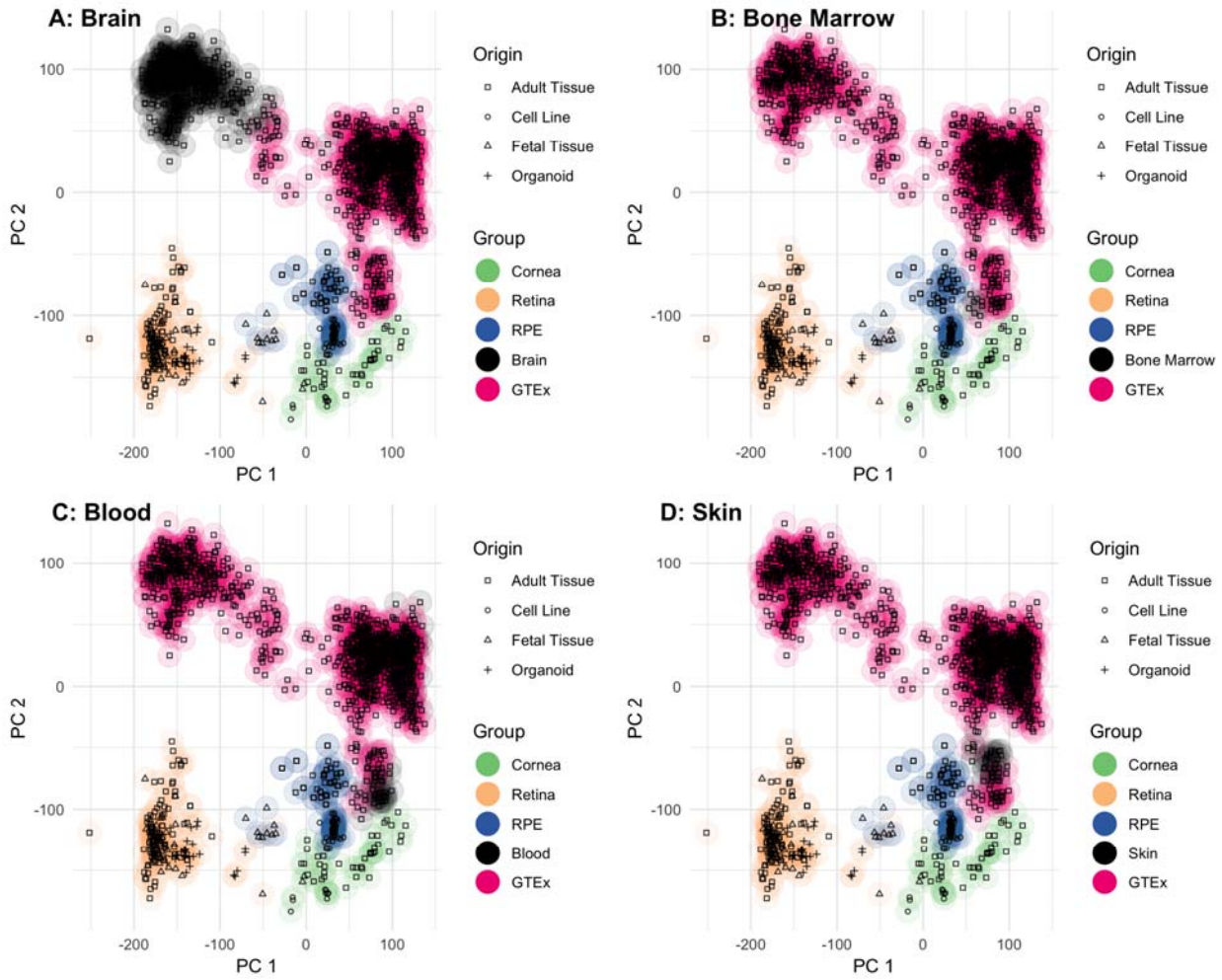
Supplemental Figure 3: Adult retina samples from Ratnapriya et al. (MGS 1 - 4, not-AMD is MGS 1, AMD is MGS 2 through 4) cluster independently from all other adult retina samples collected.



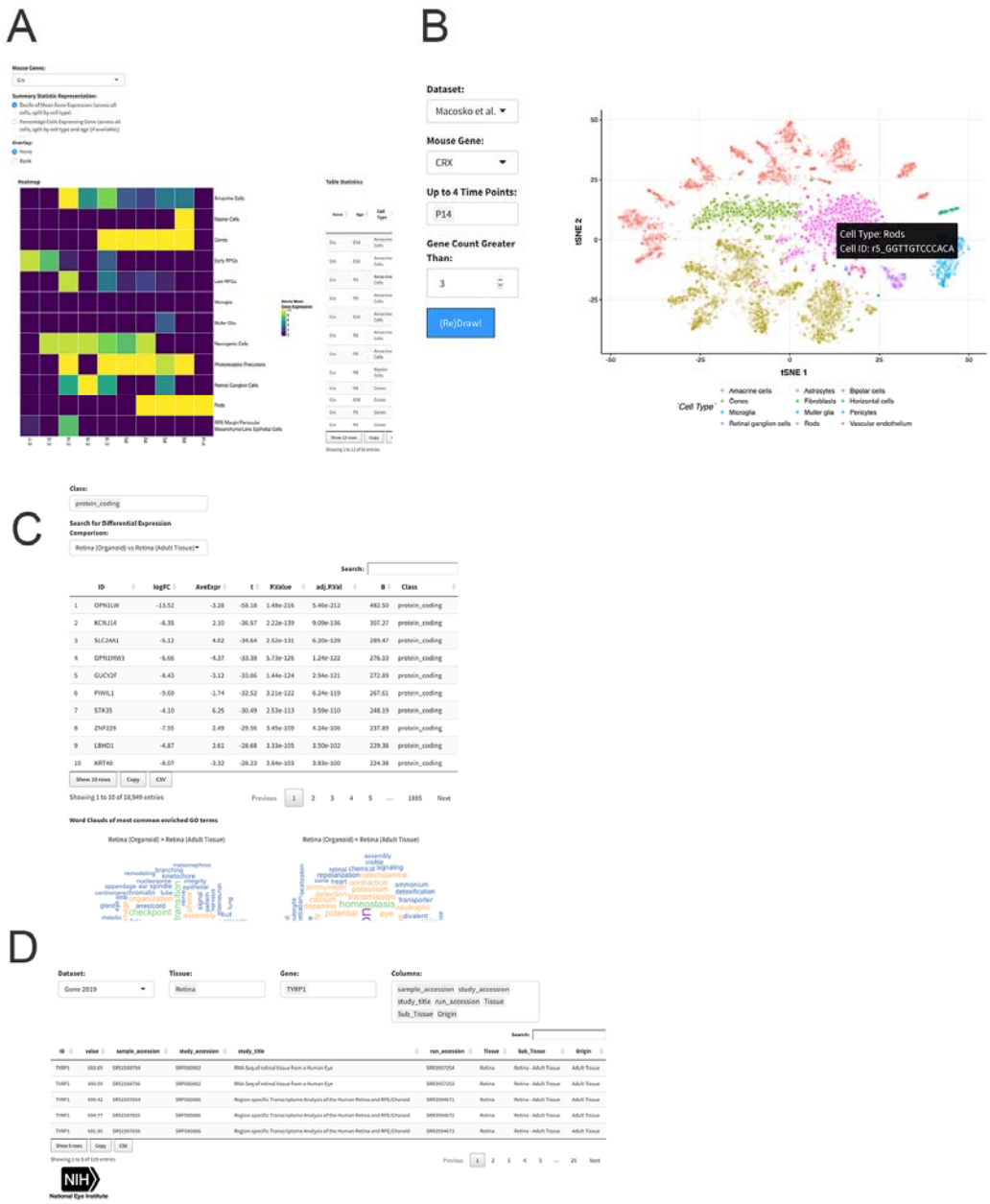
Supplemental Figure 4: Power curve to assess ability to detect >= 1 log2(FoldChange) in gene expression between n samples (x-axis) in each group.



