Identifying core biological processes distinguishing human eye tissues with systems-level gene expression analyses and weighted correlation networks

Bryan, John M†
Hufnagel, Robert B†
Brooks, Brian P†
McGaughey, David M†*

* Corresponding author
† Ophthalmic Genetics and Visual Function Branch, National Eye Institute, National Institutes of Health, United States of America

Abstract

The human eye has several specialized tissues which direct, capture, and pre-process information to provide vision. RNA-seq gene expression analyses have been used extensively, for example, to profile specific eye tissues and in large consortium studies, like the GTEx project, to study tissue-specific gene expression patterning. However, there has not been an integrated study of multiple eye tissues expression patterning with other human body tissues. We have collated current publicly available healthy human RNA-seq datasets and a substantial subset of the GTEx project RNA-seq datasets and processed all in a consistent bioinformatic workflow. We use this fully integrated dataset to probe the relatedness and biological processes between the cornea, retina, RPE-choroid complex, and the rest of the human tissues with differential expression, clustering, and GO term enrichment tools. We also leverage our large collection of retina and RPE-choroid tissues to build the first human weighted gene correlation networks and use them to highlight known biological pathways and eye gene disease enrichment. Finally, we make these data, analyses, and visualizations available via a powerful interactive web application (https://eyeintegration.nei.nih.gov/).

Introduction

The human eye is a highly specialized organ using several distinct tissues to focus and capture light and begin processing it into visual information. Light passes through the cornea and the lens which focus the light onto the retina (1). The rod and cone photoreceptors of the retina capture the light and transmits visual information through the optic nerve to the brain (2). The retinal pigment epithelium (RPE) is responsible for absorbing scattered light and providing nutrition, maintaining ionic homeostasis, and waste product processing for the photoreceptors, as well as mediating immune function for the retina and eye (3). The RPE and outer neural retina is supported and connected to the vascular system of the body via the choroid (4).

Many genetic disorders affect the function of the various eye tissues and cause vision perturbation or loss. The genetics of eye diseases range from monogenic Mendelian disorders to complex multi-gene system perturbations that are modified by environmental influences. While at least 316 identified genes underlying retinal disease have been identified, recent comprehensive next generation sequencing studies fail to find the cause of a variety of inherited retinal diseases like cone-rod dystrophies or retinitis pigmentosa 40-60% of the time (5–7). In an example of complex disease, age-related macular degeneration (AMD), which is believed to be caused by dysfunction of the RPE and choroid, genome-wide association studies (GWAS) have identified dozens of genomic locations associated with the disease. Still it is very difficult to pinpoint the causative gene or genes (8).

An important tool in understanding basic biology and unravelling the causes of disease has been the analysis of gene expression profiles. The Genotype-Tissue Expression (GTEx) Project has compiled nearly 10,000 individual tissue human RNA-seq samples and shared the data via a powerful and easy-to-use web portal (9). GTEx data has been used to help filter variants in GWAS studies, to build networks to identify candidate testis cancer genes, to help identify pathogenic mutations in an epilepsy cohort, and to identify a genetic variant linking folate homeostasis to warfarin response (10–14). Notably, the eye was not included as a tissue for this project. Because the vision community has been adopting RNA-seq for profiling different components of the eye, there is a large and growing set of useful transcriptome data. However, each study uses different bioinformatic processes to analyze their transcriptomes and the full genome-wide expression values are difficult to obtain, analyze, and visualize across studies. Therefore, utility of these resources ought to be optimized to similar effect as for other tissues.

We have collated all publicly available human eye tissue RNA-seq data and processed it with a robust and consistent bioinformatics process. We also have brought in a substantial portion of the GTEx project RNA-seq data to provide a comparison set to the eye tissues. Our full data-set holds 1027 samples. This comprehensive and consistently processed pan-eye and human data set allows for several novel analyses: first, to probe the relationships within cornea, retina, and RPE tissues and between eye tissues and other human tissues; second, to look for overarching patterns in gene expression and shared biology in differentially expressed genes between the eye tissues; and, finally, we use the large collated retina and RPE samples to build gene correlation networks for both. To maximize utility of this project to all researchers, we have also created a web application that allows quick and powerful access to the expression profiles of nearly 20,000 genes across 177 human eye tissue RNA-seq sets and 853 GTEx tissue RNA-seq sets and the two gene networks (https://eyeIntegration/nei.nih.gov/).

Results

Hundreds of individual human eye tissue RNA-seq datasets publicly available across twenty-one research studies

To identify all publicly available human eye tissue RNA-seq datasets, the Sequence Read Archive (SRA) was queried on January 19th 2017 with the R package SRAdb for human transcriptomic studies with the keywords 'RPE', 'macula', 'fovea', 'retina', 'choroid', 'sclera', 'iris', 'lens', 'cornea', and 'eye' across numerous fields in the SRA (15). This inclusive search identified 603 samples across 53 studies. Hand searching the studies to identify human eye tissue samples that did not have chemical, pharmacological, or genetic modifications or known eye-disease pared the initial search down to 219 samples across 21 studies (Supplementary Material Table S1, Fig. 1A) (16–32). The metadata of the remaining eye samples was queried and parsed to label each sample by a tissue (cornea, retina, RPE) and origin (immortalized cell line, stem cell line, fetal tissue, adult tissue) (Fig. 1B). Before gene expression quantification and quality control to remove lower quality samples we had 110 retina, 85 RPE, 28 cornea, 16 human embryonic stem cell lines (ESC), 6 lens, and 4 eyelid tissue RNA-seq data sets.

Efficient quantification tools allow for comparison of the eye transcriptome meta-set with dozens of other human tissues

The raw sequence data was obtained from the SRA or European Nucleotide Archive (ENA) and the transcript counts were quantified with the Salmon pseudo-alignment transcript quantification (33). To improve reliability of quantification, the transcript level counts were merged to the gene level (34). We then applied quantile normalization of the TPM (transcripts per million) values on a per-tissue basis with the qsmooth tool to reduce variability between different studies (35). Outliers with extremely low median gene counts and individual samples that clustered very far apart from similar samples were removed, leaving 171 eye samples (Fig. 1A, Supplementary Material, Tables S2 and S3). Voom normalization was then applied to adjust for different library sequencing depths (36). See methods for further details.

This efficient bioinformatic process also enabled us to bring in 878 samples from the GTEx project to compare to our eye meta-set (9). We selected, when possible, 10 male and 10 female non-gender specific tissues from the GTEx, ending up with 22 tissues, including blood, brain, heart, kidney, liver, lung, and thyroid (Supplementary Material, Tables S1 and S3). All raw data from the collated eye tissues or GTEx were processed identically with the above workflow. After outlier removal, using the same workflow as the eye tissue set above, we have 853 GTEx samples across 22 tissues.

Eye tissues from disparate studies cluster according to labelled eye component and tissue or cell-line origin

Our first question was whether the collated eye tissues, which potentially have significant batch effects from merging data from disparate sources, would group together using dimensionality reduction approaches. We used the Barnes-Hut implementation of the t-Distributed Stochastic Neighbor Embedding (t-SNE), which has been shown to work well in single-cell RNA-seq study analyses as well as the GTEx study set, to visualize relationships in two dimensions between the processed eye tissues (Fig. 2A) (37–39). The DBSCAN algorithm was used on the t-SNE coordinates for each sample to identify nine distinct clusters (40).

The adult tissue retina samples clustered together, though apart from their fetal or cell line based samples. The ESC retina samples have a variety of time points (37, 47, 67, 90 days) during their differentiation; we found no clustering by those criteria (data not shown) (20). The fetal and adult cornea samples, grouped closely together, but still clustered independently (Fig. 2A, clusters 8 and 9). Human embryonic stem cells (ESC), included because they are used across several studies to differentiate into different eye tissues, clustered together, generally closer to the cell-line derived samples (Fig. 2A, cluster 5).

RPE is the only tissue with more than three different sources: fetal tissue, adult tissue immortalized cell-line, and cells differentiated from ESCs. It should be noted that the adult RPE tissues are a mixture of RPE and choroid tissue, which is a vascular layer of the eye, providing oxygen and nutrients to the RPE and outer retina. This tissue will be referred to as adult RPE/choroid. The four sources cluster into three groups, with the few RPE fetal tissues clustering with the ESC-derived RPE samples (Fig 2A, cluster 1). The RPE derived from ESC group (Fig 2A, cluster 1) is composed of samples from three studies (16, 23, 28). All three groups differentiated their RPE cells for about two to four months, according to their method section. Wu and Zeng et al. gave specific times of differentiation for the exact tissues used in the SRA metadata (40 or 100 days); we did not see any differences in clustering patterns based on length of differentiation (data not shown) (28). This close grouping of fetal and ESC RPE tissues are consistent across multiple runs of t-SNE with different perplexity parameters ranging from 35-50 (data not shown). The adult RPE/choroid tissue clusters further away from the cell-line based tissues.

Overall, the t-SNE dimensionality reduction demonstrates that the eye tissues consistently cluster in unique groups by their tissue and origin. This happens despite a variety of laboratory origins with disparate culturing conditions, tissue handling, RNA extraction, sequencing cores, and so on.

Eye tissues distinct from most human tissues

To explore the relationship of eye tissues to other tissues in the human body, we leveraged the GTEx data we reprocessed to create a pan-human two-dimensional tissue relationship map with t-SNE (Fig. 2B). DBSCAN was then used, as before, to identify clusters. 'Tissue' labels from GTEx metadata in the SRA were used with one exception; fibroblasts are labelled separately from 'Skin' as they consistently group independently of skin-punch tissues. From the t-SNE visualization (Fig. 2B) we observe most human tissues group close to each other, with the exception of brain. The eye tissues, except retina, group closer to the non-brain human tissues. While the cell-line versus tissue derived eye tissue distinctions are maintained with the pan-human set, the eye-tissues are generally more related to each other than non-eye tissues.

The t-SNE 1 and 2 dimension coordinates generated by t-SNE are sensitive to the parameter perplexity, which controls the weighing of local to global relationships (41). Figures 2A and 2B used perplexities of 35 and 45, respectively. To more consistently demonstrate the pair-wise relationships between the tissues, the t-SNE dimensions were iteratively generated with perplexities from 35 to 50. Then means were taken, grouped by sample. The individual samples were then grouped by labelled tissue type and the t-SNE coordinates were again averaged.

