SUPPLEMENTARY INFORMATION
https://doi.org/10.1038/s41588-019-0351-9

nature genetics

In the format provided by the authors and unedited.

Retinal transcriptome and eQTL analyses identify genes associated with age-related macular degeneration

Rinki Ratnapriya[1,8], Olukayode A. Sosina[1,2,8], Margaret R. Starostik[1,8], Madeline Kwicklis[1], Rebecca J. Kapphahn[3], Lars G. Fritsche[4], Ashley Walton[1], Marios Arvanitis[5], Linn Gieser[1], Alexandra Pietraszkiewicz[1], Sandra R. Montezuma[3], Emily Y. Chew[6], Alexis Battle[7], Gonçalo R. Abecasis[4], Deborah A. Ferrington[3*], Nilanjan Chatterjee[2*] and Anand Swaroop[1*]

[1]Neurobiology-Neurodegeneration & Repair Laboratory, National Eye Institute, National Institutes of Health, Bethesda, MD, USA. [2]Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, USA. [3]Department of Ophthalmology and Visual Neurosciences, University of Minnesota, Minneapolis, MN, USA. [4]Center for Statistical Genetics, Department of Biostatistics, University of Michigan, Ann Arbor, MI, USA. [5]Division of Cardiology, Johns Hopkins University School of Medicine, Baltimore, MD, USA. [6]Division of Epidemiology and Clinical Applications, National Eye Institute, National Institutes of Health, Bethesda, MD, USA. [7]Departments of Biomedical Engineering and Computer Science, Johns Hopkins University, Baltimore, MD, USA. [8]These authors contributed equally: Rinki Ratnapriya, Olukayode A. Sosina, Margaret R. Starostik. *e-mail: ferri013@umn.edu; nchatte2@jhu.edu; swaroopa@nei.nih.gov

nature genetics

SUPPLEMENTARY INFORMATION

https://doi.org/10.1038/s41588-019-0351-9

In the format provided by the authors and unedited.

Retinal transcriptome and eQTL analyses identify genes associated with age-related macular degeneration

Rinki Ratnapriya[1,8], Olukayode A. Sosina[1,2,8], Margaret R. Starostik[1,8], Madeline Kwicklis[1], Rebecca J. Kapphahn[3], Lars G. Fritsche[4], Ashley Walton[1], Marios Arvanitis[5], Linn Gieser[1], Alexandra Pietraszkiewicz[1], Sandra R. Montezuma[3], Emily Y. Chew[6], Alexis Battle[7], Gonçalo R. Abecasis[4], Deborah A. Ferrington[3]*, Nilanjan Chatterjee[2]* and Anand Swaroop[1]*

[1]Neurobiology-Neurodegeneration & Repair Laboratory, National Eye Institute, National Institutes of Health, Bethesda, MD, USA. [2]Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, USA. [3]Department of Ophthalmology and Visual Neurosciences, University of Minnesota, Minneapolis, MN, USA. [4]Center for Statistical Genetics, Department of Biostatistics, University of Michigan, Ann Arbor, MI, USA. [5]Division of Cardiology, Johns Hopkins University School of Medicine, Baltimore, MD, USA. [6]Division of Epidemiology and Clinical Applications, National Eye Institute, National Institutes of Health, Bethesda, MD, USA. [7]Departments of Biomedical Engineering and Computer Science, Johns Hopkins University, Baltimore, MD, USA. [8]These authors contributed equally: Rinki Ratnapriya, Olukayode A. Sosina, Margaret R. Starostik. *e-mail: ferri013@umn.edu; nchatte2@jhu.edu; swaroopa@nei.nih.gov

SUPPLEMENTARY NOTES

1. Sample processing and data generation

1.1 Tissue, RNA, and DNA preparation

RNA and DNA were isolated from 50-100 mg of homogenized retina tissue in TRIzol® (Invitrogen, Carlsbad, CA) according to a modified version of the manufacturer's protocol that included additional washing steps[1]. The order of extraction was randomized for all samples. RNA quality and quantity were evaluated using the Bioanalyzer 2100 RNA 6000 Nano assay (Agilent Technologies, Santa Clara, CA). Seven samples with RIN ≤ 5.0 were excluded from the study. DNA was quantified using the QuantiFluor® dsDNA System (Promega, Madison, WI).

1.2 RNA library preparation and sequencing

Processing order was randomized before libraries were constructed over two days largely in batches of 24 or 48 with the TruSeq® Stranded mRNA Library Preparation Kit (Illumina, San Diego, CA). The DNA concentration of the sequencing library was determined using the Bioanalyzer DNA 1000 assay (Agilent Technologies, Santa Clara, CA), and a pool of 12 barcoded libraries were layered on a random selection of one of the eight lanes of the Illumina flow cell bridge. Paired-end reads of 125 or 126 base pairs were obtained using the HiSeq 2500 platform (Illumina, San Diego, CA). Sequence data were processed for primary analysis to generate QC values (see Alignment, QC, and quantification below). Samples with a minimum of 10 million mapped reads were retained for downstream analysis.

1.2.1 RNA-Seq Quality Control (QC)

Of the 523 samples that were sequenced, twenty-six samples were excluded because of inconsistent or poor subject descriptors as follows: ocular history (1 sample), ambiguous (1) or missing MGS level (5), age at death < 55 years (7), and RIN < 5.0 (12). Six samples were removed after sequencing since < 10 million reads were mapped and/or less than 80% of reads aligned to the reference genome, and 10 samples were eliminated because of skewed gene body coverage over housekeeping genes. Six samples were taken out due to divergence from European (Caucasian) ancestry. Discordant *CFH* and *ARMS2* SNP calls between in-house and Michigan genotyping results were also removed from further analysis (*CFH*: 1 sample; *ARMS2*: 6 samples). Discordance between nominal gender, genetically inferred gender, and gender inferred from RNA-Seq Principal Component Analysis identified 7 mismatches, and these samples were not used for further analysis. Thus, a total of 70 unique samples were removed, and the entire QC process yielded 453 high-quality samples for gene expression analysis (105 MGS1, 175 MGS2, 112 MGS3, and 61 MGS4).

1.2.2 Alignment, QC, and quantification

Raw RNA-Seq reads were trimmed for Illumina adapters and low quality (SLIDING WINDOW 4:5; LEADING 5; TRAILING 5; MINLEN 25) in Trimmomatic (version 0.36)[2]. QC check was performed using FastQC (version 0.11.5) (see URLs). Trimmed reads were aligned to the Ensembl release 85 (GRCh38.p7)[3] human genome using STAR (version 2.5.2a)[4] with per-sample 2-pass mapping and ENCODE standard options. Additional QC metrics were calculated from Trimmomatic, FastQC and STAR using in-house Python and R scripts, including FASTQ and BAM file sizes, total number of reads, number of mapped and unmapped reads, and percentage of mapped reads. RNA-Seq data were also inspected for uniform full-length gene body coverage across housekeeping genes using RSeQC (version 2.6.4)[5, 6]. RSEM (version 1.13.1)[7] was used to obtain estimated gene- and transcript expression levels. Normalization was performed using Trimmed Mean of M-values (TMM) in Counts per Million (CPM) using edgeR (version 3.18.1)[8-10], and then converted into $\log_2$ CPM with an offset of 1. For eQTL analysis, normal quantile transformation was applied instead to $\log_2$(CPM) values. Non-autosomal genes and genes aligning on chromosomal patches/scaffolds were removed from reference transcriptome and eQTL analyses. Expression of cell-type specific markers in the retina did not show any significant changes across MGS stages, indicating no major loss of cell types during AMD progression (data not shown).

1.2.3 Reference annotation-based assembly

After individual transcriptomes were assembled using the Reference Annotation-based Transcript Assembly method within Cufflinks suite (version 2.21)[11], all assemblies were merged in Cuffmerge and a single, unique set of assembled transcripts was generated using Cuffcompare. Over 91% of transcripts in the reference annotation were captured (196,558 out of 215,929 transcripts), giving a comprehensive general view of the retina transcriptome. This transcript assembly was then processed using the following filters to identify putative lincRNA and protein-coding transcripts: (1) exon count, (2) transcript length, (3) coding potential, (4) functional protein domains, (5) distance to nearest protein-coding gene, and a transcript-level expression threshold at least 1 CPM ≥ 50% of MGS1 controls.