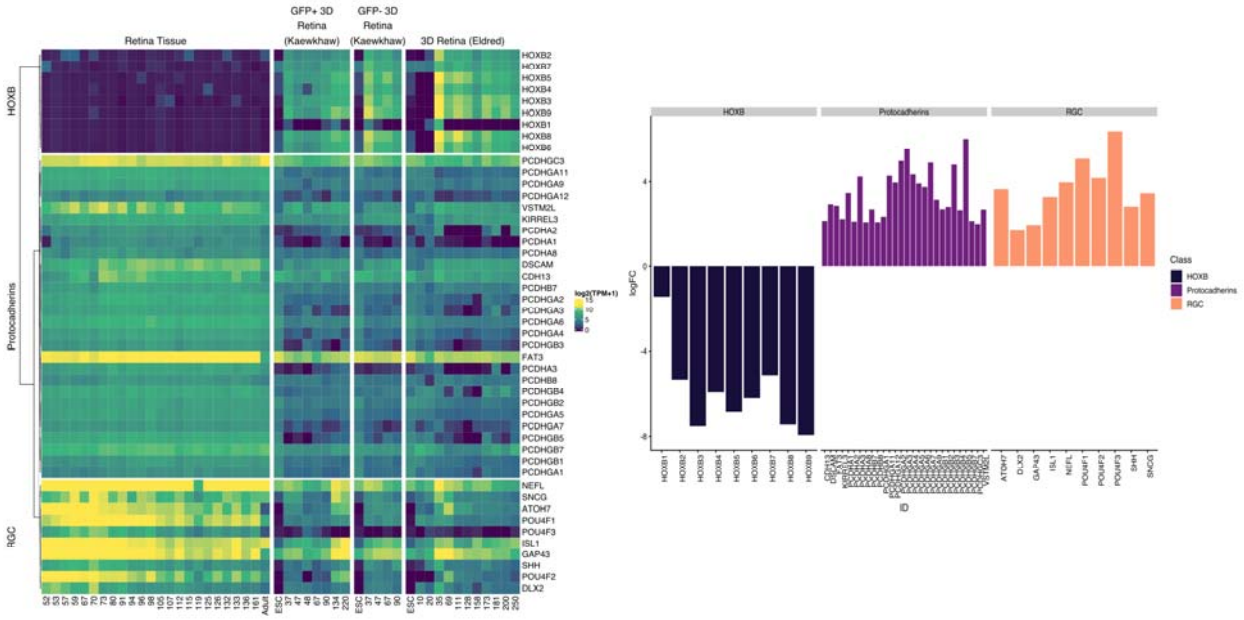
Supplemental Figure 5: Snakemake pipeline to create EiaD 2019 consists of small modular compute sections to ensure sample tracking through the full pipeline



Supplemental Figure 6: PCA plot of all samples suggests that the non-eye tissue most similar to adult retina is the brain, RPE and cornea are most similar to bone marrow, blood, and skin



Supplemental Figure 7: A. CRX mouse retina gene expression heatmap and table information from Clark et al. E11 to P14 scRNA-seq. B. t-SNE visualization of gene expression profiles of individual cells from Macosko et al. C. Data table export view. D. Differential gene expression across different tissues and with GO term enrichment.



Supplemental Figure 8: Heatmap of three sets of genes across fetal retina and organoids divided by age in days (A). Bar plot of differential expression from retina to organoid, where positive values are genes that are higher expressed in fetal tissue than organoid (B). All logFC expression values have FDR corrected p value < 0.01.

Group	Version	Pre QC Count	Count	Tissue Types (Count)
GTEx	2017	847	847	Adipose Tissue (39), Adrenal Gland (20), Blood (26), Blood Vessel (57), Brain (255), Breast (20), Colon (40), Esophagus (59), Heart (37), Kidney (17), Liver (20), Lung (20), Muscle (19), Nerve (20), Pancreas (20), Pituitary (20), Salivary Gland (20), Skin (59), Small Intestine (20), Spleen (20), Stomach (19), Thyroid (20)
GTEx	2019	1375	1314	Adipose Tissue (54), Adrenal Gland (30), Bladder (9), Blood (49), Blood Vessel (88), Bone Marrow (5), Brain (372), Breast (30), Cervix Uteri (11), Colon (58), Esophagus (87), Fallopian Tube (7), Heart (58), Kidney (27), Liver (30), Lung (29), Muscle (29), Nerve (30), Ovary (4), Pancreas (30), Pituitary (29), Prostate (5), Salivary Gland (30), Skin (83), Small Intestine (29), Spleen (30), Stomach (30), Testis (5), Thyroid (27), Uterus (4), Vagina (5)

Supplemental Table 1: EiaD holds hundreds of GTEx tissues to provide a broad comparison set

Biotype	Gene Count	Transcript Count	Definition
bidirectional_promoter_lncRNA	68	122	A non-coding locus that originates from within the promoter region of a protein-coding gene, with transcription proceeding in the opposite direction on the other strand.
unitary_pseudogene	22	21	A species-specific unprocessed pseudogene without a parent gene, as it has an active orthologue in another species.
retained_intron		6552	Alternatively spliced transcript believed to contain intronic sequence relative to other, coding, variants.
protein_coding	19012	34969	Contains an open reading frame (ORF).
processed_transcript	533	4635	Doesn't contain an ORF.
antisense	4014	5656	Has transcripts that overlap the genomic span (i.e. exon or introns) of a protein-coding locus on the opposite strand.

Biotype	Gene Count	Transcript Count	Definition
pseudogene	11	19	Have homology to proteins but generally suffer from a disrupted coding sequence and an active homologous gene can be found at another locus. Sometimes these entries have an intact coding sequence or an open but truncated ORF, in which case there is other evidence used (for example genomic polyA stretches at the 3' end) to classify them as a pseudogene. Can be further classified as one of the following.
nonsense_mediated_decay		3402	If the coding sequence (following the appropriate reference) of a transcript finishes >50bp from a downstream splice site then it is tagged as NMD. If the variant does not cover the full reference coding sequence then it is annotated as NMD if NMD is unavoidable i.e. no matter what the exon structure of the missing portion is the transcript will be subject to NMD.
IG_C_gene	14	17	Immunoglobulin (Ig) variable chain and T-cell receptor (TcR) genes imported or annotated according to the IMGT.
IG_J_gene	2	2	Immunoglobulin (Ig) variable chain and T-cell receptor (TcR) genes imported or annotated according to the IMGT.
IG_V_gene	127	127	Immunoglobulin (Ig) variable chain and T-cell receptor (TcR) genes imported or annotated according to the IMGT.
TR_C_gene	6	6	Immunoglobulin (Ig) variable chain and T-cell receptor (TcR) genes imported or annotated according to the IMGT.
TR_J_gene	6	6	Immunoglobulin (Ig) variable chain and T-cell receptor (TcR) genes imported or annotated according to the IMGT.
TR_V_gene	41	36	Immunoglobulin (Ig) variable chain and T-cell receptor (TcR) genes imported or annotated according to the IMGT.
IG_C_pseudogene	3	3	Inactivated immunoglobulin gene.
IG_J_pseudogene	1	1	Inactivated immunoglobulin gene.
IG_V_pseudogene	21	18	Inactivated immunoglobulin gene.
TR_V_pseudogene	4	4	Inactivated immunoglobulin gene.
sense_intronic	659	664	Long non-coding transcript in introns of a coding gene that does not overlap any exons.
sense_overlapping	153	180	Long non-coding transcript that contains a coding gene in its intron on the same strand.
lincRNA	4323	5881	Long, intervening noncoding (linc) RNA that can be found in evolutionarily conserved, intergenic regions.
rRNA_pseudogene	24	24	Non-coding RNA predicted to be pseudogene by the Ensembl pipeline
miRNA	130	130	Non-coding RNA predicted using sequences from Rfam and miRBase
misc_RNA	260	260	Non-coding RNA predicted using sequences from Rfam and miRBase
Mt_rRNA	2	2	Non-coding RNA predicted using sequences from Rfam and miRBase
Mt_tRNA	22	22	Non-coding RNA predicted using sequences from Rfam and miRBase
ribozyme	2	2	Non-coding RNA predicted using sequences from Rfam and miRBase
rRNA	5	5	Non-coding RNA predicted using sequences from Rfam and miRBase
scaRNA	15	15	Non-coding RNA predicted using sequences from Rfam and miRBase
scRNA	1	1	Non-coding RNA predicted using sequences from Rfam and miRBase
snoRNA	238	242	Non-coding RNA predicted using sequences from Rfam and miRBase
snRNA	104	106	Non-coding RNA predicted using sequences from Rfam and miRBase
polymorphic_pseudogene	19	22	Pseudogene owing to a SNP/DIP but in other individuals/haplotypes/strains the gene is translated.
unprocessed_pseudogene	593	585	Pseudogene that can contain introns since produced by gene duplication.
translated_processed_pseudogene	2	2	Pseudogene that has mass spec data suggesting that it is also translated.
processed_pseudogene	2537	2435	Pseudogene that lack introns and is thought to arise from reverse transcription of mRNA followed by reinsertion of DNA into the genome.
transcribed_processed_pseudogene	338	169	Pseudogene where protein homology or genomic structure indicates a pseudogene, but the presence of locus-specific transcripts indicates expression.
transcribed_unitary_pseudogene	101	38	Pseudogene where protein homology or genomic structure indicates a pseudogene, but the presence of locus-specific transcripts indicates expression.