Hierarchical clustering by Euclidean distance was done to group the tissues and a heatmap was generated (Fig. 2C) which displays the most closely related tissues. Because the hierarchical distances between cell-line derived eye tissues were inconsistent, they were removed from this analysis. We see that retina and brain tissues are individual outliers. We also see that the pituitary is grouped near RPE tissue and that fibroblasts group closely with the cornea (as denoted by the height of the dendrogram).

Differential expression analysis identifies large sets of genes distinguishing separating eye tissues

The eye tissue set collected can be separated on two major axes: tissue type (cornea, retina, or RPE) and origin (immortalized cell line, stem cell line, fetal tissue, adult tissue). Labelling each set of tissues by these two criteria gives us ten sets of eye tissues (Fig. 1B). To compare expression against non-eye tissue, we created a synthetic human 'body' expression set, by evenly combining the 22 GTEx tissues. The total number of body samples was matched to the total number of eye tissues we have by taking a random set of 8 tissues from each human body tissue category (e.g. Brain, Pituitary). There are 55 two-way combinations possible among the 11 sets.

To calculate differential expression, we modeled expression with the limma linear fit function with voom to correct for library size differences. The limma empirical Bayes function was used to identify statistically significant differentially expressed genes (36, 42). To look for global changes between the eye tissues and the body, we will first compare all of the eye tissue groups individually against the synthetic body (Table 1). A second synthetic body set was created by sampling the un-used GTEx tissues from the first synthetic body set and we found very similar differential expression values (data not shown).

The differentially expressed genes identified for each test (Table 1) was filtered to retain only genes with log fold change (logFC) < -2 or > 2 relative to the baseline tissue and with a false discovery rate (FDR) corrected p-value less than 0.01. A logFC of more than two means that the detected transcript level is more than four times as much (or one quarter as much) compared to the body tissue.

Biological term enrichment identifies eye-specific gene expression biology relating to visual function and body-specific gene expression relating to immunity and cell adhesion

As we have hundreds to thousands of genes meeting these stringent differential expression criteria across the ten comparisons we did Gene Ontology (GO) biological process term enrichment to identify systems-level patterns. We did the GO term enrichment independently on the over- and under-expressed gene sets, relative to the synthetic body set; 20 tests were performed. Overall, we found 2796 unique GO term IDs across the tests with a FDR corrected p value under 0.01 (Supplementary Materials, Table S4).

We took the top forty GO term IDs from the over and under-expressed tests (ranked by p value) and plotted them in a heatmap to identify shared GO terms among the different comparisons and to find overall trends in eye tissues gene expression relative to the synthetic body gene expression set (Figure 3). Clustering was done on both rows and columns to group together shared patterns. Like the t-SNE based clustering, the retina is an outlier for GO term enrichment. The GO terms in the first 20 rows (Fig. 3, Block 1) is driven by genes that are more highly expressed in the retina relative to other tissues. These over-expressed genes are highly enriched in GO terms relative to visual perception, light stimulus, synaptic signaling, and neurogenesis.

The next group (Fig. 3, Block 2) of enriched GO terms most strongly defines the ESC and the cornea and RPE immortalized cell lines and to a lesser extent, fetal cornea and stem cell retina tissue. These GO terms relate to cell cycle and division as well as DNA packaging and conformation. The last block (Fig 3., Block 3) is a set of GO IDs related to the body gene expression being higher than most of the eye tissues. This large block has GO terms involving migration, organismal process, adhesion, immune process, and stimulus. The full set of significantly ($p < 0.01$) enriched GO terms (2796) is available in Supplementary Materials Table S4.

Within eye tissue differential expression comparisons identify cornea, retina, RPE, and RPE/choroid gene sets

To more directly identify sets of genes enriched in particular eye tissue(s) relative to the remaining eye tissues, we compared all eye tissue differential expression pair-wise against each other and the synthetic body set (55 tests). To identify common gene sets, we used *k*-means clustering to group all genes into twenty groups; each group has a different overall gene expression pattern. We then plotted the relative gene expression for each eye tissue across the twenty *k*-means groups (Supplementary Materials, Figure S1). This produces a heatmap which identifies sets of genes that are more highly (or lowly) expressed in particular eye tissue(s) relative to the other eye tissues. We use this heatmap to identify genes defining the cornea, retina, RPE, and adult RPE/choroid and did GO term enrichment on these clusters (Table 2, Supplementary Table S5). The gene lists for each of the 20 groups are available in Supplementary File S1).

The cornea is enriched for genes involved in the extracellular matrix and collagen relative to the other eye tissues

In the GO heatmap (Fig. 3) the cornea tissues (immortalized cell line, fetal, adult) lack a highly distinguishing set of GO terms from the other eye tissues. However, there is a cluster (Supplementary Materials, Figure S1, cluster 3), with enriched fetal and adult cornea expression compared to the other tissues. This cluster contains 157 genes and top GO terms enriched for this set relate to extracellular matrix organization, collagen metabolism, and developmental processes (Table 2, Supplementary Table S5).

Adult retina and, to a lesser extent, retina stem cells enriched in visual function genes

Compared to the synthetic body set, the adult retina has many GO terms relating to visual function (Fig. 3, Block 1). This same GO enrichment is seen even when comparing adult retina against the other eye tissues, focusing on cluster 8 (Supplementary Materials, Figure S1, Table 2). This cluster is very highly expressed in adult retina and somewhat highly expressed in stem cell retina, relative to the other eye tissues.

RPE, excluding hTERT RPE, is highly enriched in genes relating to pigmentation and visual perception,

Like cornea, the non-immortalized RPE tissues do not have a distinct block of GO terms (Fig. 3). In the *k*-means heatmap (Supplementary Materials, Figure S1) we see that cluster 14 is more highly expressed in stem cell RPE, fetal RPE, and adult RPE/choroid. The hTERT immortalized cell line RPE is not highly expressed for this gene set. The 92 genes in this cluster are enriched in GO terms for visual perception, melanin processing, and vitamin A metabolism (Table 2).

Compared to other eye tissues, adult RPE/choroid is enriched for genes involved in immune function and adhesion

The cluster with genes highly expressed in adult RPE/choroid compared to the other eye tissues (number 10), has 229 genes. As this cluster is not highly expressed in the other RPE tissues, this cluster may define the choroid. These genes are strongly enriched in immune function and adhesion (Table 2).

hTERT RPE immortalized cell line has substantial gene expression differences relative to RPE derived from ESCs

As we had seen that the hTERT RPE clusters apart from the other RPE tissues, and there is a benefit to examining the differences between an immortalized RPE cell line model versus a differentiated RPE cell line model, we looked directly at differences in expression between hTERT RPE and stem cell RPE. We identified what genes and GO terms make these two cell lines different. There are over 1323 genes with a more than four-fold expression difference between RPE derived from human ESCs and the ATCC hTERT RPE immortalized cell line and 1572 with four-fold lower expression (Supplementary Materials Table S6). The five genes most highly expressed in RPE derived from human ESCs relative to the ATCC hTERT RPE immortalized cell line are *TTR* (Transthyretin), *DCT* (Dopachrome Tautomerase), *KIF1A* (Kinesin Family Member 1A), *SFRP5* (Secreted Frizzled Related Protein 5), and *NELL2* (Neural EGFL Like 2). GO terms associated with higher stem cell RPE expression relate to ion transport and synaptic transmission, suggesting that stem cell derived RPE is a more faithful model to human biology (Fig. 4).

Highly connected genes in retina and RPE gene networks recapitulate known eye biology

To this point, we have used the full gene expression set to independently cluster samples by tissue type and origin. We then used differential expression between the eye tissues and the synthetic body set to highlight differences in GO terms. We delved further by clustering the differential expression patterns between the eye tissues to find how each eye tissue is different from each other. We can go even further, by examining the relationships of the genes to each other, within a tissue, by using gene correlation networks. These network use correlated fluctuations of all-by-all pairwise gene expression similarities to build networks of gene-to-gene relationships.

As we had collected a substantial amount of retina and RPE samples, we were able to build weighted gene correlation networks with the Weighted Gene Co-Expression Network Analysis (WGCNA) R tool (43). We also attempted to build a cornea network, but the network construction failed due to failure to both differentiate the genes cleanly into defined modules and achieve appropriate network topology within a reasonable parameter space; more cornea samples are needed (Supplementary Materials, Figure S2). The gene expression TPM values, with the full set of corrections described earlier for the differential expression analyses, were used as inputs. All retina and all RPE tissues that passed quality control steps were used to build independent retina and RPE networks. The parameters used in the WGCNA network construction are enumerated in the methods.

There are 11101 and 10843 genes in the retina and RPE networks, respectively. 9621 of the genes are shared between the retina and RPE network. The kWithin metric from WGCNA measures the intramodular connectivity. Genes with higher connectivity are, theoretically, more likely to be important in gene regulation as perturbations in them will affect the system more than less connected genes. To get a sense of what the biology was of the most connected genes in the retina network, we took the 1017 genes with a kWithin greater than 20 and did GO enrichment (Supplementary Materials, Table S7), finding the top five GO terms all relate to visual perception. We did the same with the RPE network, using the 566 genes with a kWithin greater than 20. The top five GO terms in this RPE network connected list were related to endoplasmic reticulum function (Supplementary Materials, Table S7). The most similar modules, calculated by doing hypergeometric testing of GO terms and gene names, between the retina and RPE networks are the light cyan retina module and the pink RPE module. Both of these modules, by GO term enrichment, are involved in protein targeting to the ER (Supplementary Materials, Figure S3).

Retina network module highly enriched in genes implicated in eye disease and crucial for visual function

A key advantage of WGCNA networks over correlation networks is that genes can be partitioned into modules, presumably with shared biological function within each individual module. The retina network has 27 modules, with 64 to 1922 genes in each module. The RPE network has 23 modules, with 90 to 1458 genes in each module (Supplementary Materials, Figure S4). To determine whether the modules were enriched for known gene to gene interactions, we loaded each network module gene list into STRING and calculated whether there were more interactions than expected. For 23/27 retina modules and 20/23 RPE modules, the STRING p value for interaction enrichment was < 0.01 (Supplementary Materials, Table S8). We also ran GO term enrichment for each module within each network (Supplementary Materials, Table S11 and S14). While many modules have highly significant GO term enrichment, only the ‘green’ module is highly enriched for visual perception terms. Pinelli et al. built an unweighted retina gene correlation network and identified 14 candidate photoreceptor genes based upon their network (17). All 14 are in our retina network and 9 of the 14 are in our green visual function module ($p < 2.8 \times 10^{-10}$) (Supplementary Materials, Table S9).