In order to identify lincRNA, multi-exonic transcripts of at least 200 base pairs were extracted from the transcript assembly. TransDecoder (version 2.0.1) [12] was applied to select for transcripts with a maximum open-reading frame of 75 amino acids lacking coding capacity. CPAT (version 1.2.2)[13] was used as a second independent method to assess coding potential, and only those lincRNA located at least 2 Kb away from the nearest protein-coding gene were retained. In order to determine protein-coding transcripts, all multi-exonic transcripts were extracted from the transcript assembly. TransDecoder was applied to select for transcripts with a minimum open reading frame of 50 amino acid residues of coding capacity, Pfam-based HMMER (version 3.1.b) (see URLs) was used to retain transcripts with best 1 domain e-value of ≤ 0.05 and at least one known functional domain, and CPAT was implemented to further assess coding potential. The logistic regression model and hexamer table required for CPAT were built using 10,000 coding sequences from the Consensus Coding Sequence Project[14] and 10,000 annotated noncoding sequences from GenCODE (release 25)[15]. The model was evaluated with 10-fold cross validation. A two-graph receiver operating characteristic curve was generated to select the optimum coding probability cutoff value (coding ≥ 0.3755; noncoding < 0.3755).

1.3 Genotyping

DNA from 516 samples, along with replicates as QC for 30 random samples, were genotyped using the UM_HUNT_Biobank v1.0 chip, which is based on the Illumina Infinium CoreExome-24 bead array platform (Illumina, San Diego, CA) with 547,655 markers and an additional 55,939 custom content markers. Genotype analysis was performed with Illumina GenomeStudio (module 1.9.4, algorithm GenTrain 2.0). We also performed TaqMan SNP genotyping for two variants, in *CFH* (Y402H; rs1061170) and *ARMS2* (A69S; rs10490924), using the ABI 7900HT sequence detection system (Applied Biosystems, Foster City, CA). The Y402H variant in *CFH* was assayed using a custom-made probe and the A69S variant in *ARMS2* was analyzed using a commercially available TaqMan probe (C_29934973_20). Briefly, 15-30 ng of DNA was mixed with TaqMan genotyping master mix (Applied Biosystems, Foster City, CA) and TaqMan SNP genotyping assay mix (40X; Applied Biosystems, Foster City, CA) in a total volume of 15 μl. Following PCR, allele discrimination was carried out with the ABI Prism 7900HT genetic detection system (Applied Biosystems, Foster City, CA).

1.3.1 eQTL QC and imputation

Of the genotyped samples, 20 samples were excluded from analysis: missingness > 5% in 1 sample, relatedness (2nd degree or higher) in 14 samples, and contradictions in inferred and reported sex in 5 samples. Initial QC at the SNP-level involved (1) removal of SNPs with HWE p-value $< 1 \times 10^{-6}$, (2) call rate < 95%, and (3) duplicate and non-autosomal variants. We retained 570,441 variants. Genotypes were imputed with IMPUTE2 (version 2.3.1)[16] based on the 1000 Genomes Project Phase 3 reference panel (October 2014) [17]. For our eQTL analysis, QC after imputation excluded: (1) poorly imputed variants (info < 0.3), (2) indels of length > 51 bp, (3) imputed variants with HWE $< 1 \times 10^{-6}$, (4) imputed variants with MAF < 1%, and (5) monomorphic variants. In total, 8,924,684 autosomal variants across 406 individuals were retained, and coordinates were then converted from Ensembl GRCh 37.p13 to Ensembl GRCh

38.p7 in order to match the retina RNA-Seq data. Population stratification was examined using Eigenstrat (version 7.2.1) to identify 11 significant principal components[18, 19]; 10 of these were used in the final eQTL model.

2. Batch correction

Exclusion criteria for negative control genes in SSVA included: (1) Genes within 100 Kb of linkage disequilibrium of known 34 AMD susceptibility loci identified in the most recent GWAS study for AMD[20], (2) RetNet (retinal Information Network) genes (see URLs), (3) AMD candidate genes from PubMed literature search over the last five years (see Weighted Gene-correlation Network Analysis in Methods), (4) aging- and gender-associated genes from GTEx analysis[21], (5) X and Y chromosomal genes, and (6) genes that did not meet the expression-level threshold ≥1 CPM in ≥ 10% of all samples.

3. eQTL, TWAS, and eCAVIAR

3.1 Enrichment

We examined whether there is a broader relationship between *cis*-eQTLs and AMD genetic susceptibility beyond what has been observed for known GWAS loci. A Q-Q plot for each of the GWAS datasets was generated by: (1) subsetting to International HapMap Project phase 3 (NCBI build 36, dbSNPb129) variants in the European population with MAF ≥ 5%, (2) removal of variants in the major histocompatibility complex region, and (3) removal of variants within +/- 1 Mb of the known GWAS signals. We then stratified the variants into multiple (overlapping) categories based on eQTL characteristics: (1) retina-specific eQTLs: eVariants that regulate gene expression only in retina, (2) GTEx-1 eQTLs: eQTLs that regulate gene expression in at least 1 GTEx tissue (3) GTEx-20 eQTLs: eQTLs that regulate gene expression in at least 20

GTEx tissues, and (4) GTEx-40 eQTLs: eQTLs that regulate gene expression in at least 40 GTEx tissues.

3.2 Colocalization

Fine mapping using eCAVIAR (version 2.0) was performed in the following manner: (1) for each lead variant in GWAS, a 1Mb window around it was defined as its locus, (2) for all variants within the locus, we identified/defined target genes as genes that are associated at FDR ≤ 5% with any of these variants in the eQTL study, and (3) we calculated the colocalization posterior probability (CLPP) for each variant and target gene within the loci. The most relevant target gene was then defined as the gene with the highest CLPP above the threshold of 1% within the loci. A maximum of three possible causal variants for each locus was assumed.

3.3 TWAS

The TWAS procedure required that we model gene expression with genotype. The gene expressions were modeled using either elastic net [22], mixed models, or least absolute shrinkage and selection operator (LASSO). The LASSO[23] lambda parameter was calculated using the heritability; genes for which the heritability could not be calculated used the average heritability across genes instead. Of the 18,053 genes expressed in the retinal samples, 17,345 were present in the TWAS analysis. Genes not analyzed in TWAS were located on either sex chromosomes, the mitochondrial chromosome, on scaffolds, or did not have SNPs within 1 Mb of the merged GWAS-eQTL SNP set. The mean cross-validated model fit was 0.07, and the mean heritability of the 14,353 genes for which it could be calculated was 0.127. As expected, the higher the heritability, the better the cross-validated model fit. LASSO was the best fit for approximately half of the genes, and elastic net accounted for another quarter; genes for which the mixed model provided the best fit had models that captured less variation in expression than other genes.

The TWAS statistics does not take into account LD between genes, so we performed summary-level equivalent conditional tests for each chromosome for genes that were both significant at an FDR of 5% and had a genetic expression model R^2 > than 0.01. Genes were added in a stepwise manner into the model, from lowest marginal p-value to highest, until no gene remained significant. The model prior to this saturation was used as the final conditional model; no provision was made to prevent over-fitting[24]. Of the 61 genes tested, 47 remained nominally significant at $\alpha = 0.05$; of these, 39 remained significantly associated after Bonferroni correction for multiple testing (using all 61 genes considered for the test, not just ones included in the models). A permutation test (described in Methods) was also performed; seven genes were significant after Bonferroni correction and had a gene model $R^2 > 0.01$, and three of these were outside of the GWAS loci: *PARP12*, *MTMR10*, and *SH3BGR*.

We explored the tissue specificity of these results, at least in part, using GTEx data v6. We downloaded the pre-computed TWAS weights derived from the data of 39 GTEx tissues (excluding cell lines and biological replicate tissues, such as frontal cortex and cerebellar hemisphere) from the TWAS website (http://gusevlab.org/projects/fusion/) and performed the procedure for the GTEx weights with the same set of AMD GWAS summary statistics that was used with retina. The complete results of the TWAS analysis – gene model attributes, marginal association statistics, conditional and permutation test results, and GTEx marginal associations for the retinal candidates with FDR < 0.05 – are provided in Supplementary Data S5. Please note that relatively few genes had weights available in most GTEx tissues.

LocusZoom (version 0.4.8)[29] was used to visualize the genetic architecture of the AMD GWAS around the novel gene candidates found by TWAS, as well as the known regions for which we identified significant overlapping eQTLs.

3.4 Evaluation of AMD GWAS lead variants for eQTL evidence in non-retina tissues

Of the 52 lead variants from AMD-GWAS[20], 41 were analyzed in our study. Those not found were either not in the reference dataset used for imputation (6 variants) or did not pass our MAF threshold (5 variants, MAF threshold; 1%). Matrix eQTL (version 2.1.1)[25] was then used to obtain the marginal associations using the same *cis* criteria, which were then corrected for multiple testing using the Bonferroni approach at the type I error rate of 5%.