Biotype	Gene Count	Transcript Count	Definition
transcribed_unprocessed_pseudogene	680	327	Pseudogene where protein homology or genomic structure indicates a pseudogene, but the presence of locus-specific transcripts indicates expression.
TEC	755	803	To be Experimentally Confirmed. This is used for non-spliced EST clusters that have polyA features. This category has been specifically created for the ENCODE project to highlight regions that could indicate the presence of protein coding genes that require experimental validation, either by 5' RACE or RT-PCR to extend the transcripts, or by confirming expression of the putatively-encoded peptide with specific antibodies.
non_stop_decay		10	Transcript that has polyA features (including signal) without a prior stop codon in the CDS, i.e. a non-genomic polyA tail attached directly to the CDS without 3' UTR. These transcripts are subject to degradation.
3prime_overlapping_ncRNA	25	28	Transcript where ditag and/or published experimental data strongly supports the existence of short non-coding transcripts transcribed from the 3'UTR.
non_coding	2	2	Transcript which is known from the literature to not be protein coding.
macro_lncRNA	1	1	Unspliced lncRNA that is several kb in size.

Supplemental Table 2: Dozens of different types of gene and transcript types quantified

sample accession	Tissue	Sub Tissue	Origin	Age Days	Kept	study title	sample attribute	study abstract	study accession	mapping rate
SRS846894	Cornea	Cornea - Cell Line Endothelium	Cell Line	.	Kept	Transcriptomic analysis of cultured corneal endothelial cells as a validation for their use in cell-replacement therapy	source_name: corneal endothelial cell line ‖ cell type: endothelial ‖ imortalization: telemorase ‖ culture medium: F99	The corneal endothelium plays a primary role in maintaining corneal homeostasis and clarity, and must be surgically replaced with allogenic donor corneal endothelium in the event of visually significant dysfunction. However, a worldwide shortage of donor corneal tissue has led to a search for alternative sources of transplantable tissue. Cultured human corneal endothelial cells (HCEnC) have been shown to restore corneal clarity in experimental models of corneal endothelial dysfunction in animal models, but characterization of cultured HCEnC remains incomplete. To this end, we utilized next-generation RNA sequencing technology to compare the transcriptomic profile of ex vivo human corneal endothelium (evHCEnC) with that of primary HCEnC and HCEnC lines, and to determine the utility of cultured and immortalized corneal endothelial cells as models of in vivo corneal endothelium. Multidimensional analyses of the transcriptome datasets demonstrated that primary HCEnC have a closer relationship to evHCEnC than do immortalized HCEnC. Subsequent analyses showed that the majority of the genes specifically expressed in HCEnC (not expressed in ex vivo corneal epithelium or fibroblasts) demonstrated a marked variability of expression in cultured cells compared with evHCEnC. In addition, genes associated with either corneal endothelial cell function or corneal endothelial dystrophies were investigated. Significant differences in gene expression and protein levels were observed in the cultured cells compared with evHCEnC for each of the genes tested except for AGBL1 and LOXHD1, which were not detected by RNA-seq or qPCR. Our transcriptomic analysis suggests that at a molecular level primary HCEnC most closely resemble evHCEC and thus represent a viable therapeutic option for managing corneal endothelial dysfunction. Our findings also suggest that investigators should perform an assessment of the entire transcriptome of cultured HCEnC prior to determination of the potential clinical utility of the cultured HCEnC for the management of corneal endothelial cell failure. Overall design: Transcriptomes from ex vivo corneal endothelium, primary cultures and three cell lines were compared. Three samples of each endothelial cell group were submitted for RNA sequencing for a total of 15 samples. The transcriptome for the ex vivo corneal endothelium was used as the reference (i.e., proxy for in vivo corneal endothelium). Transcript abundances for a subset of genes associated with corneal endothelial cell function or disease were validated with qPCR and western blot. Samples of ex vivo endothelium used for validation were independent replicates not used for RNA-sequencing.	SRP055101	83.700
SRS3493151	Retina	Retina - Adult Tissue AMD MGS 3	Adult Tissue	.	Kept	Integrated analysis of genetic variants regulating retinal transcriptome (GREx) identifies genes underlying age-related macular degeneration	id: GSM3191271: R42015-399pf_382-IR_L6; Homo sapiens; RNA-Seq ‖ sample: SRS3493151 ‖ run: SRR7461134 ‖ patient_number: 382 ‖ r_id: 382_3 ‖ donor: 07-1068 ‖ A69S_rs10490924: G/G ‖ Y402H_rs1061170: C/C ‖ os_od: OD ‖ age: 95 ‖ sex: F ‖ mgs_level: 3 ‖ cause_of_death: acute cardiac event ‖ death_category: Cardiovascular ‖ cataracts: NA ‖ cataract_timing: NA ‖ ocular_history: IOL pseudophakic ‖ compiled_medical_history: arthritis, HTN, high chol ‖ smoking_history: NA ‖ cholesterol: yes ‖ heart_disease: NA ‖ hypertension: yes ‖ postmortem_interval_hrs: 13.77 ‖ rna_isolation_date: 7/6/15 ‖ rna_isolation_batch: isobatch2 ‖ rin: 8 ‖ library_sequenced_date: 1/29/16	Age-related macular degeneration (AMD) is a complex multifactorial disease with at least 34 loci contributing to genetic susceptibility. To gain functional understanding of AMD genetics, we generated transcriptional profiles of retina from 453 individuals including both controls and cases at distinct stages of AMD. We integrated retinal transcriptomes, covering 13,662 protein-coding and 1,462 noncoding genes, with genotypes at over 9 million common single nucleotide polymorphisms (SNPs) for expression quantitative trait loci (eQTL) analysis of a tissue not included in Genotype-Tissue Expression (GTEx) and other large datasets. Cis-eQTL analysis revealed 10,474 genes under genetic regulation, including 4,541 eQTLs detected only in the retina. We then integrated the AMD-genome-wide association studies (GWAS) data with eQTLs and ascertained target genes at six loci. Furthermore, using transcriptome wide association analysis (TWAS), we identified 23 additional AMD-associated genes, including RLBP1, HIC1 and PARP12. Our studies expand the genetic landscape of AMD leading to direct targets for biological evaluation and establish the Genotype-Retina Expression (GREx) database as a resource for post-GWAS interpretation of retina-associated traits including glaucoma and diabetic retinopathy. Overall design: Retinal samples from 523 aged post-mortem human subjects from a spectrum of age-related macular degeneration (AMD) were RNA-seq profiled.	SRP151763	50.100