There are 617 genes within the green retina module and 178 of these have a kWithin greater than 20. Many of the top connected genes have known visual function or are implicated in retinal diseases. To demonstrate the strong enrichment of known eye function genes in this module we divided the genes in the green module into four categories: known to play a role in eye disease, having GO terms relating to visual function, both, or neither (Fig. 5, Supplementary Materials, Table S10). From RetNet (http://www.sph.uth.tmc.edu/RetNet/) we have a list of 331 genes that have been implicated in retinal diseases (5). There are 178 genes with kWithin > 20 in the green module; 14 of those genes are also in RetNet, 17 have a vision GO term, 31 have both, and the remaining 116 genes are neither in RetNet nor have a vision-related GO term.

The human phenotype ontology (HPO) project is conceptually similar to gene ontology, except that they map abnormal human phenotype terms onto a graph and match them to genes (44). This provides a way to identify enrichment of abnormal human phenotypes. As there is no functioning package in R to systematically calculate HPO enrichment, we did bootstrapping and hypergeometric testing (see methods), looking for enrichment overall at the module level and for individual HPO terms within each module, respectively. The green module is highly enriched for HPO terms relating to eye disease, with terms like nyctalopia, abnormal electroretinogram, photophobia, cone-rod dystrophies, and blindness among the top terms (Fig. 5, Supplementary Materials, Table S12).

Other highly significant GO terms in the remaining retina network modules also match known retina function. GO terms enriched relate to ion transport (greenyellow), developmental processes (darkorange, greenyellow, tan), mitochondrial function (midnight blue), and metabolism (turquoise) (Supplementary Materials, Table S11). The retina network darkgrey module also contains several genes implicated in retina diseases like *ELOVL4*, *OPN1SW*, *SLC24A1*, and *PDE6A* (see Supplementary Materials, Table S13 for full list). Additionally, the green, tan, brown, and blue modules are, overall, enriched for HPO disease terms (Supplementary Materials, Figure S5).

Retina green module identifies visual transduction pathway and core upstream regulators

The green module was further analyzed for known biological networks components, which were generated through the use of Ingenuity Pathways Analysis (Ingenuity® Systems, www.ingenuity.com). Visual transduction was the most significant pathway present, with 16 components present in the green module. These components function predominantly in rod and cone photoreceptors in the conversion of photic energy to neural signaling in the retina (Supplementary Materials, Figure S6A and data not shown). Regulatory component analysis projected that *CRX* and *NRL* were predicted among the regulators of gene expression in the green module, upstream of several genes implicated in retinal photoreceptor degeneration also present in the green module (Supplementary Materials, Figure S6B). These two transcription factors drive rod photoreceptor differentiation and maintenance beginning in embryogenesis, and dysfunction of either of these is associated with retinal degeneration (45). In sum, the green module is enriched for photoreceptor function and recapitulates specific components of known biological and gene regulatory networks that are important causes of retinal disease.

RPE/choroid network contains many modules related to cell metabolism

Unlike the retina network, there are no strongly associated GO terms relating to visual function. However, there are numerous modules with strongly significant GO terms relating to metabolic processes and active transcription and translation (blue, brown, dark turquoise, green, light cyan, light green, red, turquoise). One module (yellow) relates to catabolism, one to immune function (tan), one to the endoplasmic reticulum (ER) (pink), and two the mitochondria (dark green, dark yellow) (Supplementary Materials, Table S14). Among the top HPO terms across the RPE modules are ones relating to anemia (pink), optic disc pallor (green), and respiration (dark green). Overall, the green, midnightblue, turquoise, lightyellow, magenta, and brown RPE modules are enriched for HPO terms (Supplementary Materials, Figure S5).

Retina and RPE networks in retinal diseases and AMD

Higher connected genes are theoretically more important in the function of the retina and RPE. From RetNet we have 331 genes that are associated with retinal diseases (though some unknown proportion affect the retina via the RPE). From a recent large AMD GWAS study, there is a list of 33 loci strongly associated with AMD, and thus likely related to RPE or choroid dysfunction (8). To see whether these retina or RPE gene lists have higher connectivity relative to the other genes in the networks we used density plots of the kWithin value to see whether we see any left-ward (less connectivity) or right-ward (more connectivity) shifts in our gene list kWithin connectivity.

We see that the RetNet gene list has a higher connectivity than non-RetNet genes in the retina module; this right-ward shift is highly significant ($p = 3.26 \times 10^{-8}$). The connectivity of the RetNet gene list in the RPE network is significantly different than the non-RetNet genes ($p = 0.28$). 53 RetNet genes are in the green retina module, which is a 4.1 fold enrichment over chance. The darkgrey module has a similar enrichment in RetNet genes with 10, which is a 3.8 fold enrichment over chance.

The 33 genes associated with AMD have a higher connectivity the remaining genes in the RPE network ; this right-ward shift is also significant ($p = 0.049$). Like the RetNet retinal disease gene list in the RPE network, the 33 AMD genes are not significantly more connected than the other genes in the retina network ($p = 0.49$) (Supplementary Materials, Figure S7).

Discussion

We collected all publicly available human eye RNA-seq datasets, the largest collection to date, and carefully performed a lengthy series of normalization and quality control procedures to robustly quantify gene expression within three major eye tissues and between the eye and other human tissues. We used the gene expression data to accurately cluster samples by tissue and origin and further demonstrated that only 75 genes can be effectively used to cluster. We used differential gene expression analysis with GO term enrichment to identify biological processes that best distinguish the eye tissues both from each other and from a synthetic human expression set. We then leveraged the large sets of retina and RPE tissues to build the first human weighted gene correlation networks for retina and RPE. We demonstrated the power of the networks to highlight genes known to be crucial in eye biology. Finally, we make the data and analyses available in a powerful web application (https://eyeIntegration.nei.nih.gov).

The structures of the eye are epithelial, neuroepithelial, and neural crest in origin. We were expecting some of the eye tissues to cluster closely with the skin, but instead we found that the retina was a very unique tissue, that transformed fibroblasts most closely matched the cornea, and the RPE was nearest the pituitary. Embryologic origins and specialized functions likely create these similarities and divisions, respectively. Cornea is derived from the same surface ectoderm as skin and from neural crest cells, while retina and RPE are derived from the neural tube epithelium from the ventral diencephalon, along with the hypothalamus and posterior pituitary. Corneal epithelium is replenished by limbal stem cells that remain into adulthood, which may explain the proximity of corneal and ESC clusters. That retina was separated from other ocular and non-ocular tissues likely related to the exclusivity and high expression burden of the visual transduction cycle in cone and rod photoreceptors.

The systems-level study of differential gene expression across cornea, retina, RPE, and RPE-choroid tissue highlights core functions of these tissues. Cornea-specific genes specify the structural aspect of the cornea with extracellular matrix organization and collagen metabolism and catabolism. The corneal epithelium is replenished continuously with limbal stem cells, which may be reflected in the enrichment of GO terms relating to development. The retinal tissues are strongly defined by genes involved in visual processes. The RPE and RPE-choroid tissues are also distinguished, with the former being more involved in visual processes and pigmentation while the latter is involved with immune system processes.

The creation of the first human retina and RPE weighted gene correlation networks has allowed us to identify dozens of modules with co-regulated genes. It is important to stress that these networks were built only with gene expression information and were optimized using network-specific metrics, such as how well the topological overlap matrix placed genes into well-defined modules. Only afterwards did we evaluate the significance of connected genes and modules to GO terms and known eye biology.

It is striking that the some of the most significant GO terms, by p-value and enrichment, in the retina network are associated with a single 617-gene module underlying visual function. This module represents the visual transduction pathway, which is relatively unique to the retina and is associated with isolated and nonsyndromic retinal degenerative conditions.

As the RPE has a high-energy role in transferring nutrients and clearing waste products for the photoreceptors of the retina, it is not surprising that a plurality of the modules are enriched for genes important in RNA translation, protein modification and production, catabolism, and mitochondrial function. The enrichment of highly connected AMD associated genes in the RPE network further emphasizes the value of this network.

Finally, the value of this extensive and carefully curated data-set is enhanced by the creation of the eyeIntegration web app (http://eyeIntegration.nei.nih.gov; Supplementary Materials, Figure S8). The site serves two roles, first as an interactive extension of this manuscript and second as a platform for researchers to identify interesting genes in eye function via searchable gene expression plots across many tissues, 55 pair-wise differential expression tests, and two gene networks. We also make the source code and accompanying data-sets fully and freely available for other researchers (see methods). The unravelling of eye biology and function has been furthered by genetic eye diseases, animal models, and functional assays. We hope that this open data sharing and powerful web application will provide a fourth way to decipher eye biology in health and disease.

Materials and Methods

Identification of normal human eye RNA-seq data-sets and tissue labelling

The entire SRA data was downloaded as a SQL file on January 19th, 2017 with the SRAdb R package. The following keywords were used in a partial-matching case-insensitive (e.g. 'retina' would match 'RETINAL') search: 'RPE', 'macula', 'fovea', 'retina', 'choroid', 'sclera', 'iris', 'lens', 'cornea', and 'eye.' These keywords were matched against the following fields in the SRA: 'study_abstract', 'experiment_name', 'study_name', 'sample_ID', 'sample_name', 'study_title', 'study_description' in human samples with a 'library_source' of 'transcriptomic' and filtering out miRNA studies. Study titles, abstracts, and other fields were checked by hand for inclusion in this study for whether they were genuine eye studies of normal (non-disease, non-mutated, no chemical modification) human eye tissue. The SRA metadata for the GTEx project was also pulled by searching for the study_accession 'SRP012682.' Our script enabling search of the SRA for eye tissues is provided as 'sraDB_search_select.R'

For reproducibility, the meta-data for each sample was parsed with our script 'parse_sample_attribute.R' to label the eye tissue (cornea, lens, eye-lid, retina, RPE, ESC) and its origin (immortalized cell-line, cell-line derived from ESC, fetal tissue, or adult tissue). This script has been written to handle the wide variety of metadata usage by the 21 research projects and the script must be edited to handle new eye samples. The GTEx tissue were labeled by tissue or sub tissue by parsing the GTEx SRA metadata for 'histological type' and 'body site', respectively with the 'parse_sample_attribute.R' script.