We compared our findings to that of Strunz and colleagues[26] which includes eQTLs from liver samples of 588 individuals and GTEx (v7). For this comparison, we used 31 SNPs with MAF ≥ 5% that were common to both studies. For each variant, eQTL analysis was performed for all genes that are present within a 1Mb window and expressed in the two tissues (Supplementary Data 3). We also tested 37 AMD-associated variants (with MAF ≥ 1%) that were analyzed in the retina and detected in at least one GTEx (v7) tissue. For each SNP-gene combination, we list all the GTEx tissues that had p-values less than or equal to that of retina (Supplementary Data 3), or if no GTEx tissue had p-value lower than the retina, we listed all tissues with their respective p-values.

4. Gene expression analysis

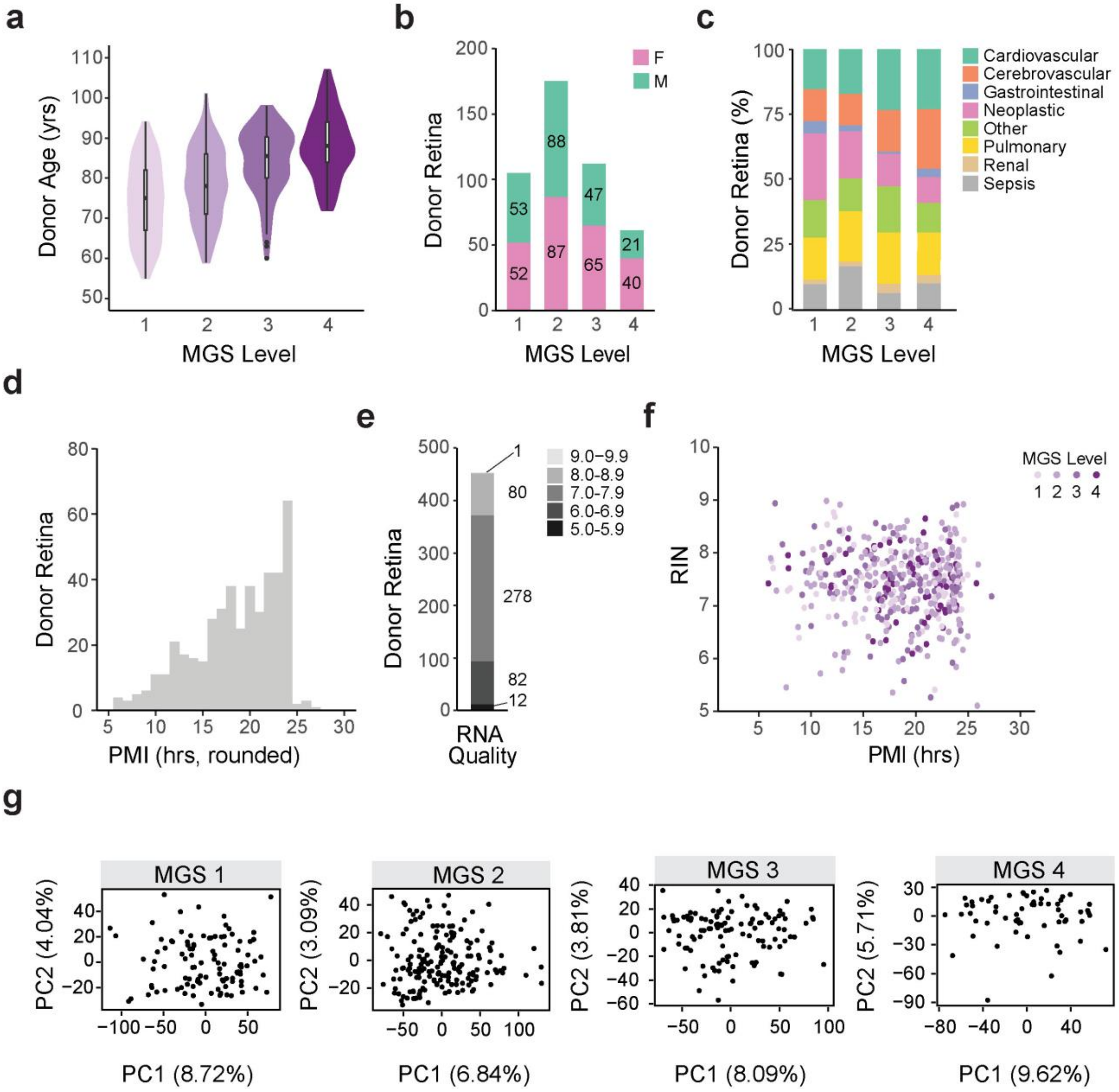
4.1 GSEA

We focused on gene sets that passed a significance threshold of FDR q-value ≤ 0.25 and on key genes that appeared in at least 25% of gene sets in common functional categories using Leading Edge Analysis (Supplementary Data S6). Comparison of early AMD to controls identified 38 significantly enriched gene sets, all upregulated and generally relating to cell killing (3), metabolism (12), and the immune system (15). The largest of these categories involved immune system processes (13) with an average normalized enrichment score (avg. NES) of 2.4 and 80 key genes. Comparison of intermediate AMD to controls identified 6 upregulated and 60 downregulated significantly enriched gene sets comprising metabolism (2), cell killing (2), and cellular component organization (3). Comparison of advanced AMD to controls identified 44 upregulated and 15 downregulated significantly enriched gene sets including those relating to metabolism (21), cell component organization (9), immune system (6), and stress response (4). Additionally, we identified downregulated gene sets that were predominant and largely exclusive to intermediate AMD and associated with synapses in cell communication (14, avg. NES = -2.2), nervous system development (9, avg. NES = -2.4), biological regulation (4, avg. NES = -2.3), and establishment/maintenance of cell polarity (3, avg. NES = -2.4) (Supplementary Table S6).

4.2 Comparison of transcriptomes across retina and GTEx tissues

The bioinformatics pipeline used to analyze RNA-Seq data in this study mainly differed from that of GTEx v7 in gene quantification methods and gene annotation version. To understand the consequences of using different pipelines and to ensure appropriate tissue comparisons between studies, multidimensional scaling plots and hierarchical clustering dendrograms were generated based on normalized gene expression levels from the different pipelines. Statistical methods used to generate the multi-dimensional scaling (MDS) plot itself were obtained from GTEx[27, 28]. Three comparisons were made based on the following data sets: (1) Raw GTEx v7 data processed through our pipeline, (2) Raw GTEx v7 and retina data processed through our pipeline, and (3) GTEx v7 gene-level TPM count data provided on the GTEx online portal.

Raw GTEx v7 data were processed through our pipeline as previously mentioned in 1.2.2 Alignment, QC, and quantification. In addition, we used similar methods that GTEx had applied to detect samples outliers[27, 28]. PCA-based outlier detection was performed in the first two principal components by using Mahalanobis distance to center the data. Outliers were identified using a threshold of three standard deviations.



Supplementary Figure 1

Characteristics of retina donor samples (n = 453) used in this study.

(a) Violin plots showing the age distribution, in years, of donors across the four MGS stages. The boxplot within each violin plot depicts the median, and the lower and upper hinges correspond to the first and third quartiles, respectively. Outlying data are represented by individual points that extend beyond 1.5 × interquartile range below the first quartile or above the third quartile. The mean age of donors was 80 years (range 55-107), and the mean donor age increased with AMD severity: 74 years (range 55-94) in MGS1, 78 years (59-101) in MGS2, 84 years (60-98) in MGS3, and 88 years (range 72-107) in MGS4.

(b) Distribution of gender across the four MGS stages. Gender was distributed almost evenly in MGS1 to MGS3, with almost twice as many females as males in MGS4.