sample accession	Tissue	Sub Tissue	Origin	Age Days	Kept	study title	sample attribute	study abstract	study accession	mapping rate
SRS1597869	RPE	RPE - Adult Tissue	Adult Tissue	.	Kept	Region-specific Transcriptome Analysis of the Human Retina and RPE/Choroid	gap_accession: phs001151 || submitter handle: NEI_RetinaRPE_Choroid || biospecimen repository: NEI_RetinaRPE_Choroid || study name: Region-specific Transcriptome Analysis of the Human Retina and RPE/Choroid || study design: Control Set || biospecimen repository sample id: S16 || submitted sample id: S16 || submitted subject id: 3 || gap_sample_id: 1975628 || gap_subject_id: 1622644 || sex: female || body site: eye || histological type: RPE/choroid || analyte type: RNA || is tumor: No || subject is affected: No || molecular data type: RNA Seq (NGS) || gap_consent_code: 1 || gap_consent_short_name: GRU-NPU	Proper spatial differentiation of retinal cell types is necessary for normal human vision. Many retinal diseases, such as Best disease and male germ cell associated kinase (MAK)-associated retinitis pigmentosa, preferentially affect distinct topographic regions of the retina. While much is known about the distribution of cell types in the retina, the distribution of molecular components across the posterior pole of the eye has not been well-studied. To investigate regional difference in molecular composition of ocular tissues, we assessed differential gene expression across the temporal, macular, and nasal retina and retinal pigment epithelium (RPE)/choroid of human eyes using RNA-Seq. RNA from temporal, macular, and nasal retina and RPE/choroid from four human donor eyes was extracted, poly-A selected, fragmented, and sequenced as 100 bp read pairs. Digital read files were mapped to the human genome and analyzed for differential expression using the Tuxedo... (for more see dbGaP study page.)	SRP088886	84.000
SRS1747723	ESC	ESC - Stem Cell Line	Stem Cell	.	low salmon mapping rate	Transcriptomes during lens differentiation of human embryonic stem cells	isolate: human embryonic stem cells || age: not applicable || biomaterial_provider: Zhongshan Ophthalmic Centre, 54th Xianlie Road, Guangzhou, China || sex: not applicable || tissue: not applicable || cell_line: H9 human embryonic stem cell || cell_type: human embryonic stem cell || BioSampleModel: Human	In order to examine the fundamental mechanisms governing lens cells differentiation, we analyzed the transcriptome changes during the differentiation of human embryonic stem cells (hESCs) into lens cells and lentoid bodies. The differentiation of hESCs was induced by a sequential treatments of growth factors. In briefly,Noggin (100ng/ul) was applied from day0 to day6; then a combination ofbFGF (100ng/ul) and BMP4/7 (20 ng/ml)was addedfrom day 7 to day 18, followed by bFGF (100ng/ul) and Wnt3a(20 ng/ml) from day19 to day32. Cells at day 0, day 6, day 18 and day 32 were collected for analysis of paired-end RNA sequencing using Illuminia Hiseq 2500. The results revealed dynamic transcription network during lentoid bodies differentiation. We observed differential expression of genes involved insignaling pathways, which were considered to be necessary for lens development. These results provide a valuable resource for studying the mechanisms regulating in vitro lentoid body differentiation of hESCs and getting a glimpse of signaling pathway network in lens embryonic development.	SRP091605	17.900
SRS3492989	Retina	Retina - Adult Tissue AMD MGS 4	Adult Tissue	.	Kept	Integrated analysis of genetic variants regulating retinal transcriptome (GREx) identifies genes underlying age-related macular degeneration	id: GSM3191108: R42015-211pf_232-IR_L3; Homo sapiens; RNA-Seq | sample: SRS3492989 | run: SRR7460971 | patient_number: 232 | r_id: 232_4 | donor: 15-0040 | A69S_rs10490924: G/G | Y402H_rs1061170: C/T | os_od: OS | age: 103 | sex: M | mgs_level: 4 | cause_of_death: pneumonia | death_category: Pulmonary | cataracts: NA | cataract_timing: NA | ocular_history: confirmed phakic (OU), AMD | compiled_medical_history: NA | smoking_history: NA | cholesterol: NA | heart_disease: NA | hypertension: NA | postmortem_interval_hrs: 17.68 | rna_isolation_date: 6/24/15 | rna_isolation_batch: isobatch2 | rin: 7.6 | library_sequenced_date: 11/20/15	Age-related macular degeneration (AMD) is a complex multifactorial disease with at least 34 loci contributing to genetic susceptibility. To gain functional understanding of AMD genetics, we generated transcriptional profiles of retina from 453 individuals including both controls and cases at distinct stages of AMD. We integrated retinal transcriptomes, covering 13,662 protein-coding and 1,462 noncoding genes, with genotypes at over 9 million common single nucleotide polymorphisms (SNPs) for expression quantitative trait loci (eQTL) analysis of a tissue not included in Genotype-Tissue Expression (GTEx) and other large datasets. Cis-eQTL analysis revealed 10,474 genes under genetic regulation, including 4,541 eQTLs detected only in the retina. We then integrated the AMD-genome-wide association studies (GWAS) data with eQTLs and ascertained target genes at six loci. Furthermore, using transcriptome wide association analysis (TWAS), we identified 23 additional AMD-associated genes, including RLBP1, HIC1 and PARP12. Our studies expand the genetic landscape of AMD leading to direct targets for biological evaluation and establish the Genotype-Retina Expression (GREx) database as a resource for post-GWAS interpretation of retina-associated traits including glaucoma and diabetic retinopathy. Overall design: Retinal samples from 523 aged post-mortem human subjects from a spectrum of age-related macular degeneration (AMD) were RNA-seq profiled.	SRP151763	48.400
SRS3493218	Retina	Retina - Adult Tissue AMD MGS 2	Adult Tissue	.	Kept	Integrated analysis of genetic variants regulating retinal transcriptome (GREx) identifies genes underlying age-related macular degeneration	id: GSM3191338: R42016-403pf_442-IR_L7; Homo sapiens; RNA-Seq | sample: SRS3493218 | run: SRR7461201 | patient_number: 442 | r_id: 442_2 | donor: 15-1220 | A69S_rs10490924: G/G | Y402H_rs1061170: T/T | os_od: OS | age: 92 | sex: F | mgs_level: 2 | cause_of_death: complete heart block | death_category: Cardiovascular | cataracts: yes | cataract_timing: 2005 & 2011 (unsure which eye first) | ocular_history: confirmed pseudophakic (OU), cataracts (OU) | compiled_medical_history: CABG - 1979, hypertension, hyperlipidemia | smoking_history: NA | cholesterol: yes | heart_disease: NA | hypertension: yes | postmortem_interval_hrs: 18.82 | rna_isolation_date: 8/15/16 | rna_isolation_batch: isobatch2 | rin: 8.4 | library_sequenced_date: 9/23/16	Age-related macular degeneration (AMD) is a complex multifactorial disease with at least 34 loci contributing to genetic susceptibility. To gain functional understanding of AMD genetics, we generated transcriptional profiles of retina from 453 individuals including both controls and cases at distinct stages of AMD. We integrated retinal transcriptomes, covering 13,662 protein-coding and 1,462 noncoding genes, with genotypes at over 9 million common single nucleotide polymorphisms (SNPs) for expression quantitative trait loci (eQTL) analysis of a tissue not included in Genotype-Tissue Expression (GTEx) and other large datasets. Cis-eQTL analysis revealed 10,474 genes under genetic regulation, including 4,541 eQTLs detected only in the retina. We then integrated the AMD-genome-wide association studies (GWAS) data with eQTLs and ascertained target genes at six loci. Furthermore, using transcriptome wide association analysis (TWAS), we identified 23 additional AMD-associated genes, including RLBP1, HIC1 and PARP12. Our studies expand the genetic landscape of AMD leading to direct targets for biological evaluation and establish the Genotype-Retina Expression (GREx) database as a resource for post-GWAS interpretation of retina-associated traits including glaucoma and diabetic retinopathy. Overall design: Retinal samples from 523 aged post-mortem human subjects from a spectrum of age-related macular degeneration (AMD) were RNA-seq profiled.	SRP151763	43.800