Efficient quantification of gene expression across 1027 samples

Two studies had their raw RNA-seq data accessioned with dbGaP (9, 19). We obtained access to these studies under dbGaP study #115588. Raw sequence data for these two studies were pulled and converted to fastq with the sratoolkit (2.8.0) fastq-dump tool. The remaining raw fastq data was pulled from NCBI via ftp, with the wget calls created by the script 'sra_to_fastq.R'. The one exception was the E-MTAB-4377 resource which was only available as bam files as of January 19th 2017 from European Bioinformatics Institutes ArrayExpress archive (17). The bam files were downloaded, then converted to fastq with the Picard SamToFastq (2.1.1) program (https://broadinstitute.github.io/picard/).

The raw fastq read files were loaded into salmon (0.7.2) with –seqBias and –gcBias flags against the Gencode Release 25 protein-coding transcript sequences fasta file to perform transcript-level quantification (33, 46). The Gencode gene names are used across this study. To improve specificity of the gene expression, transcripts with low abundance across all tissues were removed from the fasta file, and Salmon was re-run as per Soneson et. al (47). The filtered fasta file is provided in the source code as 'gencode.v25.pc_transcripts.commonTx.fa.gz.' and the Salmon script as 'run_salmon.sh.' To improve sensitivity and specificity, the transcript-level quantifications were merged to the gene-level and the length scaled transcripts per kilobase million (TPM) calculations were done with the R library tximport (1.2.0) (34) in our 'calculate_lengthScaledTPM.R' script.

Multi-step process to remove samples with low overall gene expression counts, quantile normalize samples by tissue, then cluster to identify outliers

A multi-stage process was then used on the full data set to remove outlier samples (either because of overall low gene expression levels or from clustering with the incorrect tissue group) and genes with zero to extremely low expression across the entire data-set were removed. While we found several mislabeled GTEx samples, this has been noticed before (39). Samples with a median TPM value < 50 were removed as these were outliers in terms of overall gene expression coverage. This step alone removed all of the lens samples, 20 RPE, 15 retina, and 16 ESC samples (Supplementary Materials, Table S2 and Figure S2). To alleviate potential batch effects between the samples from different studies, the TPM values were quantile normalized within tissues and globally simultaneously with the qsmooth algorithm (35) (Supplementary Materials, Figure S7).

Finally, the remaining samples were dimensionality reduced with t-SNE, then clustered with DBSCAN. The performance of t-SNE is sensitive to the perplexity parameter, which weighs local versus global relationships. We found for our study that perplexities ranging from 30-50 performed the most reliably (data not shown). For the all-sample t-SNE we used a perplexity of 45. For the eye-only sample t-SNE, we used a perplexity of 35. The t-SNE coordinates were clustered by DBSCAN with the eps parameter set to 1.3. The cluster assignments from DBSCAN were then aggregated to the tissue and origin level, to identify small numbers of samples that clustered with other tissues; these are likely sample swaps. These outliers were removed. The script for this process is 'outlier_identification.Rmd.'

Differential gene expression analysis with pair-wise testing

A synthetic pan-human gene expression set was created by randomly sampling 8 tissues from each of the 22 GTEx tissue samples. This was used with the nine different eye tissue-origin sample sets and the ESC set, totaling 11 different groups. All 55 pair-wise tests (11 choose 2 equals 55) were done with the limma package with voom library size normalization, using the quantile-normalized TPM values as the input (36, 42). The script 'differential_expression.Rmd' contains the code for these steps.

GO, HPO, and STRING enrichment

For GO enrichment, the biomaRt package was used, in R, to get the entrez IDs from the 'dec2016' 'hsapiens_gene_ensembl' mart. The GOstats package, in R, was used to calculate GO enrichment by the hypergeometric test, only keeping over-enriched terms. The background gene list across the different tests was defined as all genes in the original TPM expression matrix. The function for this analysis is provided as 'GO_enrichment.R.'

For HPO enrichment, no working R package was available. To identify modules that mapped to a higher than expected number of HPO terms we used bootstrapping, comparing the number of HPO terms mapped to a module (proportional to its size) against a bootstrap distribution of the same metric. To analyze overabundance of HPO terms in a module we used hypergeometric testing, comparing the number of HPO terms in a module against the background of all genes and their associated HPO terms. The 'ALL_SOURCES_FREQUENT_FEATURES_genes_to_phenotype.txt' file from 'Build #124' was downloaded on April 4th, 2017 from http://compbio.charite.de/jenkins/job/hpo.annotations.monthly/lastStableBuild/ . This file links gene names to HPO terms. The script that did the hypergeometric testing is provided as 'HPO_enrichment_function.R.'

STRING enrichment p-values were computed with the STRINGdb R package. We placed all genes in each module, up to 400 (the max input possible for STRINGdb). For modules with more than 400 genes (7 retina modules and 10 RPE modules), we used the 400 genes in the module with the highest kWithin connectivity. The script for this is 'stringDB.R.'

Tissue-level gene block analysis with KMeans clustering and gene ontology enrichment

The differential gene expression patterns across the 55 pair-wise tests were grouped into twenty clusters, each holding groups of genes with shared expression patterns. The grouping was done with the *k*-means algorithm, in R, with 10,000 iterations and the 'MacQueen' algorithm. The cluster assignments for each gene was joined with the eye-tissue TPM values for the gene. The TPM values were averaged for each eye tissue, then the overall gene expression in each cluster was averaged. The TPM values, averaged by tissue, then cluster, were plotted in a heatmap. The code for this analysis is in 'kmeans_de_cluster_heatmap.Rmd' and the cluster assignments for each gene are available as 'DE_Kmeans_cluster_Gene_Lists.zip'.

Gene network construction with WGCNA

Weighted co-expression networks were constructed separately on both retina and RPE samples using the Weighted Gene Co-Expression Network Analysis (WGCNA) framework with the corresponding *WGCNA* R package. TPM expression matrices were used for the construction of both networks. Genes with consistently low levels of expression (less than 30 TPM in at least 5% of samples for the retina network, less than 40 TPM in at least 5% of samples for the RPE network) were removed prior to network construction. We found that less stringent cut-offs for low expression resulted in poor clustering of these genes (data not shown).

Average-linkage hierarchical clustering and t-distributed Stochastic Neighbor Embedding (t-SNE) were used to assess batch issues stemming from sample origin and study source, using the *WGCNA* and *Rtsne* R packages, respectively. Following the observation of batch effects, the *ComBat* R package was used to correct for batch issues stemming from an interaction variable between sample origin and study source. Following batch correction, a $\log_2$-transformation was applied to each expression matrix. the following transformation was applied to each expression matrix:

$$f(lsTPM) = \log_2(lsTPM + 1)$$

WGCNA identifies co-expression patterns using a weighted correlation matrix. The un-weighted correlation matrix is raised to a soft-thresholding power (β) in order to satisfy the scale-free law (43). This means that $p(i)$, the probability that a node has degree i, follows a power law distribution $p(i) \sim i^{-n}$. In choosing β for each of the networks, it is suggested by the WGCNA developers to choose a β which produces a negative correlation between $\log(i)$ and $\log(p(i))$, with $R^2 > 0.8$. Using the *pickSoftThreshold* function in the *WGCNA* R package, a range of soft-thresholding powers (β) were evaluated for both networks. The suggested criteria were met with soft-thresholding powers of 4 and 7 for the retina and RPE networks, respectively. Each co-expression network was constructed in the following manner using the $\log_2$-transformed expression matrices:

1. Compute a Pearson correlation matrix of all gene pairs: $S = [S_{ij}]$, where $S_{ij} = |cor(i, j)|$, where i and j are distinct genes.
2. Compute an adjacency matrix as: $A = [a_{ij}]$, where $a_{ij} = power(S_{ij}, \beta) = [S_{ij}]^{\beta}$
3. Compute an unsigned topological overlap matrix (TOM) as: $TOM_{ij} = \frac{a_{ij}+\sum_{u\neq i,j} a_{iu}a_{uj}}{\min(k_i,k_j)+1-a_{ij}}$, where $k_i = \sum_{u\neq i} a_{iu}$, and $k_j = \sum_{u\neq j} a_{ju}$
4. Define a dissimilarity matrix as $d_{ij} = 1 - TOM_{ij}$. Use average-linkage hierarchical clustering on the dissimilarity matrix to cluster the genes.
5. Use the *cutreeDynamic* function to place genes into distinct modules. For this function, parameters of *deepSplit = 0* and *minClusterSize = 30* were used.

The script used to generate the networks is provided as 'WGCNA_networks.Rmd.'

Identifying Similar Modules Across Retina and RPE Networks

Similarities in module compositions between the retina and RPE networks were assessed. This was performed through pair-wise cross-network comparison of retina and RPE modules in terms of the genes that were assigned to each pair of modules, as well as the GO terms that were associated with the modules being compared. For each cross-network module comparison, the number of overlapping genes was calculated and subjected to a hypergeometric test to assess significance. This process was repeated with examining overlap in GO terms between modules. In both analyses, p-values were adjusted using the FDR correction method.

Web app, other tools, and source code

The fastq file transfer and salmon quantification were run in the bash environment. The salmon-based RNA-seq quantification and random forest calculations used the computational resources of the NIH HPC Biowulf cluster (http://hpc.nih.gov).

All other statistical analyses and visualization was done in the R environment (see 'session_info_R.txt' for packages used and versions). The heatmaps were made with the superheat package. All other figures were made with ggplot2.

The interactive web app was built with the R Shiny framework and hosted on a R Shiny Server (https://shiny.rstudio.com). Plotly (https://plot.ly) was used to turn ggplot images into interactive images. The visNetwork R package was used to visualize the network modules. For the purpose of limiting the number of edges to a number that would be tractable for interactive visualization, the network edges were filtered so that each node would have its k-nearest within-module genes (k-strongest edges to genes in the same module) remain in the network, for a range of k values.

The source code for the web page is available at https://gitlab.com/davemcg/Human_eyeIntegration_App. The scripts mentioned in the methods underlying the data processing and analysis for this paper are available as supplemental file scripts.zip and the data used in the scripts is available at Zenodo (10.5281/zenodo.569870).

Acknowledgements

This research was supported by the Intramural Research Program of the National Eye Institute, National Institutes of Health. This work utilized the computational resources of the NIH HPC Biowulf cluster (http://hpc.nih.gov). We would like to thank the 21 eye research projects whose public sharing of their raw sequencing data and metadata made this project possible.