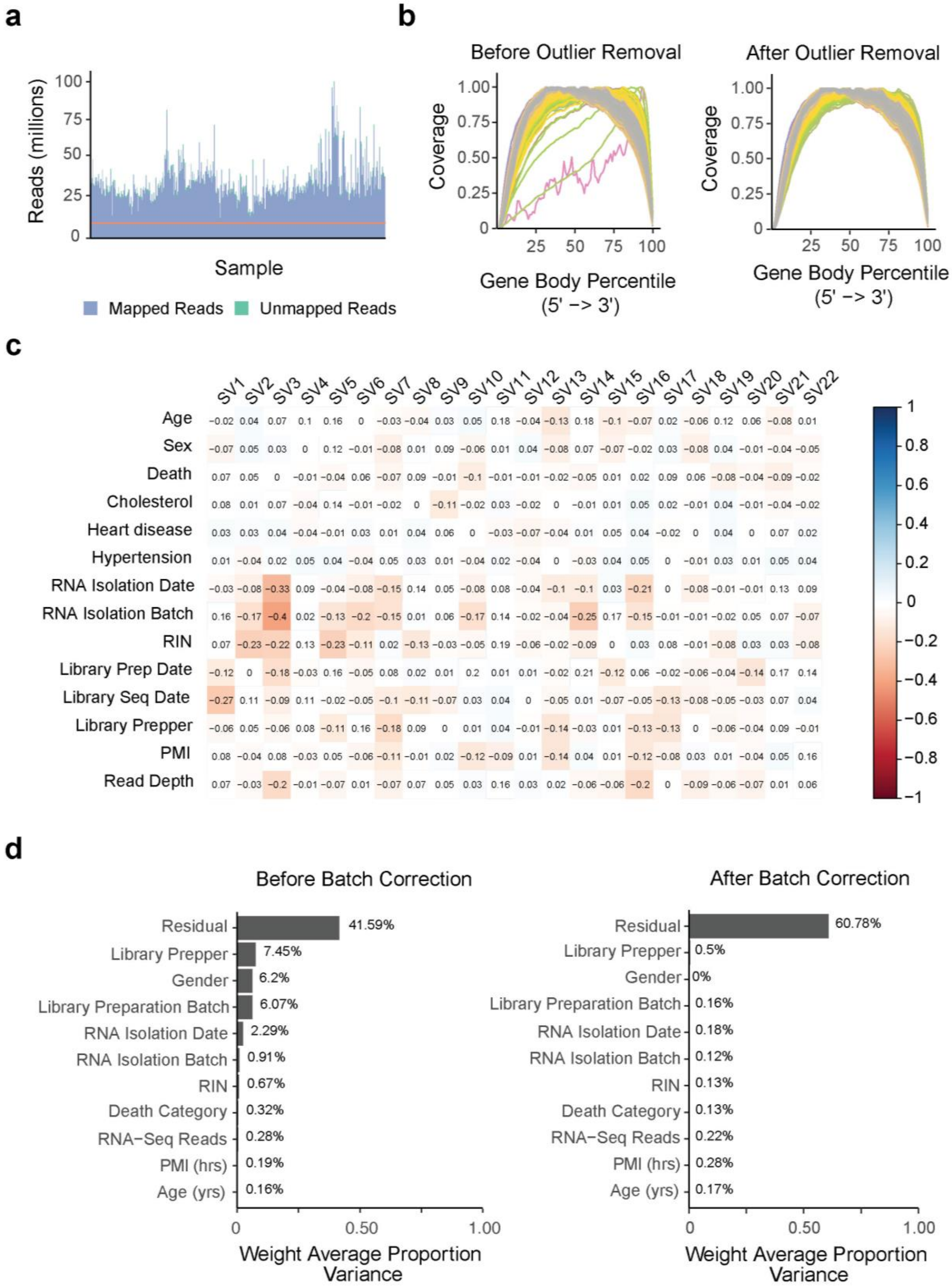
(c) The cause of death across the four MGS stages. Donors within each MGS stage were grouped into 8 categories based on the reported cause of death to determine that causes of death were not conflated with donor age or MGS stage.

(d) Distribution of post-mortem interval (PMI), in hours. PMI was defined as the mean time lapse from death to enucleation and tissue cryopreservation. Mean PMI was 18.66 hours.

(e) Quality of RNA, as defined by the RNA Integrity Number (RIN), used for RNA-Seq. Mean RIN was 7.42 ± 0.6 (5.1-9).

(f) Scatterplot of RNA integrity (RIN) versus post-mortem interval (PMI).

(g) PCA plots of donors within each MGS level (105 MGS1, 175 MGS2, 112 MGS3, and 61 MGS4 donors) based on normalized gene expression levels.