sample accession	Tissue	Sub Tissue	Origin	Age Days	Kept	study title	sample attribute	study abstract	study accession	mapping rate
SRS1305277	Cornea	Cornea - Adult Tissue	Adult Tissue	.	Kept	Transcriptome profiling of human keratoconus corneas through RNA sequencing identifies collagen synthesis disruption and downregulation of core elements of TGF-ÃƒÂ¸, Hippo, and Wnt pathways	source_name: Cornea || disease state: non-KTCN || study: discovery || tissue: cornea	To understand better the factors contributing to keratoconus (KTCN), we used RNA sequencing to perform a transcriptome profile of human KTCN corneas. Over 82% of the genes and almost 75% of the transcripts detected as differentially expressed in KTCN and non-KTCN corneas were confirmed in the replication study using another set of samples. We used these differentially expressed genes to generate a network of KTCN-deregulated genes. We found an extensive disruption of collagen synthesis and maturation pathways, as well as downregulation of the core elements of the TGF-ÃƒÂ¸, Hippo, and Wnt signaling pathways influencing corneal organization. We identified long noncoding RNAs (lncRNAs) and conducted a computational analysis of their potential functions, and found that lncRNAs regulated the processing and expression of the aforementioned genes. This first comprehensive transcriptome profiling of human KTCN corneas points further to a complex etiology of KTCN. Overall design: Transcription profiling of 25 KTCN and 25 non-KTCN corneas using RNA-Seq	SRP070148	46.500
E-MTAB-4377.RNA49	Retina	Retina - Adult Tissue	Adult Tissue	.	Kept	RNAseq 50 Normal Human Retina		RNA-seq of post-mort retina donor without clinically relevant visual impairment. Ploy-A enriched. 75-nt paired-end. Short time lapse between tissue sampling and cDNA generation.	E-MTAB-4377	81.600
SRS3493229	Retina	Retina - Adult Tissue AMD MGS 3	Adult Tissue	.	Kept	Integrated analysis of genetic variants regulating retinal transcriptome (GREx) identifies genes underlying age-related macular degeneration	id: GSM3191346: R42016-110pf_45-IR_L3; Homo sapiens; RNA-Seq | sample: SRS3493229 | run: SRR7461209 | patient_number: 45 | r_id: 45_3 | donor: 14-1029 | A69S_rs10490924: G/G | Y402H_rs1061170: T/T | os_od: OD | age: 90 | sex: M | mgs_level: 3 | cause_of_death: respiratory failure | death_category: Pulmonary | cataracts: NA | cataract_timing: NA | ocular_history: cataracts (OU), confirmed pseudophakic | compiled_medical_history: HTN, Non-Hodgkins lymphoma, prostate cancer, bladder cancer, UTI, light smoker (20yrs) | smoking_history: yes | cholesterol: NA | heart_disease: NA | hypertension: yes | postmortem_interval_hrs: 16.02 | rna_isolation_date: 12/19/14 | rna_isolation_batch: isobatch1 | rin: 8.5 | library_sequenced_date: 3/8/16	Age-related macular degeneration (AMD) is a complex multifactorial disease with at least 34 loci contributing to genetic susceptibility. To gain functional understanding of AMD genetics, we generated transcriptional profiles of retina from 453 individuals including both controls and cases at distinct stages of AMD. We integrated retinal transcriptomes, covering 13,662 protein-coding and 1,462 noncoding genes, with genotypes at over 9 million common single nucleotide polymorphisms (SNPs) for expression quantitative trait loci (eQTL) analysis of a tissue not included in Genotype-Tissue Expression (GTEx) and other large datasets. Cis-eQTL analysis revealed 10,474 genes under genetic regulation, including 4,541 eQTLs detected only in the retina. We then integrated the AMD-genome-wide association studies (GWAS) data with eQTLs and ascertained target genes at six loci. Furthermore, using transcriptome wide association analysis (TWAS), we identified 23 additional AMD-associated genes, including RLBP1, HIC1 and PARP12. Our studies expand the genetic landscape of AMD leading to direct targets for biological evaluation and establish the Genotype-Retina Expression (GREx) database as a resource for post-GWAS interpretation of retina-associated traits including glaucoma and diabetic retinopathy. Overall design: Retinal samples from 523 aged post-mortem human subjects from a spectrum of age-related macular degeneration (AMD) were RNA-seq profiled.	SRP151763	50.500
SRS2582161	Retina	Retina - Fetal Tissue	Fetal Tissue	73	Kept	Molecular anatomy of the developing human retina	D73P		SRP119766	74.500

Supplemental Table 3: Full metadata for 10 random eye samples. Full metadata available as supplementary file "metadata.csv."

Sub Tissue	Cluster	Count
Retina - Adult Tissue	1	107
RPE - Cell Line	2	50
Cells - EBV-transformed lymphocytes	3	30
Pancreas	4	30
Lens - Stem Cell Line	5	2
RPE - Fetal Tissue	5	7
RPE - Stem Cell Line	5	16
Cornea - Adult Tissue	6	25
Cornea - Endothelium	7	16
Cornea - Fetal Endothelium	7	2

Sub Tissue	Cluster	Count
Cornea - Stem Cell Endothelium	8	4
ESC - Stem Cell Line	9	6
Lens - Stem Cell Line	9	2
Retina - 3D Organoid Stem Cell	9	5
Retina - Fetal Eye	9	3
Retina - 3D Organoid Stem Cell	10	47
Retina - RGC Stem Cell	10	3
RPE - Adult Tissue	11	48
Retina - Fetal Tissue	12	35
Brain - Amygdala	13	30
Brain - Anterior cingulate cortex (BA24)	13	27
Brain - Caudate (basal ganglia)	13	28
Brain - Cerebellar Hemisphere	13	1
Brain - Cortex	13	28
Brain - Frontal Cortex (BA9)	13	29
Brain - Hippocampus	13	29
Brain - Hypothalamus	13	28
Brain - Nucleus accumbens (basal ganglia)	13	30
Brain - Putamen (basal ganglia)	13	28
Brain - Spinal cord (cervical c-1)	13	28
Brain - Substantia nigra	13	28
Pituitary	13	1
Lung	14	29
Whole Blood	15	19
Brain - Cerebellar Hemisphere	16	28
Brain - Cerebellum	16	29
Brain - Cortex	16	1
Thyroid	17	27
Heart - Atrial Appendage	18	30
Heart - Left Ventricle	18	28
Skin - Not Sun Exposed (Suprapubic)	19	29
Skin - Sun Exposed (Lower leg)	19	24
Muscle - Skeletal	20	29
Prostate	20	1

Sub Tissue	Cluster	Count
Artery - Aorta	21	30
Artery - Coronary	21	27
Artery - Tibial	21	28
Adipose - Subcutaneous	22	27
Adipose - Visceral (Omentum)	22	27
Artery - Coronary	22	3
Breast - Mammary Tissue	22	8
Esophagus - Gastroesophageal Junction	22	1
Small Intestine - Terminal Ileum	22	1
Nerve - Tibial	23	30
Bladder	24	4
Breast - Mammary Tissue	24	22
Cervix - Endocervix	24	2
Fallopian Tube	24	2
Minor Salivary Gland	24	23
Prostate	24	4
Vagina	24	1
Pituitary	25	28
Kidney - Cortex	26	27
Stomach	27	21
Bladder	28	5
Colon - Sigmoid	28	29
Colon - Transverse	28	29
Esophagus - Gastroesophageal Junction	28	28
Esophagus - Muscularis	28	28
Small Intestine - Terminal Ileum	28	27
Stomach	28	9
Adrenal Gland	29	30
Small Intestine - Terminal Ileum	30	1
Spleen	30	30
Cells - Transformed fibroblasts	31	30
Liver	32	30
Cervix - Ectocervix	33	2
Esophagus - Mucosa	33	30