References

1. Bourne,W.M. (2003) Biology of the corneal endothelium in health and disease. *Eye*, **17**, 912–918.

2. Wallace,V.A. (2011) Concise Review: Making a Retina—From the Building Blocks to Clinical Applications. *STEM CELLS*, **29**, 412–417.

3. Bharti,K., Nguyen,M.-T.T., Skuntz,S., Bertuzzi,S. and Arnheiter,H. (2006) The other pigment cell: specification and development of the pigmented epithelium of the vertebrate eye. *Pigment Cell Res. Spons. Eur. Soc. Pigment Cell Res. Int. Pigment Cell Soc.*, **19**, 380–394.

4. Hayreh,S.S. (1975) Segmental nature of the choroidal vasculature. *Br. J. Ophthalmol.*, **59**, 631–648.

5. Daiger,S., Rossiter,B., Greenberg,J., Christoffels,A. and Hide,W. (1998) Data services and software for identifying genes and mutations causing retinal degeneration. *Invest Ophthalmol Vis Sci*, **39**, S295.

6. Ellingford,J.M., Barton,S., Bhaskar,S., Williams,S.G., Sergouniotis,P.I., O'Sullivan,J., Lamb,J.A., Perveen,R., Hall,G., Newman,W.G., *et al.* (2016) Whole Genome Sequencing Increases Molecular Diagnostic Yield Compared with Current Diagnostic Testing for Inherited Retinal Disease. *Ophthalmology*, **123**, 1143–1150.

7. Carss,K.J., Arno,G., Erwood,M., Stephens,J., Sanchis-Juan,A., Hull,S., Megy,K., Grozeva,D., Dewhurst,E., Malka,S., *et al.* (2017) Comprehensive Rare Variant Analysis via Whole-Genome Sequencing to Determine the Molecular Pathology of Inherited Retinal Disease. *Am. J. Hum. Genet.*, **100**, 75–90.

8. Fritsche,L.G., Igl,W., Bailey,J.N.C., Grassmann,F., Sengupta,S., Bragg-Gresham,J.L., Burdon,K.P., Hebbring,S.J., Wen,C., Gorski,M., *et al.* (2016) A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants. *Nat. Genet.*, **48**, 134–143.

9. GTEx Consortium (2013) The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.*, **45**, 580–585.

10. Okbay,A., Beauchamp,J.P., Fontana,M.A., Lee,J.J., Pers,T.H., Rietveld,C.A., Turley,P., Chen,G.-B., Emilsson,V., Meddens,S.F.W., *et al.* (2016) Genome-wide association study identifies 74 loci associated with educational attainment. *Nature*, **533**, 539–542.

11. Murthy,M.N., Blauwendraat,C., UKBEC, Guelfi,S., IPDGC, Hardy,J., Lewis,P.A. and Trabzuni,D. (2017) Increased brain expression of GPNMB is associated with genome wide significant risk for Parkinson's disease on chromosome 7p15.3. *Neurogenetics*, 10.1007/s10048-017-0514-8.

12. Wang,C., Gu,Y., Zhang,K., Xie,K., Zhu,M., Dai,N., Jiang,Y., Guo,X., Liu,M., Dai,J., *et al.* (2016) Systematic identification of genes with a cancer-testis expression pattern in 19 cancer types. *Nat. Commun.*, **7**, 10499.

13. Muona,M., Berkovic,S.F., Dibbens,L.M., Oliver,K.L., Maljevic,S., Bayly,M.A., Joensuu,T., Canafoglia,L., Franceschetti,S., Michelucci,R., *et al.* (2015) A recurrent de novo mutation in KCNC1 causes progressive myoclonus epilepsy. *Nat. Genet.*, **47**, 39–46.

14. Daneshjou,R., Gamazon,E.R., Burkley,B., Cavallari,L.H., Johnson,J.A., Klein,T.E., Limdi,N., Hillenmeyer,S., Percha,B., Karczewski,K.J., *et al.* (2014) Genetic variant in folate homeostasis is associated with lower warfarin dose in African Americans. *Blood*, **124**, 2298–2305.

15. Zhu,Y., Stephens,R.M., Meltzer,P.S. and Davis,S.R. (2013) SRAdb: query and use public next-generation sequencing data from within R. *BMC Bioinformatics*, **14**, 19.

16. Peng,S., Gan,G., Qiu,C., Zhong,M., An,H., Adelman,R.A. and Rizzolo,L.J. (2013) Engineering a blood-retinal barrier with human embryonic stem cell-derived retinal pigment epithelium: transcriptome and functional analysis. *Stem Cells Transl. Med.*, **2**, 534–544.

17. Pinelli,M., Carissimo,A., Cutillo,L., Lai,C.-H., Mutarelli,M., Moretti,M.N., Singh,M.V., Karali,M., Carrella,D., Pizzo,M., *et al.* (2016) An atlas of gene expression and gene co-regulation in the human retina. *Nucleic Acids Res.*, **44**, 5773–5784.

18. Li,M., Jia,C., Kazmierkiewicz,K.L., Bowman,A.S., Tian,L., Liu,Y., Gupta,N.A., Gudiseva,H.V., Yee,S.S., Kim,M., *et al.* (2014) Comprehensive analysis of gene expression in human retina and supporting tissues. *Hum. Mol. Genet.*, **23**, 4001–4014.

19. Whitmore,S.S., Wagner,A.H., DeLuca,A.P., Drack,A.V., Stone,E.M., Tucker,B.A., Zeng,S., Braun,T.A., Mullins,R.F. and Scheetz,T.E. (2014) Transcriptomic analysis across nasal, temporal, and macular regions of human neural retina and RPE/choroid by RNA-Seq. *Exp. Eye Res.*, **129**, 93–106.

20. Kaewkhaw,R., Kaya,K.D., Brooks,M., Homma,K., Zou,J., Chaitankar,V., Rao,M. and Swaroop,A. (2015) 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 Dayt. Ohio*, **33**, 3504–3518.

21. Frausto,R.F., Le,D.J. and Aldave,A.J. (2016) Transcriptomic Analysis of Cultured Corneal Endothelial Cells as a Validation for Their Use in Cell Replacement Therapy. *Cell Transplant.*, **25**, 1159–1176.

22. Darrow,E.M., Huntley,M.H., Dudchenko,O., Stamenova,E.K., Durand,N.C., Sun,Z., Huang,S.-C., Sanborn,A.L., Machol,I., Shamim,M., *et al.* (2016) Deletion of DXZ4 on the human inactive X chromosome alters higher-order genome architecture. *Proc. Natl. Acad. Sci. U. S. A.*, **113**, E4504-4512.

23. Hu,G., Huang,K., Yu,J., Gopalakrishna-Pillai,S., Kong,J., Xu,H., Liu,Z., Zhang,K., Xu,J., Luo,Y., *et al.* (2012) Identification of miRNA Signatures during the Differentiation of hESCs into Retinal Pigment Epithelial Cells. *PLOS ONE*, **7**, e37224.

24. Radeke,M.J., Radeke,C.M., Shih,Y.-H., Hu,J., Bok,D., Johnson,L.V. and Coffey,P.J. (2015) Restoration of mesenchymal retinal pigmented epithelial cells by TGFβ pathway inhibitors: implications for age-related macular degeneration. *Genome Med.*, **7**, 58.

25. Ouyang,H., Xue,Y., Lin,Y., Zhang,X., Xi,L., Patel,S., Cai,H., Luo,J., Zhang,M., Zhang,M., *et al.* (2014) WNT7A and PAX6 define corneal epithelium homeostasis and pathogenesis. *Nature*, **511**, 358–361.

26. Gill,K.P., Hung,S.S.C., Sharov,A., Lo,C.Y., Needham,K., Lidgerwood,G.E., Jackson,S., Crombie,D.E., Nayagam,B.A., Cook,A.L., *et al.* (2016) Enriched retinal ganglion cells derived from human embryonic stem cells. *Sci. Rep.*, **6**, 30552.

27. Chen,Y., Huang,K., Nakatsu,M.N., Xue,Z., Deng,S.X. and Fan,G. (2013) Identification of novel molecular markers through transcriptomic analysis in human fetal and adult corneal endothelial cells. *Hum. Mol. Genet.*, **22**, 1271–1279.

28. Wu,W., Zeng,Y., Li,Z., Li,Q., Xu,H. and Yin,Z.Q. (2016) Features specific to retinal pigment epithelium cells derived from three-dimensional human embryonic stem cell cultures — a new donor for cell therapy. *Oncotarget*, **7**, 22819–22833.

29. Chng,Z., Peh,G.S.L., Herath,W.B., Cheng,T.Y.D., Ang,H.-P., Toh,K.-P., Robson,P., Mehta,J.S. and Colman,A. (2013) High Throughput Gene Expression Analysis Identifies Reliable Expression Markers of Human Corneal Endothelial Cells. *PLOS ONE*, **8**, e67546.

30. Farkas,M.H., Grant,G.R., White,J.A., Sousa,M.E., Consugar,M.B. and Pierce,E.A. (2013) 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*, **14**, 486.

31. Santaguida,S., Vasile,E., White,E. and Amon,A. (2015) Aneuploidy-induced cellular stresses limit autophagic degradation. *Genes Dev.*, 10.1101/gad.269118.115.

32. Tian,L., Kazmierkiewicz,K.L., Bowman,A.S., Li,M., Curcio,C.A. and Stambolian,D.E. (2015) Transcriptome of the human retina, retinal pigmented epithelium and choroid. *Genomics*, **105**, 253–264.

33. Patro,R., Duggal,G., Love,M.I., Irizarry,R.A. and Kingsford,C. (2017) Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods*, **advance online publication**.

34. Soneson,C., Love,M.I. and Robinson,M.D. (2016) Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research*, **4**, 1521.

35. Hicks,S.C., Okrah,K., Paulson,J.N., Quackenbush,J., Irizarry,R.A. and Bravo,H.C. (2016) Smooth Quantile Normalization. *bioRxiv*, 10.1101/085175.

36. Law,C.W., Chen,Y., Shi,W. and Smyth,G.K. (2014) voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.*, **15**, R29.

37. Maaten,L. van der and Hinton,G. (2008) Visualizing Data using t-SNE. *J. Mach. Learn. Res.*, **9**, 2579–2605.

38. Ntranos,V., Kamath,G.M., Zhang,J.M., Pachter,L. and Tse,D.N. (2016) Fast and accurate single-cell RNA-seq analysis by clustering of transcript-compatibility counts. *Genome Biol.*, **17**, 112.