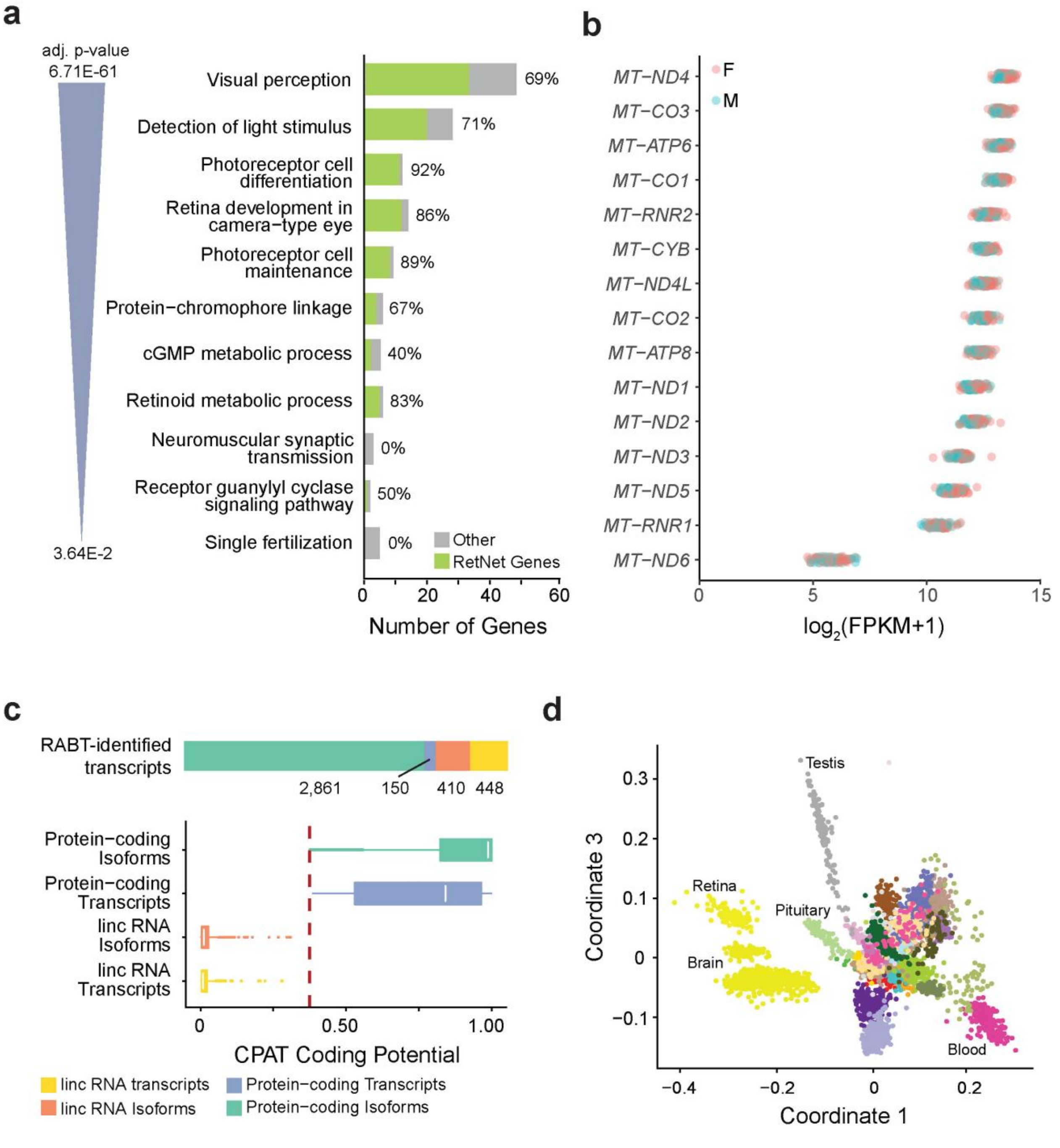


	SV1	SV2	SV3	SV4	SV5	SV6	SV7	SV8	SV9	SV10	SV11	SV12	SV13	SV14	SV15	SV16	SV17	SV18	SV19	SV20	SV21	SV22
Age	−0.02	0.04	0.07	0.1	0.16	0	−0.03	−0.04	0.03	0.05	0.18	−0.04	−0.13	0.18	−0.1	−0.07	0.02	−0.06	0.12	0.06	−0.08	0.01
Sex	−0.07	0.05	0.03	0	0.12	−0.01	−0.08	0.01	0.09	−0.06	0.01	0.04	−0.08	0.07	−0.07	−0.02	0.03	−0.08	0.04	−0.01	−0.04	−0.05
Death	0.07	0.05	0	−0.01	−0.04	0.06	−0.07	0.09	−0.01	−0.1	−0.01	−0.01	−0.02	−0.05	0.01	0.02	0.09	0.06	−0.08	−0.04	−0.09	−0.02
Cholesterol	0.08	0.01	0.07	−0.04	0.14	−0.01	−0.02	0	−0.11	−0.02	0.03	−0.02	0	−0.01	0.01	0.05	0.02	−0.01	0.04	−0.01	−0.04	−0.02
Heart disease	0.03	0.03	0.04	−0.04	−0.01	0.03	0.01	0.04	0.06	0	−0.03	−0.07	−0.04	0.01	0.05	0.04	−0.02	0	0.04	0	0.07	0.02
Hypertension	0.01	−0.04	0.02	0.05	0.04	−0.04	0.05	0.03	0.01	−0.04	0.02	−0.04	0	−0.03	0.04	0.05	0	−0.01	0.03	0.01	0.05	0.04
RNA Isolation Date	−0.03	−0.08	−0.33	0.09	−0.04	−0.08	−0.15	0.14	0.05	−0.08	0.08	−0.04	−0.1	−0.1	0.03	−0.21	0	−0.08	−0.01	−0.01	0.13	0.09
RNA Isolation Batch	0.16	−0.17	−0.4	0.02	−0.13	−0.2	−0.15	0.01	0.06	−0.17	0.14	−0.02	−0.04	−0.25	0.17	−0.15	−0.01	−0.01	−0.02	0.05	0.07	−0.07
RIN	0.07	−0.23	−0.22	0.13	−0.23	−0.11	0.02	−0.13	−0.03	−0.05	0.19	−0.06	−0.02	−0.09	0	0.03	0.08	−0.01	−0.08	0.03	0.03	−0.08
Library Prep Date	−0.12	0	−0.18	−0.03	0.16	−0.05	0.08	0.02	0.01	0.2	0.01	0.01	−0.02	0.21	−0.12	0.06	−0.02	−0.06	−0.04	−0.14	0.17	0.14
Library Seq Date	−0.27	0.11	−0.09	0.11	−0.02	−0.05	−0.1	−0.11	−0.07	0.03	0.04	0	−0.05	0.01	−0.07	−0.05	−0.13	−0.08	−0.05	−0.03	0.07	0.04
Library Prepper	−0.06	0.05	−0.06	0.08	−0.11	0.16	−0.18	0.09	0	0.01	0.04	−0.01	−0.14	−0.03	0.01	−0.13	−0.13	0	−0.06	−0.04	0.09	−0.01
PMI	0.08	−0.04	0.08	−0.03	0.05	−0.06	−0.11	−0.01	0.02	−0.12	−0.09	0.01	−0.14	0.04	0.01	−0.12	−0.08	0.03	0.01	−0.04	0.05	0.16
Read Depth	0.07	−0.03	−0.2	−0.01	−0.07	0.01	−0.07	0.07	0.05	0.03	0.16	0.03	0.02	−0.06	−0.06	−0.2	0	−0.09	−0.06	−0.07	0.01	0.06

Supplementary Figure 2

RNA-Seq QC metrics (n = 453).

(a) Number of RNA-Seq reads that mapped to the human reference genome Ensembl 38.85. The red horizontal line denotes 10 million reads.
(b) Normalized mean per-base 5’ to 3’ gene body coverage of housekeeping genes. Left: before outlier removal. Right: after outlier removal.
(c) Pearson’s correlation between 22 significant surrogate variables identified in SSVA and possible documented sources of variation. A p-value of 0.05 was used as the significance threshold. Correlation coefficients are labeled in black and color-coded such that positive correlations are displayed in blue and negative correlations in red. Color intensity is proportional to the correlation coefficients. RIN: RNA Integrity Number; PMI: post-mortem interval.
(d) Principal variance component analysis (PVCA) of the retina gene expression data set. Residual represents the remaining variance in the data set not attributed to the specified batch and biological variables. Left: before batch correction. Right: after batch correction. RIN: RNA Integrity Number; PMI: post-mortem interval.



Supplementary Figure 3

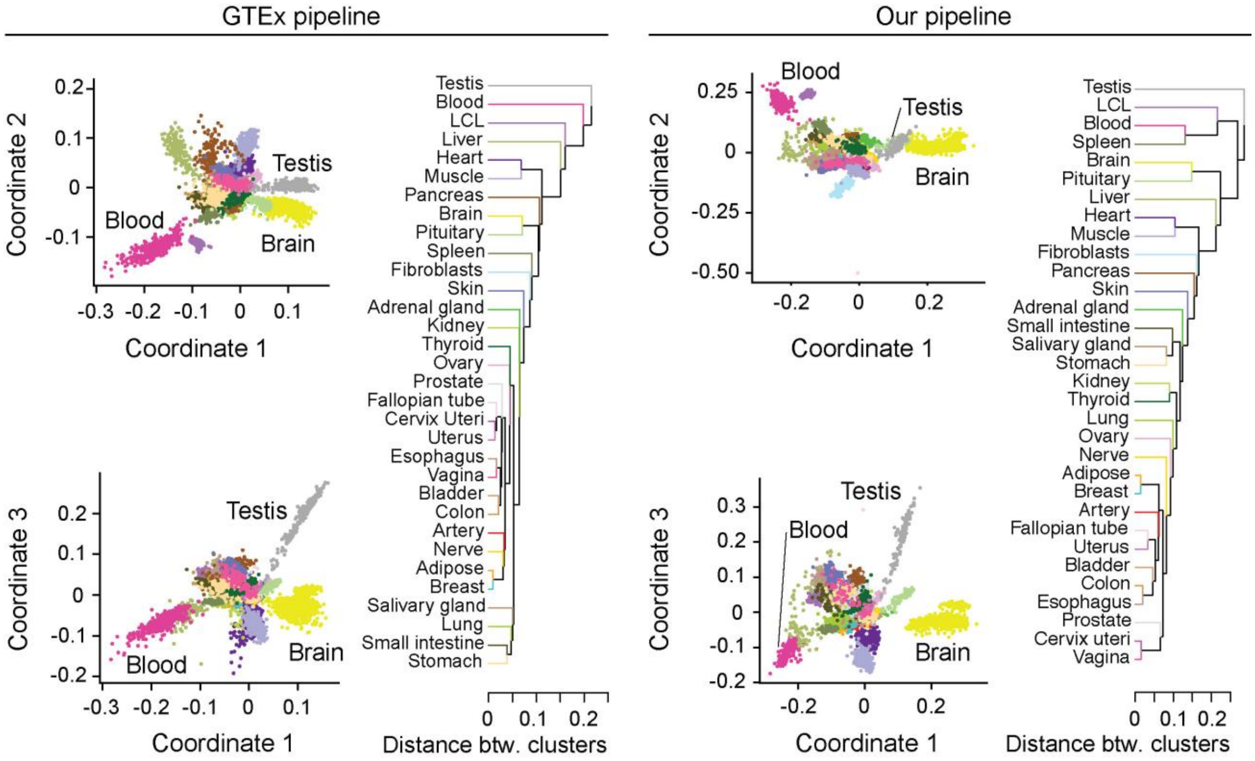
Reference transcriptome of the human retina (n = 105 MGS1 control donor retinas).

(a) Gene Ontology (GO) Biological Process pathway enrichment analysis of 186 high abundance genes (≥ 100 FPKM) in the retina. The bars represent the number of genes identified in each pathway, highlighting in green the number of inherited retinal diseasecausing genes in the RetNet database of ocular diseases (percentage indicated to the right of bar). Redundancy of enriched GO terms was removed using a similarity cutoff of 0.40. Enrichment was determined by hypergeometric tests; a Benjamini-Hochberg adjusted p-value ≤ 0.05 was used as the significance threshold.

(b) Scatter plot of mitochondrial gene expression based on $\log_2$ (FPKM+1) values among 53 males and 52 females.

(c) Novel transcript discovery using reference annotation-based transcript assembly. Top: Number of putative novel protein-coding and lincRNA isoforms and transcripts. Bottom: Coding Potential Assessment Tool (CPAT) coding probability score of putative novel protein-coding and lincRNA isoforms and transcripts. The dotted red vertical line denotes the calculated coding probability cutoff of 0.3755. We discovered a total of 410 and 2,861 lincRNA and protein-coding isoforms, respectively, and a total of 150 and 448 lincRNA and protein-coding transcripts, respectively. Boxplots depict the median (white line), and the lower and upper hinges correspond to the first and third quartiles, respectively. Individual points represent outlying data that extend beyond 1.5 × interquartile range below the first quartile or above the third quartile.