Sub Tissue	Cluster	Count
Minor Salivary Gland	33	7
Vagina	33	3
Cornea - Cell Line Endothelium	34	9
Cells - Leukemia cell line (CML)	35	5
Cervix - Ectocervix	36	4
Cervix - Endocervix	36	3
Fallopian Tube	36	5
Ovary	36	4
Testis	36	1
Uterus	36	4
Vagina	36	1
Testis	37	4

Supplemental Table 4: Counts of Sub Tissues in each tSNE - dbscan based cluster group

ID	Description	GeneRatio	pvalue	qvalue	geneID
GO:0030198	extracellular matrix organization	51/668	3.11e-17	0.000	VCAM1 FOXC1 P4HA1 LOX NID2 COL1A2 COL27A1 ICAM1 DDR2 WT1 COL14A1 POSTN COL3A1 PDPN HAS2 SMOC2 COL5A1 TGFBI COL6A3 DCN CDH1 TTR SULF1 ECM2 ADAMTS5 CCDC80 ITGB4 ITGA8 COL4A6 FOXF1 DPP4 ITGB6 FN1 COL1A1 ITGB3 SERPINE1 FMOD NPNT MMP19 CREB3L1 A2M BGN COL8A1 TNXB GFAP SPINT1 MMP7 ELF3 THBS1 COL17A1 MYH11
GO:0007156	homophilic cell adhesion via plasma membrane adhesion molecules	28/592	1.54e-12	0.000	PCDHGB5 PCDHGA6 PCDHGA7 PCDHGA3 PCDHGA5 PCDHGA2 PCDHGB3 PCDHGA4 PCDHGA11 PCDHGA9 PCDHGB2 PCDHA3 PCDHGA12 PCDHB7 PCDHGB4 PCDHGA1 PCDHGB1 PCDHGB7 FAT3 DSCAM PCDHA1 PCDHGC3 PCDHB8 VSTM2L KIRREL3 PCDHA2 PCDHA8 CDH13
GO:0050878	regulation of body fluid levels	50/668	6.53e-11	0.000	CYP26B1 FZD6 GUCA1B LYN COL1A2 XDH ADM DOCK6 TFPI2 COL3A1 PDPN HEG1 HAS2 SLC7A11 HK2 EDNRB AQP4 CLDN4 PLAU PLAT ENTPD2 F2RL1 HRG SCNN1A CD9 CLDN1 APOE FOXB1 EHD2 COL1A1 ITGB3 SERPINE1 TRPV4 PRTN3 PRKCG SCNN1B CAV1 KDF1 NFE2 A2M SCNN1G AQP1 GRHL3 ALOX12B TP63 THBS1 AGR2 ANXA8 FLG2 KRT1

ID	Description	GeneRatio	pvalue	qvalue	geneID
GO:0061448	connective tissue development	34/668	1.85e-10	0.000	HOXB3 HOXC4 HOXA5 HOXA3 BMP5 BMPR1B COL27A1 MGP GDF7 WT1 WNT11 ALX1 COL14A1 PLA2G16 PRRX1 MSX2 SNAI2 COL5A1 FOXA1 TGFBI COL6A3 STC1 SULF1 EGR1 MATN2 FRZB OSR2 COL1A1 EFEMP1 PITX1 TRPV4 NKX3-2 ACTA2 CHI3L1
GO:0031589	cell-substrate adhesion	37/668	3.08e-09	0.000	VCAM1 NID2 GBP1 POSTN COL3A1 PDPN HAS2 SMOC2 PTN ALOX15 ACTN3 HOXA7 ECM2 PLAU EPHA1 TNFRSF12A CCDC80 ITGB4 HRG ITGA8 FOXF1 RHOD CCL28 ANGPT1 ITGB6 FN1 COL1A1 ITGB3 SERPINE1 NPNT WNT4 COL8A1 CEACAM6 THBS1 AGR2 APOD COL17A1
GO:0030335	positive regulation of cell migration	45/668	1.04e-08	0.000	HGF LGALS3 IGFBP5 EGF LYN VEGFD ICAM1 DDR2 WNT11 ANXA1 POSTN PDPN HAS2 SEMA3C GPNMB SNAI2 SMOC2 PTP4A1 PLAU EPHA1 FPR2 F2RL1 IL1R1 PTGS2 FOXF1 RHOD FAM110C ANGPT1 EDN3 SRPX2 FOXC2 FN1 COL1A1 ITGB3 SERPINE1 TRPV4 ENPP2 GRB7 ANXA3 CAV1 CEACAM6 S100A14 RAB25 THBS1 NR4A3
GO:0001503	ossification	38/668	2.00e-08	0.000	IFITM1 HGF GPC3 EGR2 IGFBP5 FOXC1 HOXA3 DHRS3 BMP5 BMPR1B MGP DDR2 WNT11 IGF2 TOB1 MSX2 FHL2 GPNMB SNAI2 IGFBP3 FGFR2 PTN ASPN STC1 ALOX15 CEBPB PTGS2 PHOSPHO1 KREMEN2 OSR2 ISG15 COL1A1 NPNT DLX5 ALPL CREB3L1 WNT4 TP63
GO:0050678	regulation of epithelial cell proliferation	32/668	3.05e-07	0.000	GPC3 HOXA5 BMP5 KLF9 SIX1 PAX2 VEGFD XDH HAS2 EGR3 SNAI2 EDNRB FGFR2 PTN SULF1 DLX6 ALDH1A2 SERPINF1 APOE OSR2 ESRP2 ZFP36 ITGB3 DLX5 HTR2B CAV1 KDF1 CCL2 TP63 THBS1 NR4A3 NR4A1
GO:0030099	myeloid cell differentiation	29/592	8.11e-05	0.006	HIST1H4L HIST1H4F POU4F2 HIST1H3F POU4F1 HIST1H3I HIST1H3C HIST1H3B HIST1H3J PRKCB HIST1H4C SIGLEC15 HIST1H4A TREM2 HIST1H4I HIST1H3A LILRB4 FES C1QC HIST1H3G HIST1H4D CSF1R HIST1H4B TYROBP HIST1H4J GPR68 EVI2B CCL3 SPI1
GO:0030099	myeloid cell differentiation	29/668	6.08e-04	0.008	HIST1H4K HOXB8 HOXB7 PIR GPC3 LGALS3 HOXA5 LYN UBD INHBA EPO HOXA7 CEBPB B2M F2RL1 LIF MT1G IL31RA MB ISG15 ZFP36 HOXA9 PRTN3 NFE2 CA2 IRF4 FOS THBS1 NR4A3
GO:0050769	positive regulation of neurogenesis	28/592	1.20e-03	0.043	POU4F2 MYB P2RY12 HEYL E2F1 VWC2L ASCL1 TRIM67 NEFL DSCAM ISLR2 ALKAL2 PAX6 GDF6 CX3CR1 FES OLIG2 STMN2 SHH RIT2 NEUROG3 CUX2 CPNE6 ADRA2B IRX3 DKK1 ITPKA DLX1

Supplemental Table 5: Top GO terms enriched between fetal retina and organoid retina, after paring of redundant terms with REVIGO[70]

Acknowledgements

We would like the thank the dozens of groups who provided the raw data required to create this project. We keep a running list of the projects and associated citations at https://github.com/davemcg/eyeIntegration_app/blob/master/inst/citations.md and strongly encourage anyone who uses EiaD and eyeIntegration to cite relevant projects. We would also like to thank Kapil Bharti, Robert Hufnagel, and Brian Brooks for their continuous set of critiques and suggestions in the development of eyeIntegration app over the past two years. Tiziana Cogliati was especially helpful in the editing of this manuscript. We also would like to thank the two anonymous reviewers for their careful reading and constructive criticisms. Finally, this work utilized the computational resources of the NIH HPC Biowulf cluster (http://hpc.nih.gov).

Funding

This research was supported by the Intramural Research Program of the National Eye Institute, National Institutes of Health.