39. Taskesen,E. and Reinders,M.J.T. (2016) 2D Representation of Transcriptomes by t-SNE Exposes Relatedness between Human Tissues. *PLOS ONE*, **11**, e0149853.

40. Ester,M., Kriegel,H., Sander,J. and Xu,X. (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. In. AAAI Press, pp. 226–231.

41. Wattenberg,M., Viégas,F. and Johnson,I. (2016) How to Use t-SNE Effectively. *Distill*, 10.23915/distill.00002.

42. 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 Res.*, **43**, e47.

43. Langfelder,P. and Horvath,S. (2008) WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, **9**, 559.

44. Köhler,S., Vasilevsky,N.A., Engelstad,M., Foster,E., McMurry,J., Aymé,S., Baynam,G., Bello,S.M., Boerkoel,C.F., Boycott,K.M., *et al.* (2017) The Human Phenotype Ontology in 2017. *Nucleic Acids Res.*, **45**, D865–D876.

45. Swaroop,A., Kim,D. and Forrest,D. (2010) Transcriptional regulation of photoreceptor development and homeostasis in the mammalian retina. *Nat. Rev. Neurosci.*, **11**, 563–576.

46. J,H., A,F., Jm,G., E,T., M,D., F,K., Bl,A., D,B., A,Z., S,S., *et al.* (2012) GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.*, **22**, 1760–1774.

47. Soneson,C., Matthes,K.L., Nowicka,M., Law,C.W. and Robinson,M.D. (2016) Isoform prefiltering improves performance of count-based methods for analysis of differential transcript usage. *Genome Biol.*, **17**, 12.

Legends to Figures

Figure 1
Identifying 177 unique human eye and ESC samples across 16 studies and four tissue types

A. Counts for unique cornea, ESC, retina, and RPE (choroid) human RNA-seq samples by study accession
B. Counts by tissue and origin. * is adult RPE – choroid

Figure 2
Gene expression information sufficient to both accurately cluster eye and GTEx tissue independently and demonstrates that eye tissues are generally more closely related to other than other body tissues

A. Dimensionality reduction by t-SNE of human eye tissues and ESC, colored by clustering assignment, and labelled by tissues in cluster. Shape of point corresponds to tissue origin
B. Eye tissues with GTEx tissues, colored by clustering assignment, labelled by tissues in cluster
C. Pair-wise euclidean distance between each tissue. Closer tissues have a smaller height in the dendrogram and are more yellow in color. More distant tissues have a larger height in the dendrogram and are more blue.

Figure 3
Major differences in systems relating to visual function, active cell division, adhesion, and immunity between the eye tissues and the other tissues in the human body

Top 80 GO Terms (40 with eye > body and 40 with body > eye) across eye-tissue to body differential expression tests. Yellow is more significant, blue is less. Hierarchical clustering of both rows and columns place more related GO terms and tissue comparison sets together.

Figure 4
Genes crucial in eye function are highly differentially expressed between stem cell derived RPE and hTERT RPE

A. The top 5 genes overexpressed in ESC-derived RPE and immortalized cell line hTERT RPE along
B. Word cloud of enriched GO terms

Figure 5
Retina network green module highly enriched for important visual function genes

A. Top 50 connected genes in green module in the retina network. Colored by group (see C.)
B. Word cloud of top GO terms in green retina module
C. kWithin connectivity (higher is more connected) for top 20 connected genes, labelled to indicate whether the gene is in RetNet, has a GO term relating to visual function, both, or none.

Legends to Tables

Table 1

	Log Fold Change	
Comparison	**< -2**	**> 2**
Cornea (adult) vs Body (adult)	1969	1249
Cornea (fetal) vs Body (adult)	172	873
Cornea (immortalized cell) vs Body (adult)	2611	1475
ESC (stem cell) vs Body (adult)	2738	2177
Retina (adult) vs Body (adult)	2607	1978
Retina (stem cell) vs Body (adult)	3443	2622
RPE - choroid (adult) vs Body (adult)	1200	1258
RPE (fetal) vs Body (adult)	1510	1402
RPE (immortalized cell) vs Body (adult)	2446	1398
RPE (stem cell) vs Body (adult)	2270	1308

Thousands of highly differentially expressed genes between the eye tissues and the synthetic body

Number of genes with logFC < -2 or > 2 (0.25 or 4 fold, p value < 0.01) between each eye tissue against the synthetic body set

Table 2

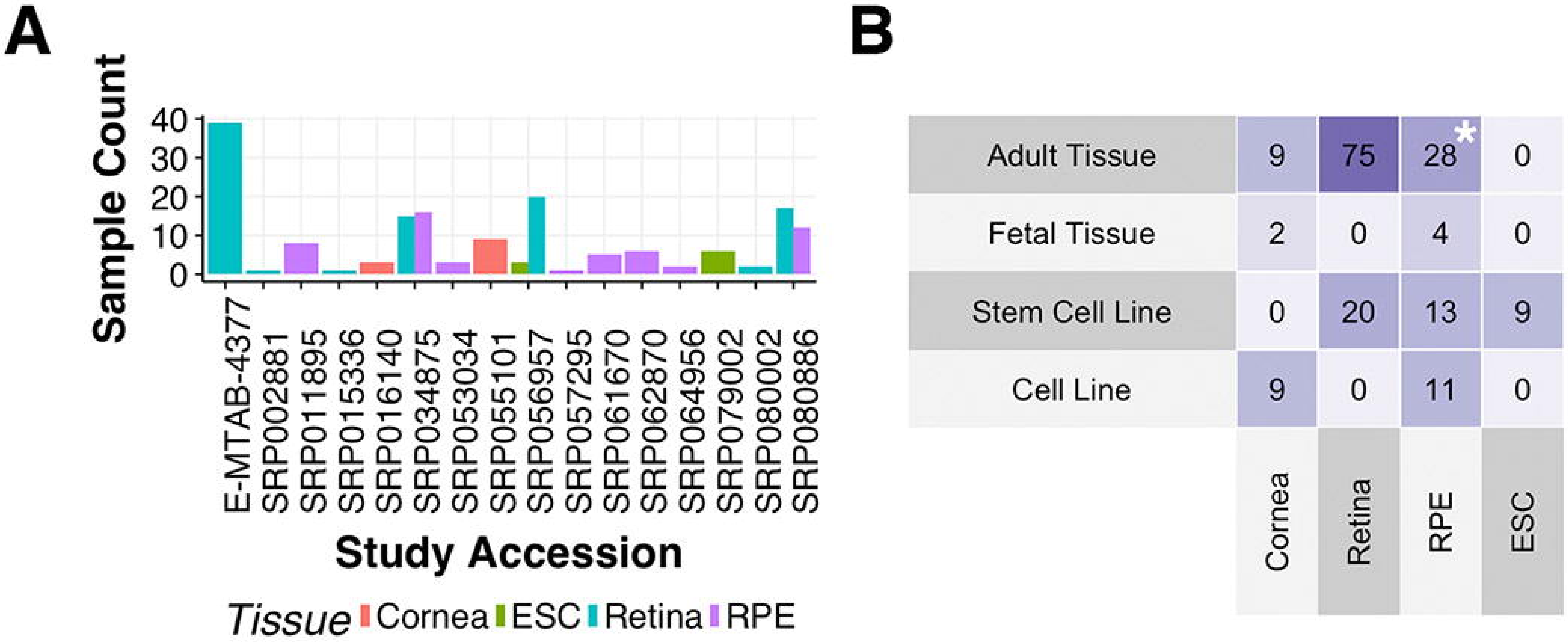
Tissue(s)-Specific Set	Cluster	GO BP ID	P value (FDR)	Odds Ratio	Term
Cornea	3	GO:0030198	1.16e-13	11.27	extracellular matrix organization
Cornea	3	GO:0030574	1.07e-11	30.51	collagen catabolic process
Cornea	3	GO:0032502	1.50e-09	3.20	developmental process
Adult Retina	8	GO:0050953	8.43e-65	38.80	sensory perception of light stimulus
Adult Retina	8	GO:0007601	1.50e-62	37.74	visual perception
Adult Retina	8	GO:0060041	3.87e-19	17.14	retina development in camera-type eye
non-immortalized RPE	14	GO:0007601	1.57e-10	18.39	visual perception
non-immortalized RPE	14	GO:0042438	1.70e-07	104.91	melanin biosynthetic process
non-immortalized RPE	14	GO:0006776	1.39e-05	163.77	vitamin A metabolic process
Adult RPE - choroid	10	GO:0002376	1.35e-20	5.06	immune system process
Adult RPE - choroid	10	GO:0006952	5.03e-18	5.39	defense response
Adult RPE - choroid	10	GO:0007155	1.64e-13	4.25	cell adhesion

Top GO terms for tissue-specific cluster groups relate to eye tissue specific function