(d) Multidimensional scaling plot of samples across tissues based on normalized gene expression levels. We plotted 105 MGS1 control retinas and 6,421 GTEx samples across all body sites.

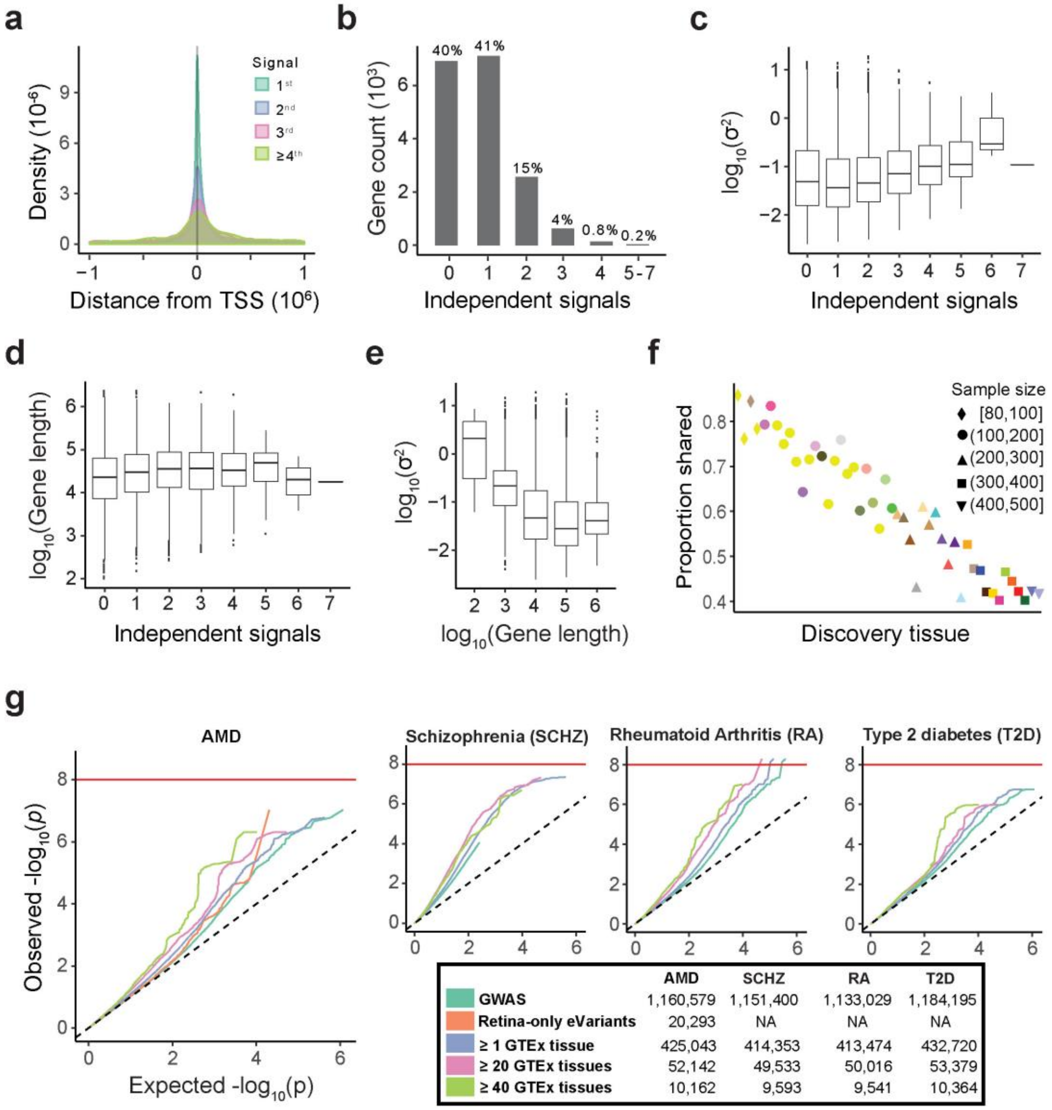


Supplementary Figure 4

Comparison of RNA-Seq analysis pipelines using GTEx data without retina (n = 6,421 samples across all body sites).

Multidimensional scaling plots and hierarchical clustering dendrograms of samples across tissues based on normalized gene expression levels. Left: based on our bioinformatics pipeline. Right: based on GTEx v7 gene-level TPM count data. These comparisons suggest that the relationship between tissues was not affected by the analysis pipeline.

Our RNA-seq analysis pipeline was based on the most recent literature recommendations for RNA-Seq analysis (as described in Methods) and mainly differed from that of GTEx in gene quantification methods and in gene annotation version. We therefore downloaded the raw GTEx data and processed these through our bioinformatics pipeline to generate the MDS plot. Statistical methods used to generate the MDS plot itself were obtained from GTEx. In addition, we explored whether similar findings could be obtained using a different analysis pipeline. We also plotted MDS plots from expression data provided on the GTEx online portal. MDS plots and hierarchical clustering dendrograms generated from different pipelines were comparable.



	AMD	SCHZ	RA	T2D
🟩 GWAS	1,160,579	1,151,400	1,133,029	1,184,195
🟧 Retina-only eVariants	20,293	NA	NA	NA
🟦 ≥ 1 GTEx tissue	425,043	414,353	413,474	432,720
🟪 ≥ 20 GTEx tissues	52,142	49,533	50,016	53,379
🟩 ≥ 40 GTEx tissues	10,162	9,593	9,541	10,364

Supplementary Figure 5

eQTL analysis of human retina (n = 406).

(a) The relationship between the strength of each *cis*-eQTL's association and the distance of its eVariant from its eGene's transcription start site (TSS)**.**

(b) The distribution of *cis*-independent signals for each autosomal gene. Thus approximately 60% of genes in the retina were found to be under genetic control with the majority of the genes having one independent signal (41%).

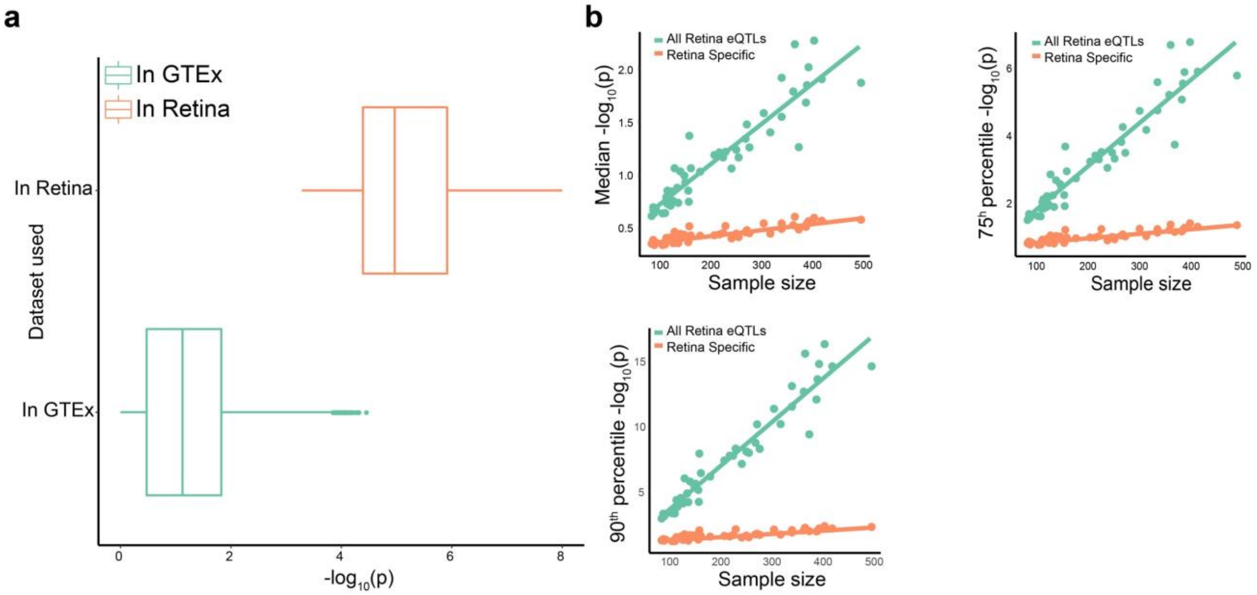
(c) Distribution of the amount of variability left unexplained in gene expression levels after correction for other covariates used in the model stratified by the number of independent signals found per gene.