Bibliography

1. Chen Y, Huang K, Nakatsu MN, Xue Z, Deng SX, Fan G. Identification of novel molecular markers through transcriptomic analysis in human fetal and adult corneal endothelial cells. *Hum Mol Genet*. 2013;22(7):1271-1279. doi:10.1093/hmg/dds527

2. Chng Z, Peh GSL, Herath WB, et al. High Throughput Gene Expression Analysis Identifies Reliable Expression Markers of Human Corneal Endothelial Cells. *PLOS ONE*. 2013;8(7):e67546. doi:10.1371/journal.pone.0067546

3. Chung DD, Frausto RF, Lin BR, Hanser EM, Cohen Z, Aldave AJ. Transcriptomic Profiling of Posterior Polymorphous Corneal Dystrophy. *Invest Ophthalmol Vis Sci*. 2017;58(7):3202-3214. doi:10.1167/iovs.17-21423

4. Frausto RF, Le DJ, Aldave AJ. Transcriptomic Analysis of Cultured Corneal Endothelial Cells as a Validation for Their Use in Cell Replacement Therapy. *Cell Transplant*. 2016;25(6):1159-1176. doi:10.3727/096368915X688948

5. Kabza M, Karolak JA, Rydzanicz M, et al. Collagen synthesis disruption and downregulation of core elements of TGF-β, Hippo, and Wnt pathways in keratoconus corneas. *European Journal of Human Genetics*. 2017;25(5):582-590. doi:10.1038/ejhg.2017.4

6. Ouyang H, Xue Y, Lin Y, et al. WNT7A and PAX6 define corneal epithelium homeostasis and pathogenesis. *Nature*. 2014;511(7509):358-361. doi:10.1038/nature13465

7. Song Q, Yuan S, An Q, et al. Directed differentiation of human embryonic stem cells to corneal endothelial cell-like cells: A transcriptomic analysis. *Experimental Eye Research*. 2016;151:107-114. doi:10.1016/j.exer.2016.08.004

8. Han C, Li J, Wang C, et al. Wnt5a Contributes to the Differentiation of Human Embryonic Stem Cells into Lentoid Bodies Through the Noncanonical Wnt/JNK Signaling Pathway. *Invest Ophthalmol Vis Sci*. 2018;59(8):3449-3460. doi:10.1167/iovs.18-23902

9. Aldiri I, Xu B, Wang L, et al. The Dynamic Epigenetic Landscape of the Retina During Development, Reprogramming, and Tumorigenesis. *Neuron*. 2017;94(3):550-568.e10. doi:10.1016/j.neuron.2017.04.022

10. Farkas MH, Grant GR, White JA, Sousa ME, Consugar MB, Pierce EA. Transcriptome analyses of the human retina identify unprecedented transcript diversity and 3.5 Mb of novel transcribed sequence via significant alternative splicing and novel genes. *BMC Genomics*. 2013;14(1):486. doi:10.1186/1471-2164-14-486

11. Hoshino A, Ratnapriya R, Brooks MJ, et al. Molecular Anatomy of the Developing Human Retina. *Developmental Cell*. 2017;43(6):763-779.e4. doi:10.1016/j.devcel.2017.10.029

12. Kaewkhaw R, Kaya KD, Brooks M, et al. Transcriptome Dynamics of Developing Photoreceptors in Three-Dimensional Retina Cultures Recapitulates Temporal Sequence of Human Cone and Rod Differentiation Revealing Cell Surface Markers and Gene Networks. *Stem Cells*. 2015;33(12):3504-3518. doi:10.1002/stem.2122

13. Kaewkhaw R, Swaroop M, Homma K, et al. Treatment Paradigms for Retinal and Macular Diseases Using 3-D Retina Cultures Derived From Human Reporter Pluripotent Stem Cell Lines. *Invest Ophthalmol Vis Sci*. 2016;57(5):ORSFl1-ORSFl11. doi:10.1167/iovs.15-17639

14. Li M, Jia C, Kazmierkiewicz KL, et al. Comprehensive analysis of gene expression in human retina and supporting tissues. *Hum Mol Genet*. 2014;23(15):4001-4014. doi:10.1093/hmg/ddu114

15. Mustafi D, Kevany BM, Bai X, et al. Transcriptome analysis reveals rod/cone photoreceptor specific signatures across mammalian retinas. *Hum Mol Genet*. 2016;25(20):4376-4388. doi:10.1093/hmg/ddw268

16. Pinelli M, Carissimo A, Cutillo L, et al. An atlas of gene expression and gene co-regulation in the human retina. *Nucl Acids Res*. 2016;44(12):5773-5784. doi:10.1093/nar/gkw486

17. Whitmore SS, Wagner AH, DeLuca AP, et al. Transcriptomic analysis across nasal, temporal, and macular regions of human neural retina and RPE/choroid by RNA-Seq. *Experimental Eye Research*. 2014;129:93-106. doi:10.1016/j.exer.2014.11.001

18. Darrow EM, Huntley MH, Dudchenko O, et al. Deletion of DXZ4 on the human inactive X chromosome alters higher-order genome architecture. *PNAS*. 2016;113(31):E4504-E4512. doi:10.1073/pnas.1609643113

19. Harenza JL, Diamond MA, Adams RN, et al. Transcriptomic profiling of 39 commonly-used neuroblastoma cell lines. *Scientific Data*. 2017;4:170033. doi:10.1038/sdata.2017.33

20. Hu G, Huang K, Yu J, et al. Identification of miRNA Signatures during the Differentiation of hESCs into Retinal Pigment Epithelial Cells. *PLOS ONE*. 2012;7(7):e37224. doi:10.1371/journal.pone.0037224

21. Nozawa R-S, Boteva L, Soares DC, et al. SAF-A Regulates Interphase Chromosome Structure through Oligomerization with Chromatin-Associated RNAs. *Cell*. 2017;169(7):1214-1227.e18. doi:10.1016/j.cell.2017.05.029

22. Oberstein A, Shenk T. Cellular responses to human cytomegalovirus infection: Induction of a mesenchymal-to-epithelial transition (MET) phenotype. *PNAS*. 2017;114(39):E8244-E8253. doi:10.1073/pnas.1710799114

23. Peng S, Gan G, Qiu C, et al. Engineering a Blood-Retinal Barrier With Human Embryonic Stem Cell-Derived Retinal Pigment Epithelium: Transcriptome and Functional Analysis. *STEM CELLS Translational Medicine*. 2013;2(7):534-544. doi:10.5966/sctm.2012-0134

24. Radeke MJ, Radeke CM, Shih Y-H, et al. Restoration of mesenchymal retinal pigmented epithelial cells by TGFβ pathway inhibitors: Implications for age-related macular degeneration. *Genome Medicine*. 2015;7(1):58. doi:10.1186/s13073-015-0183-x

25. Saini JS, Corneo B, Miller JD, et al. Nicotinamide Ameliorates Disease Phenotypes in a Human iPSC Model of Age-Related Macular Degeneration. *Cell Stem Cell*. 2017;20(5):635-647.e7. doi:10.1016/j.stem.2016.12.015

26. Samuel W, Jaworski C, Postnikova OA, et al. Appropriately differentiated ARPE-19 cells regain phenotype and gene expression profiles similar to those of native RPE cells. *Mol Vis*. 2017;23:60-89. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5360456/. Accessed January 17, 2019.