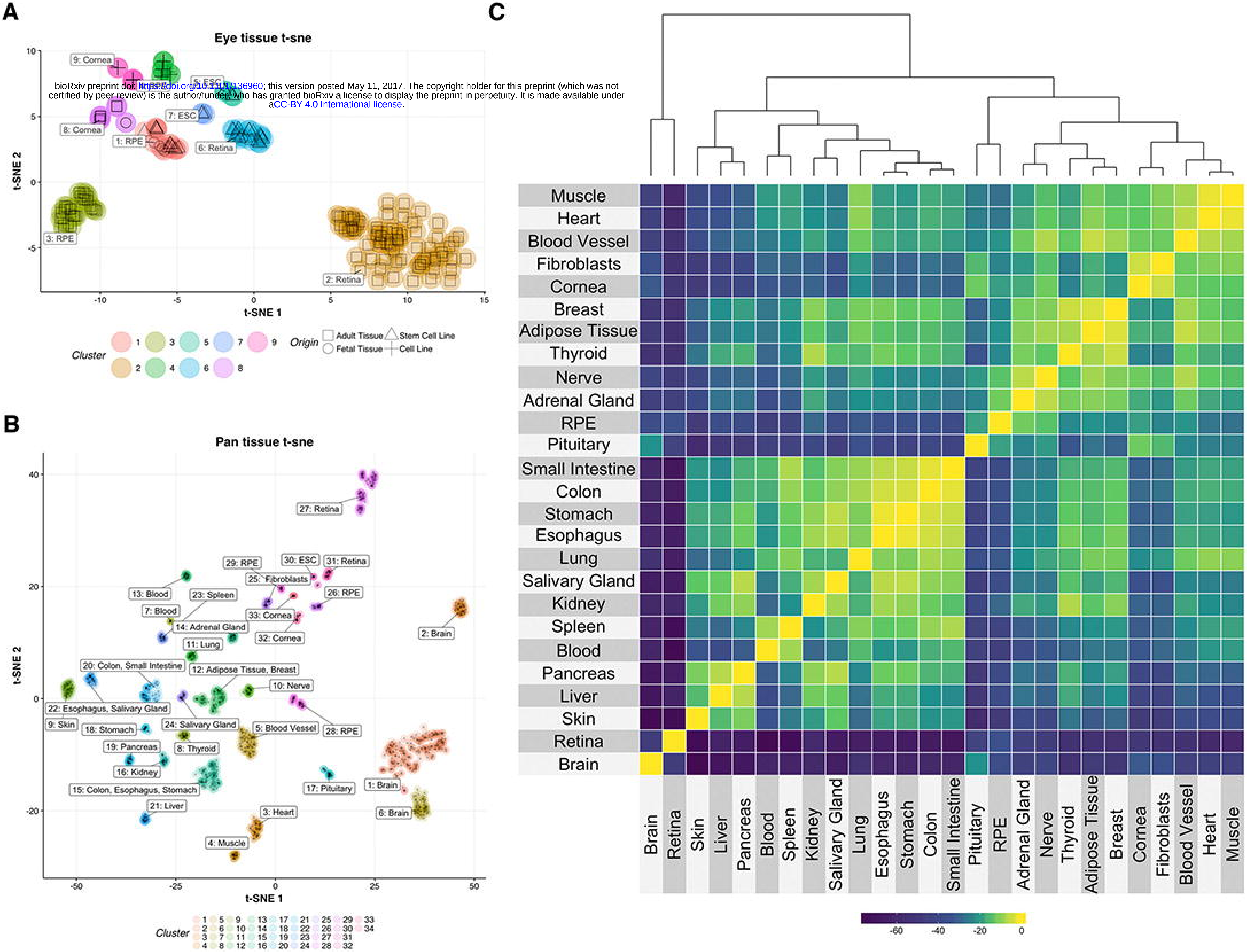
Three representative GO terms were selected for GO term enrichment done on the differentially expressed Kmeans cluster sets 3, 8, 10, and 14 (see Supplementary Materials, Figure S1) which represent over-expressed genes in cornea, adult retina, adult RPE/choroid, and non-immortalized RPE, respectively

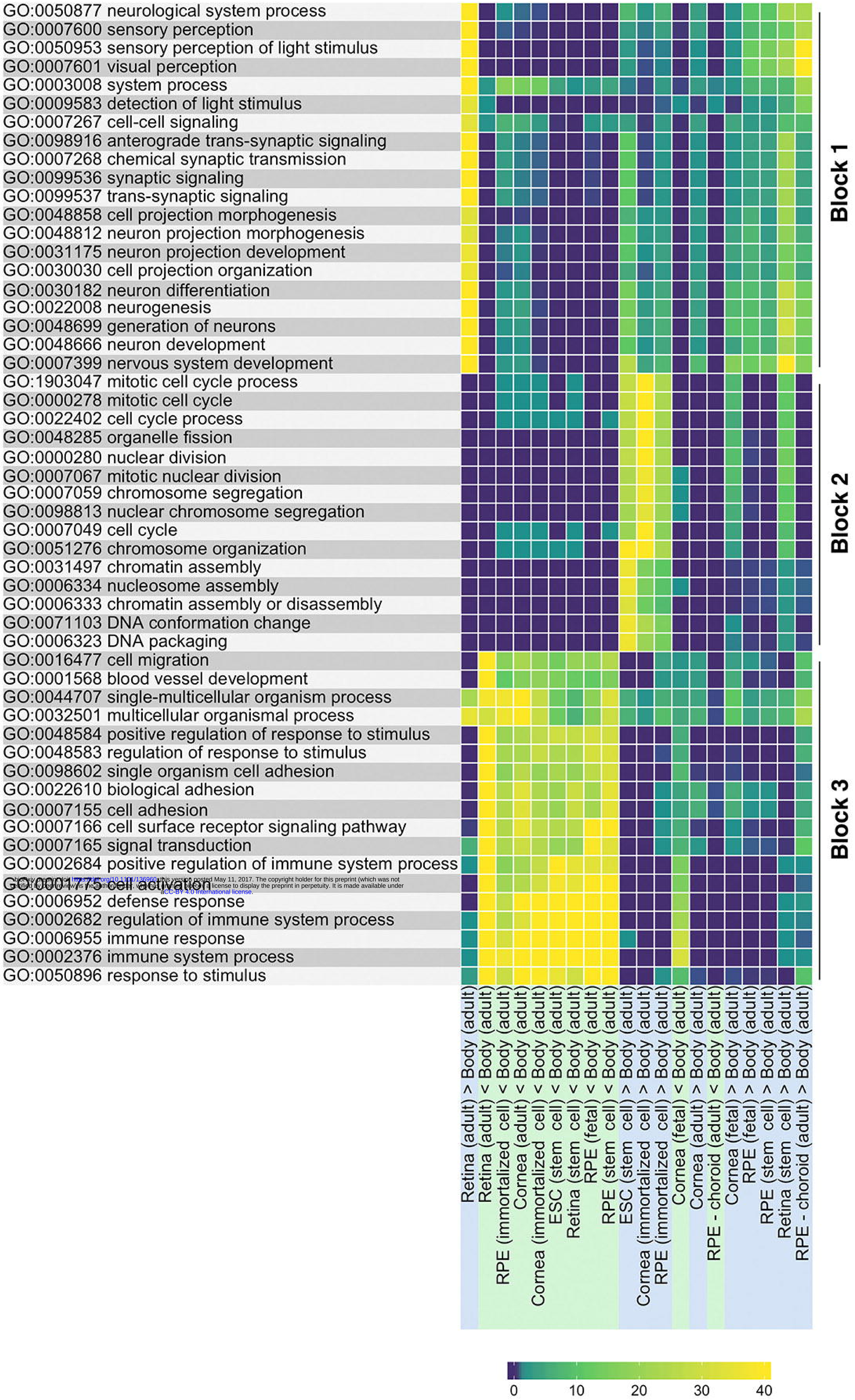
Abbreviations

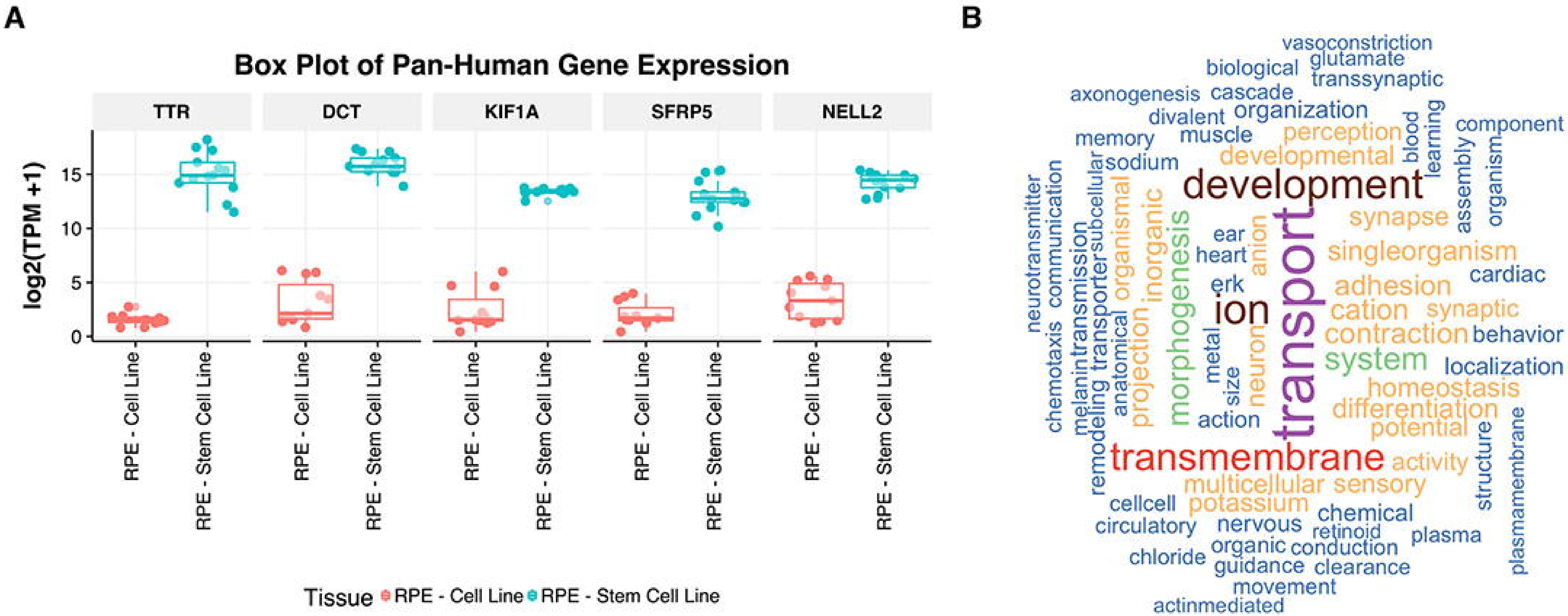
Age-related macular degeneration (AMD)
Human Embryonic Stem Cells (ESCs)
Gene Ontology (GO)
Gene Tissue Expression Project (GTEx)
Human Phenotype Ontology (HPO)
log Fold Change (logFC)
Retinal Pigment Epithelium (RPE)
Sequence Read Archive (SRA)
length-scaled Transcripts Per Million (TPM)
t-Distributed Stochastic Neighbor Embedding (t-SNE)
Weighted Gene Co-Expression Network Analysis (WGCNA)