(d) Distribution of gene length stratified by the number of independent signals found per gene.

(e) Distribution of the amount of variability left unexplained in gene expression levels after correction for other covariates used in the model ordered by gene length.

(f) Proportion of *cis*-eQTLs discovered in GTEx that were replicated in the retina (y-axis), ordered by sample size in discovery tissue (x-axis). The color and shape of the points represent the sample size of the replication tissue.

(g) Q-Q plot indicating the relationship between the observed -$\log_{10}$ p-values (observed p-values, based on two-sided t-tests, obtained from the most recent GWAS study for AMD[20]) for each stratum relative to its expected null distribution. Each stratum, except for the GWAS one, classifies the eVariants by how many tissues they regulate at least one gene in. This analysis is shown for AMD, schizophrenia, rheumatoid arthritis, and Type 2 diabetes. Boxplots (c-e) show the median; the lower and upper hinges correspond to the first and third quartiles, respectively. Outlying data are represented by individual points that extend beyond 1.5 × interquartile range below the first quartile or above the third quartile.

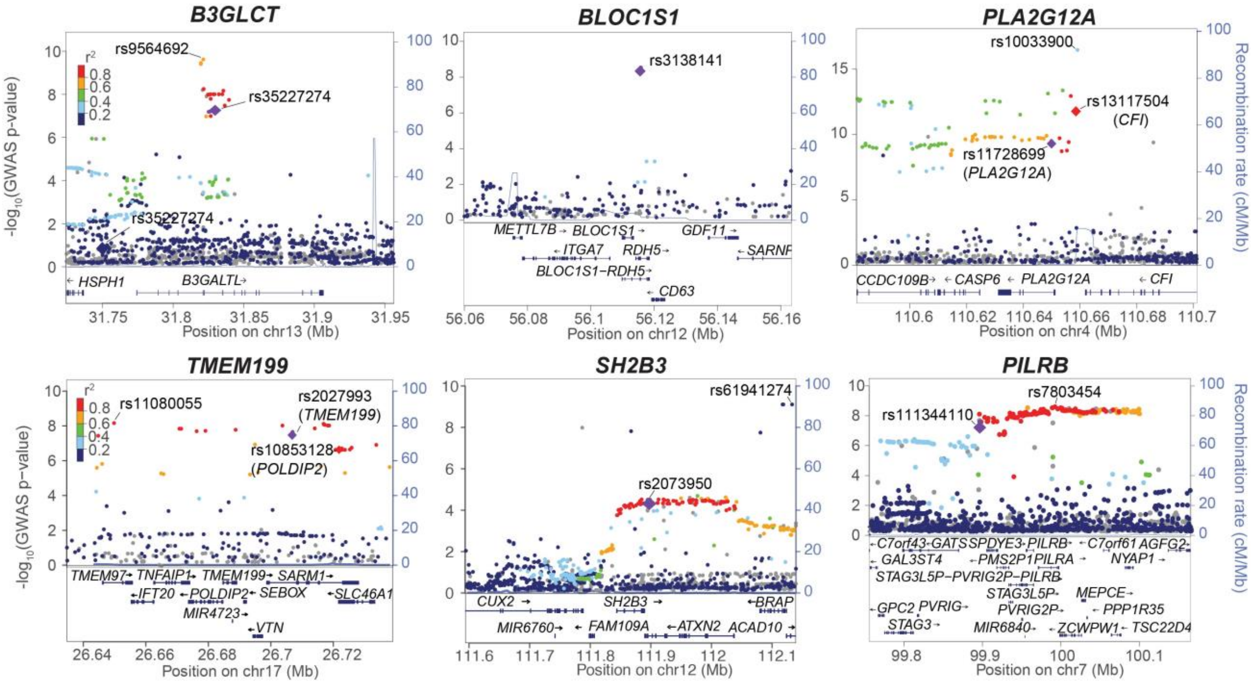


Supplementary Figure 6

Comparison of retina-specific eQTLs across GTEx.

(a) Boxplots showing minimum p-values across GTEx tissues for eQTLs detected only in the retina, after correcting for the number of tissues the eQTLs were tested in. As a comparison, distribution of *p*-values in the retina analysis for the same eQTLs are also shown. The distribution of *p*-values between retina and other tissues is expected given that these SNPs, by definition, are significant eQTLs in retina, but not in other tissues. Boxplots show the median; the lower and upper hinges correspond to the first and third quartiles, respectively. Outlying data are represented by individual points that extend beyond 1.5 × interquartile range below the first quartile or above the third quartile. All plots are based on p-values obtained from two-sided t-tests using 406 individuals.

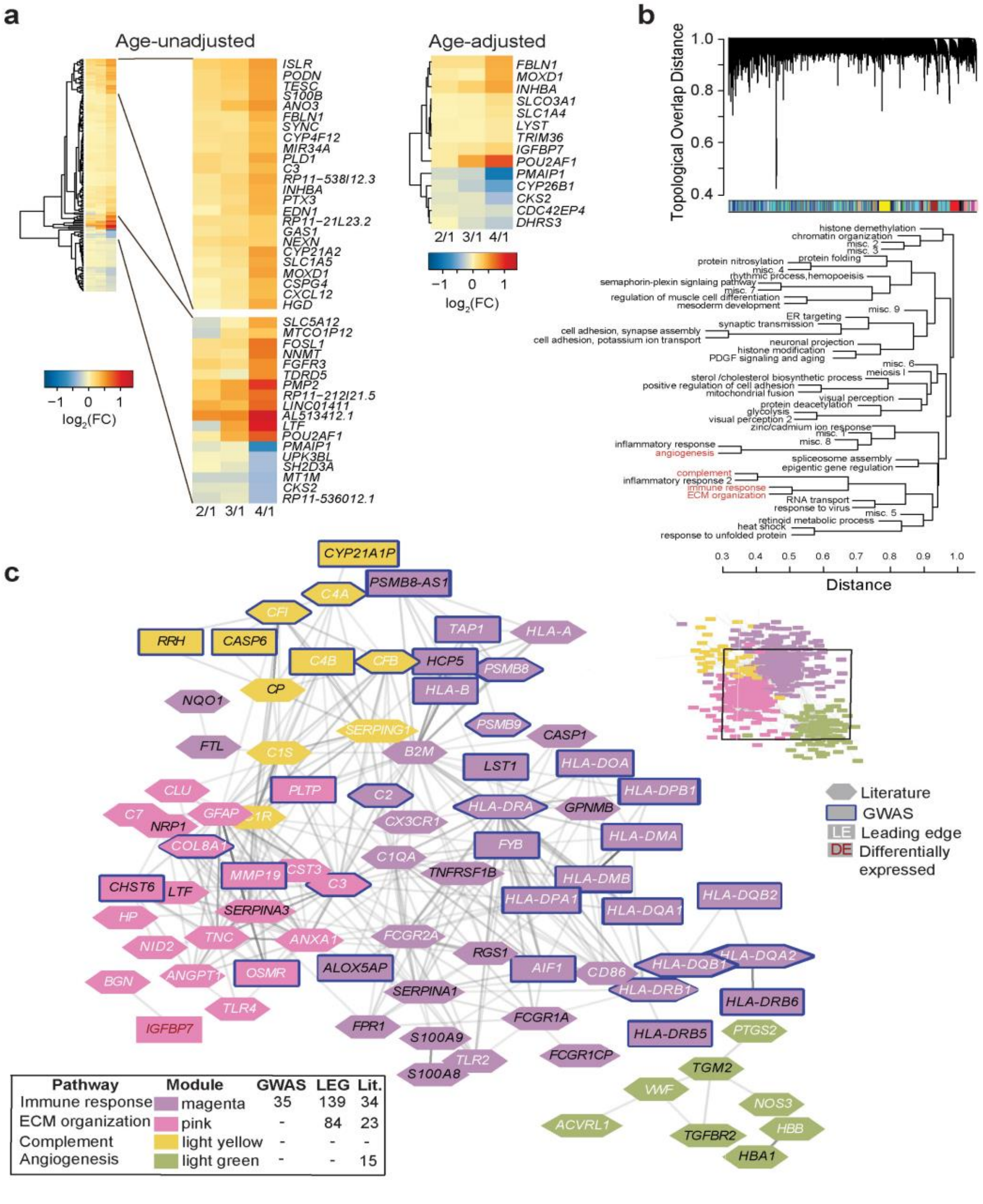
(b) Median, 75th, and 90th percentile of -log$_{10}$(p-values) of retina-specific *cis*-eQTLs in different non-retina tissues against their respective sample sizes. These plots were generated to explore whether SNPs that were not detected as significant eQTLs in non-retina tissues using the stringent p-value threshold could still reveal some enrichment towards lower p-values than what is expected by chance. We also compared this trend for all eQTLs detected, regardless of whether they were retina-specific or not. A weak trend towards lower p-values in tissues with large sample sizes for retina-specific eQTLs was observed. However, this trend was much weaker compared to that observed for all eQTLs. It appears that retina-specific eQTLs have stronger effects in the retina though possibility of weak effects of these eQTLs in other tissues cannot be ruled out.