27. Santaguida S, Vasile E, White E, Amon A. Aneuploidy-induced cellular stresses limit autophagic degradation. *Genes Dev*. 2015;29(19):2010-2021. doi:10.1101/gad.269118.115

28. Shao Z, Wang H, Zhou X, et al. Spontaneous generation of a novel foetal human retinal pigment epithelium (RPE) cell line available for investigation on phagocytosis and morphogenesis. *Cell Proliferation*. 2017;50(6):e12386. doi:10.1111/cpr.12386

29. Shih Y-H, Radeke MJ, Radeke CM, Coffey PJ. Restoration of Mesenchymal RPE by Transcription Factor-Mediated Reprogramming. *Invest Ophthalmol Vis Sci*. 2017;58(1):430-441. doi:10.1167/iovs.16-20018

30. Smith JR, Todd S, Ashander LM, et al. Retinal Pigment Epithelial Cells are a Potential Reservoir for Ebola Virus in the Human Eye. *Trans Vis Sci Tech*. 2017;6(4):12-12. doi:10.1167/tvst.6.4.12

31. Stevenson NL, Bergen DJM, Skinner REH, et al. Giantin-knockout models reveal a feedback loop between Golgi function and glycosyltransferase expression. *J Cell Sci*. 2017;130(24):4132-4143. doi:10.1242/jcs.212308

32. Tresini M, Warmerdam DO, Kolovos P, et al. The core spliceosome as target and effector of non-canonical ATM signalling. *Nature*. 2015;523(7558):53-58. doi:10.1038/nature14512

33. Wheway G, Schmidts M, Mans DA, et al. An siRNA-based functional genomics screen for the identification of regulators of ciliogenesis and ciliopathy genes. *Nature Cell Biology*. 2015;17(8):1074-1087. doi:10.1038/ncb3201

34. Au ED, Fernandez-Godino R, Kaczynksi TJ, Sousa ME, Farkas MH. Characterization of lincRNA expression in the human retinal pigment epithelium and differentiated induced pluripotent stem cells. *PLOS ONE*. 2017;12(8):e0183939. doi:10.1371/journal.pone.0183939

35. Carithers LJ, Ardlie K, Barcus M, et al. A Novel Approach to High-Quality Postmortem Tissue Procurement: The GTEx Project. *Biopreservation and Biobanking*. 2015;13(5):311-319. doi:10.1089/bio.2015.0032

36. GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature*. 2017;550(7675):204-213. doi:10.1038/nature24277

37. Ratnapriya R, Sosina OA, Starostik MR, et al. Retinal transcriptome and eQTL analyses identify genes associated with age-related macular degeneration. *Nat Genet*. February 2019. doi:10.1038/s41588-019-0351-9

38. Collado-Torres L, Nellore A, Kammers K, et al. Reproducible RNA-seq analysis using *Recount2*. *Nature Biotechnology*. 2017;35:319-321. doi:10.1038/nbt.3838

39. Lachmann A, Torre D, Keenan AB, et al. Massive mining of publicly available RNA-seq data from human and mouse. *Nature Communications*. 2018;9(1):1366. doi:10.1038/s41467-018-03751-6

40. Budak G, Dash S, Srivastava R, Lachke SA, Janga SC. Express: A database of transcriptome profiles encompassing known and novel transcripts across multiple development stages in eye tissues. *Exp Eye Res*. 2018;168:57-68. doi:10.1016/j.exer.2018.01.009

41. Kakrana A, Yang A, Anand D, et al. iSyTE 2.0: A database for expression-based gene discovery in the eye. *Nucleic Acids Res*. 2018;46(D1):D875-D885. doi:10.1093/nar/gkx837

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

43. Leinonen R, Sugawara H, Shumway M. The Sequence Read Archive. *Nucleic Acids Res*. 2011;39(Database issue):D19-D21. doi:10.1093/nar/gkq1019

44. Zhu Y, Stephens RM, Meltzer PS, Davis SR. SRAdb: Query and use public next-generation sequencing data from within R. *BMC Bioinformatics*. 2013;14:19. doi:10.1186/1471-2105-14-19

45. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25(16):2078-2079. doi:10.1093/bioinformatics/btp352

46. Harrow J, Frankish A, Gonzalez JM, et al. GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Res*. 2012;22(9):1760-1774. doi:10.1101/gr.135350.111

47. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*. 2017;14(4):417-419. doi:10.1038/nmeth.4197

48. Soneson C, Matthes KL, Nowicka M, Law CW, Robinson MD. Isoform prefiltering improves performance of count-based methods for analysis of differential transcript usage. *Genome Biol*. 2016;17:12. doi:10.1186/s13059-015-0862-3

49. Soneson C, Love MI, Robinson MD. Differential analyses for RNA-seq: Transcript-level estimates improve gene-level inferences. *F1000Research*. 2016;4:1521. doi:10.12688/f1000research.7563.2

50. Hicks SC, Okrah K, Paulson JN, Quackenbush J, Irizarry RA, Bravo HC. Smooth quantile normalization. *Biostatistics*. 2018;19(2):185-198. doi:10.1093/biostatistics/kxx028

51. Robinson MD, McCarthy DJ, Smyth GK. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. 2010;26(1):139-140. doi:10.1093/bioinformatics/btp616

52. Ritchie ME, Phipson B, Wu D, et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*. 2015;43(7):e47. doi:10.1093/nar/gkv007

53. Wright FA, Sullivan PF, Brooks AI, et al. Heritability and genomics of gene expression in peripheral blood. *Nature Genetics*. 2014;46(5):430-437. doi:10.1038/ng.2951

54. Zhang C, Zhang B, Lin L-L, Zhao S. Evaluation and comparison of computational tools for RNA-seq isoform quantification. *BMC Genomics*. 2017;18(1):583. doi:10.1186/s12864-017-4002-1

55. Law CW, Chen Y, Shi W, Smyth GK. Voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*. 2014;15:R29. doi:10.1186/gb-2014-15-2-r29

56. Yu G, Wang L-G, Han Y, He Q-Y. clusterProfiler: An R Package for Comparing Biological Themes Among Gene Clusters. *OMICS*. 2012;16(5):284-287. doi:10.1089/omi.2011.0118

57. Köster J, Rahmann S. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*. 2012;28(19):2520-2522. doi:10.1093/bioinformatics/bts480

58. Macosko EZ, Basu A, Satija R, et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*. 2015;161(5):1202-1214. doi:10.1016/j.cell.2015.05.002

59. Clark BS, Stein-O'Brien GL, Shiau F, et al. Single-Cell RNA-Seq Analysis of Retinal Development Identifies NFI Factors as Regulating Mitotic Exit and Late-Born Cell Specification. *Neuron*. May 2019. doi:10.1016/j.neuron.2019.04.010

60. Bi R, Liu P. Sample size calculation while controlling false discovery rate for differential expression analysis with RNA-sequencing experiments. *BMC Bioinformatics*. 2016;17:146. doi:10.1186/s12859-016-0994-9

61. Wattenberg M, Viégas F, Johnson I. How to Use t-SNE Effectively. *Distill*. 2016;1(10):e2. doi:10.23915/distill.00002

62. Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*. 2016;32(18):2847-2849. doi:10.1093/bioinformatics/btw313

63. Eldred KC, Hadyniak SE, Hussey KA, et al. Thyroid hormone signaling specifies cone subtypes in human retinal organoids. *Science*. 2018;362(6411):eaau6348. doi:10.1126/science.aau6348

64. Lefebvre JL, Zhang Y, Meister M, Wang X, Sanes JR. Γ-Protocadherins regulate neuronal survival but are dispensable for circuit formation in retina. *Development*. 2008;135(24):4141-4151. doi:10.1242/dev.027912

65. Lukowski S, Lo C, Sharov A, et al. Generation of human neural retina transcriptome atlas by single cell RNA sequencing. *bioRxiv*. September 2018:425223. doi:10.1101/425223

66. Lu Y, Yi W, Wu Q, et al. Single-cell RNA-seq analysis maps the development of human fetal retina. *bioRxiv*. September 2018:423830. doi:10.1101/423830

67. Peng Y-R, Shekhar K, Yan W, et al. Molecular Classification and Comparative Taxonomics of Foveal and Peripheral Cells in Primate Retina. *bioRxiv*. September 2018:428110. doi:10.1101/428110

68. Philippidou P, Dasen JS. Hox Genes: Choreographers in Neural Development, Architects of Circuit Organization. *Neuron*. 2013;80(1). doi:10.1016/j.neuron.2013.09.020

69. Hayashi S, Takeichi M. Emerging roles of protocadherins: From self□avoidance to enhancement of motility. *J Cell Sci*. 2015;128(8):1455-1464. doi:10.1242/jcs.166306

70. Supek F, Bošnjak M, Škunca N, Šmuc T. REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms. *PLOS ONE*. 2011;6(7):e21800. doi:10.1371/journal.pone.0021800