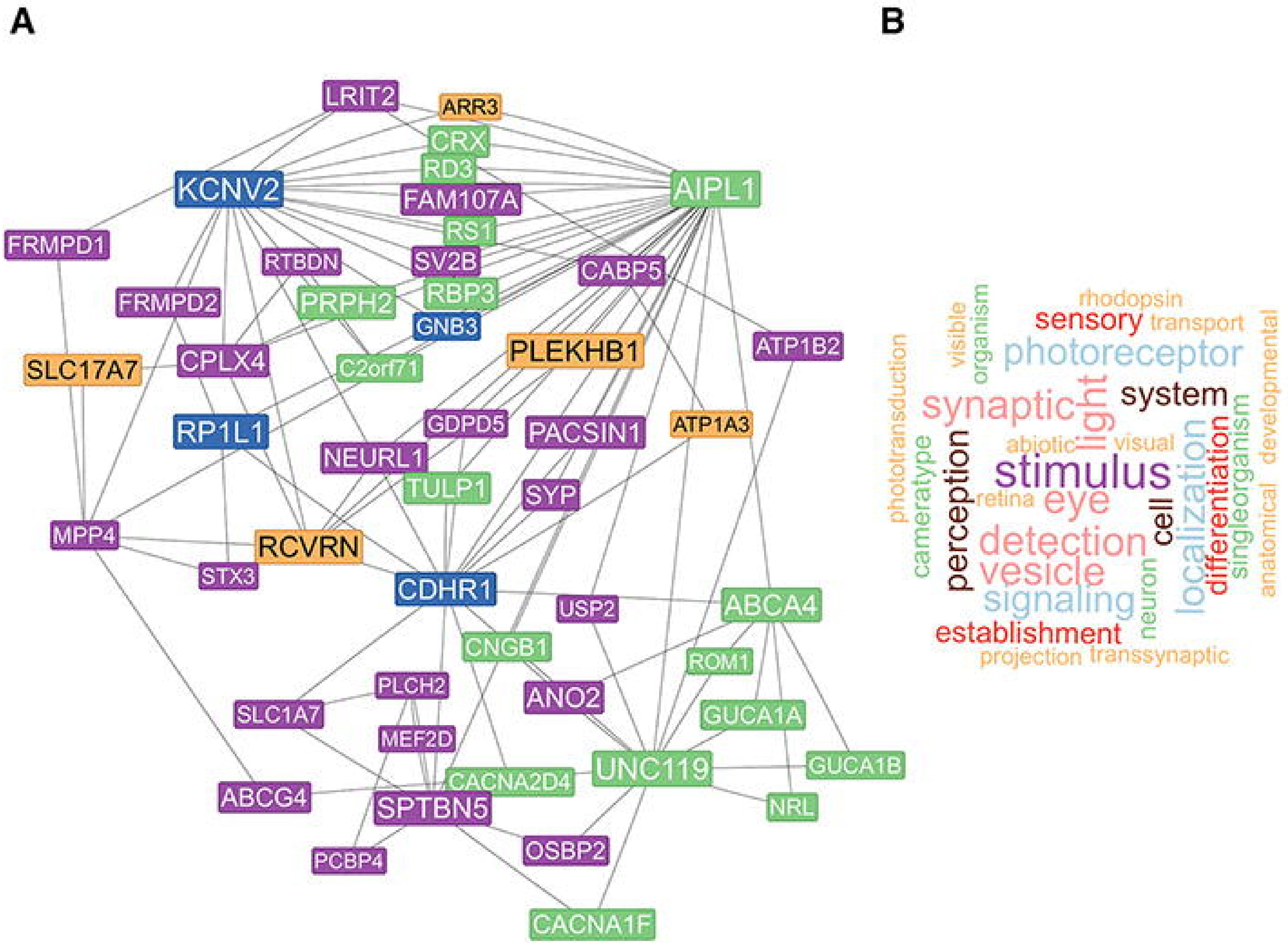


	Cornea	Retina	RPE	ESC
Adult Tissue	9	75	28*	0
Fetal Tissue	2	0	4	0
Stem Cell Line	0	20	13	9
Cell Line	9	0	11	0


A
Eye tissue t-sne
9: Cornea
5: ESC
7: ESC
8: Cornea
1: RPE
6: Retina
3: RPE
2: Retina
t-SNE 2
t-SNE 1
10
5
0
-5
-10
-5
0
5
10
15
Cluster
1
2
3
4
5
6
7
8
9
Origin
Adult Tissue
Fetal Tissue
Stem Cell Line
Cell Line
B
Pan tissue t-sne
27: Retina
29: RPE
30: ESC
31: Retina
25: Fibroblasts
26: RPE
13: Blood
23: Spleen
7: Blood
33: Cornea
14: Adrenal Gland
32: Cornea
2: Brain
11: Lung
20: Colon, Small Intestine
12: Adipose Tissue, Breast
10: Nerve
22: Esophagus, Salivary Gland
9: Skin
18: Stomach
24: Salivary Gland
5: Blood Vessel
28: RPE
19: Pancreas
8: Thyroid
16: Kidney
15: Colon, Esophagus, Stomach
17: Pituitary
1: Brain
21: Liver
3: Heart
6: Brain
4: Muscle
t-SNE 2
t-SNE 1
40
20
0
-20
-50
-25
0
25
50
Cluster
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34
C
Muscle
Heart
Blood Vessel
Fibroblasts
Cornea
Breast
Adipose Tissue
Thyroid
Nerve
Adrenal Gland
RPE
Pituitary
Small Intestine
Colon
Stomach
Esophagus
Lung
Salivary Gland
Kidney
Spleen
Blood
Pancreas
Liver
Skin
Retina
Brain
Brain
Retina
Skin
Liver
Pancreas
Blood
Spleen
Kidney
Salivary Gland
Lung
Esophagus
Stomach
Colon
Small Intestine
Pituitary
RPE
Adrenal Gland
Nerve
Thyroid
Adipose Tissue
Breast
Cornea
Fibroblasts
Blood Vessel
Heart
Muscle
-60
-40
-20
0


GO:0050877 neurological system process
GO:0007600 sensory perception
GO:0050953 sensory perception of light stimulus
GO:0007601 visual perception
GO:0003008 system process
GO:0009583 detection of light stimulus
GO:0007267 cell-cell signaling
GO:0098916 anterograde trans-synaptic signaling
GO:0007268 chemical synaptic transmission
GO:0099536 synaptic signaling
GO:0099537 trans-synaptic signaling
GO:0048858 cell projection morphogenesis
GO:0048812 neuron projection morphogenesis
GO:0031175 neuron projection development
GO:0030030 cell projection organization
GO:0030182 neuron differentiation
GO:0022008 neurogenesis
GO:0048699 generation of neurons
GO:0048666 neuron development
GO:0007399 nervous system development
GO:1903047 mitotic cell cycle process
GO:0000278 mitotic cell cycle
GO:0022402 cell cycle process
GO:0048285 organelle fission
GO:0000280 nuclear division
GO:0007067 mitotic nuclear division
GO:0007059 chromosome segregation
GO:0098813 nuclear chromosome segregation
GO:0007049 cell cycle
GO:0051276 chromosome organization
GO:0031497 chromatin assembly
GO:0006334 nucleosome assembly
GO:0006333 chromatin assembly or disassembly
GO:0071103 DNA conformation change
GO:0006323 DNA packaging
GO:0016477 cell migration
GO:0001568 blood vessel development
GO:0044707 single-multicellular organism process
GO:0032501 multicellular organismal process
GO:0048584 positive regulation of response to stimulus
GO:0048583 regulation of response to stimulus
GO:0098602 single organism cell adhesion
GO:0022610 biological adhesion
GO:0007155 cell adhesion
GO:0007166 cell surface receptor signaling pathway
GO:0007165 signal transduction
GO:0002684 positive regulation of immune system process
GO:0001775 cell activation
GO:0006952 defense response
GO:0002682 regulation of immune system process
GO:0006955 immune response
GO:0002376 immune system process
GO:0050896 response to stimulus
Block 1
Block 2
Block 3
Retina (adult) > Body (adult)
Retina (adult) < Body (adult)
RPE (immortalized cell) < Body (adult)
Cornea (adult) < Body (adult)
Cornea (immortalized cell) < Body (adult)
ESC (stem cell) < Body (adult)
Retina (stem cell) < Body (adult)
RPE (fetal) < Body (adult)
RPE (stem cell) < Body (adult)
ESC (stem cell) > Body (adult)
Cornea (immortalized cell) > Body (adult)
RPE (immortalized cell) > Body (adult)
Cornea (fetal) < Body (adult)
Cornea (adult) > Body (adult)
RPE - choroid (adult) < Body (adult)
Cornea (fetal) > Body (adult)
RPE (fetal) > Body (adult)
RPE (stem cell) > Body (adult)
Retina (stem cell) > Body (adult)
RPE - choroid (adult) > Body (adult)
0
10
20
30
40


A
Box Plot of Pan-Human Gene Expression
TTR
DCT
KIF1A
SFRP5
NELL2
log2(TPM +1)
0
5
10
15
RPE - Cell Line
RPE - Stem Cell Line
Tissue
RPE - Cell Line
RPE - Stem Cell Line
B
vasoconstriction
biological
glutamate
transsynaptic
axonogenesis
cascade
divalent
organization
memory
muscle
perception
blood
learning
component
sodium
developmental
assembly
organism
development
neurotransmitter
communication
subcellular
melanin
transmission
organismal
inorganic
morphogenesis
synapse
ear
heart
anion
singleorganism
cardiac
erk
transport
adhesion
ion
cation
synaptic
chemotaxis
remodeling
transporter
anatomical
projection
metal
neuron
contraction
behavior
system
localization
size
homeostasis
action
differentiation
potential
structure
plasmamembrane
transmembrane
activity
multicellular
sensory
cellcell
potassium
circulatory
nervous
chemical
retinoid
plasma
chloride
organic
conduction
guidance
clearance
movement
actinmediated



C

Gene	kWithin	Group
AIPL1	70.09	RetNet + Vision GO
KCNV2	69.81	RetNet
UNC119	68.72	RetNet + Vision GO
RP1L1	67.25	RetNet
PLEKHB1	66.93	Vision GO
ABCA4	65.01	RetNet + Vision GO
SPTBN5	64.20	None
RCVRN	64.01	Vision GO
PRPH2	63.46	RetNet + Vision GO
CDHR1	62.53	RetNet
CPLX4	61.45	None
PACSIN1	61.20	None
ANO2	60.87	None
FAM107A	59.95	None
RBP3	59.59	RetNet + Vision GO
SLC17A7	59.40	Vision GO
CRX	59.19	RetNet + Vision GO
TULP1	59.09	RetNet + Vision GO
NEURL1	58.38	None
SYP	58.32	None