Supplementary Figure 7

Manhattan plots at known AMD loci.

LocusZoom (version 0.4.8)[29]-generated Manhattan plot using the IAMDGC GWAS results (two-sided z-tests with no adjustments for multiple testing, n_{cases} = 16,144, $n_{controls}$ = 17,832)[20] of GWAS regions encompassing the candidates that fell within known AMD loci and were shown to be associated through multiple methods of analysis, as specified by Table 1. The top variants for the independent eQTL signals determined by the conditional analysis are displayed as diamonds and labeled. The SNP with the strongest GWAS signal in the region is also identified in each plot. Coloration of the points is determined by strength of linkage disequilibrium (LD) with respect to the top variant of the strongest eQTL signal. If LD information provided to LocusZoom was absent for that SNP, one of its proxies according to LDLink [30] (R^2 > 0.99) was used. Recombination rate is shown as a blue line.



Pathway	Module	GWAS	LEG	Lit.
Immune response	magenta	35	139	34
ECM organization	pink	-	84	23
Complement	light yellow	-	-	-
Angiogenesis	light green	-	-	15

Supplementary Figure 8

Differential expression and WGCNA analysis.

(a) Heatmap showing the expression pattern of differentially expressed genes by comparing advanced AMD (n = 61) to controls (n = 105) with and without adjusting for age at the significance threshold of ≤ FDR 0.20.

(b) We identified 47 modules, each containing between 16 and 4,847 genes (for 18,053 genes total). Top: Dendrogram of genes with topological overlap used as distance (shown on y-axis). The color bar below indicates which module the genes belong to. Bottom: Hierarchical clustering of module expression eigenvalues (eigengenes). The modules involved in complement (yellow), angiogenesis (light green), immune activation (magenta), and extracellular matrix (pink) are highlighted in red. These modules were adjacent to each other according to eigenvalue-based hierarchical clustering.

(c) Two of these modules were particularly interesting as they were enriched for literature (pink FDR = 2.21×10^{-3} via a hypergeometric test; magenta FDR = 1.37×10^{-9}, n_{genes} = 18,053 used in test) and leading edge (pink FDR = 1.10×10^{-3}; magenta FDR = 1.33×10^{-26}) candidate genes. Additionally, the magenta module was enriched for genes from the GWAS loci (FDR = 2.38×10^{-4}). The pink module also contained three DE- (*FBLN1*, *MOXD1*, *IGFBP7*) and two AMD-associated genes (*COL8A1* and *MMP19*). GO analysis of the magenta and pink module highlighted extracellular matrix organization and immune response pathways, respectively, which were previously implicated in AMD pathology. These modules interacted closely with two other modules; the light green (also enriched for literature genes, FDR = 8.30×10^{-3}) and light yellow, which were enriched for angiogenesis and complement GO terms, respectively. We show only genes that fall in either literature, GWAS, or differentially expressed groups and are strongly correlated with another such candidates (adjacency > 0.05).

Supplementary Table 1. Summary of eQTL, eCAVIAR and TWAS analyses for prioritizing variants and target genes across AMD-GWAS loci.

AMD Locus	**Lead GWAS SNP**	**Chr:Position**	**GWAS_pval**	**eQTL_pval**	**Target gene(s)**	**% Variability Explained**	**Significant TWAS genes in the locus (FDR < 0.05)**
B3GALTL	rs9564692	13:31821240	3.31×10^{-10}	2.36×10^{-11}*	*B3GLCT*†	10.47	*B3GLCT* (1.37×10^{-4})
RDH5/CD63	rs3138141	12:56115778	4.3×10^{-9}	5.69×10^{-19}*	*BLOC1S1*†, *RP11-644F5.10*	17.80	*BLOC1S1* (6.36×10^{-6}), *RP11-644F5.10* (2.89×10^{-6})
SLC16A8	rs8135665	22:38476276	5.53×10^{-11}	1.56×10^{-3}	*CTA-228A9.3*†	2.45	*CTA-228A9.3* (1.26×10^{-5})
ACAD10	rs61941274	12:112132610	1.07×10^{-9}	8.95×10^{-2}	*SH2B3*†	0.71	*SH2B3* (2.16×10^{-2})
PILRB/PILRA	rs7803454	7:99991548	4.76×10^{-9}	3.57×10^{-77}*	*PILRB*, *STAG3L5P*, *PILRA*, *ZCWPW1*, *TSC22D4*	57.51	*MEPCE* (5.83×10^{-6}), *PMS2P1* (1.11×10^{-5}), *STAG3L5P-PVRIG2P-PILRB* (1.88×10^{-5}), *PILRB* (1.88×10^{-5})
TMEM97/VTN	rs11080055	17:26649724	1.04×10^{-8}	8.37×10^{-19}*	*POLDIP2*, *SLC13A2***, *TMEM199*†	17.65	*TMEM199* (2.37×10^{-5}), *POLDIP2* (8.27×10^{-5})
CFI	rs10033900	4:110659067	5.35×10^{-17}	3.98×10^{-7}*	*PLA2G12A*	6.17	*CFI* (3.31×10^{-10}), *PLA2G12A* (4.53×10^{-10})
KMT2E/SRPK2	rs1142	7:104756326	1.35×10^{-9}	6.49×10^{-6}*	*CTB-152G17.6***	4.91	
NPLOC4/TSPAN10	rs6565597	17:79526821	1.45×10^{-11}	1.91×10^{-5}*	*ARL16*	4.43	*ANAPC11*‡ (4.03×10^{-3})
C2/CFB/SKIV2L	rs114254831	6:32155581	9.4×10^{-12}	4.70×10^{-6}*	*HLA-DQB1*	5.06	*SKIV2L* (1.78×10^{-31})
APOE	rs429358	19:45411941	2.39×10^{-42}	2.85×10^{-3}	*CTB-129P6.7*, *TOMM40*†	2.18	
APOE	rs73036519	19:45748362	3.14×10^{-7}	3.80×10^{-2}	*ZNF180*, *TOMM40*†	1.06	
C2/CFB/SKIV2L	rs116503776	6:31930462	1.17×10^{-103}	3.71×10^{-4}	*DXO*	3.09	*SKIV2L* (1.78×10^{-37})
CETP	rs5817082	16:56997349	3.56×10^{-19}	1.18×10^{-3}	*NLRC5*	2.57	*HERPUD1* (9.66×10^{-5})
CETP	rs17231506	16:56994528	2.18×10^{-18}	6.56×10^{-3}	*HERPUD1*	1.81	*HERPUD1* (9.66×10^{-5})
COL8A1	rs55975637	3:99419853	1.30×10^{-8}	1.32×10^{-2}	*NIT2*	1.51	*TOMM70* (2.55×10^{-2})
CFH	rs10922109	1:196704632	9.6×10^{-618}	7.44×10^{-3}	*KCNT2*	1.76	*KCNT2* (1.04×10^{-20})
CFH	rs570618	1:196657064	2.0×10^{-590}	1.42×10^{-2}	*CFH*	1.48	*KCNT2* (1.04×10^{-20})

CFH	rs187328863	1:196380158	1.06×10^{-68}	2.63×10^{-2}	*ZBTB41*	1.22	*KCNT2* (1.04×10^{-20})
CFH	rs61818925	1:196815450	6.03×10^{-165}	3.21×10^{-1}	*ZBTB41*	0.24	*KCNT2* (1.04×10^{-20})
CNN2	rs67538026	19:1031438	2.58×10^{-8}	9.21×10^{-11}*	*TMEM259*	9.87	
RAD51B	rs61985136	14:68769199	1.56×10^{-10}	2.15×10^{-2}	*PIGH*	1.30	*TMEM229B* (2.64×10^{-2})
RAD51B	rs2842339	14:68986999	1.36×10^{-6}	5.60×10^{-2}	*ZFYVE26*	0.90	*TMEM229B* (2.64×10^{-2})
*MMP9******	NA	NA	NA	NA	NA	NA	*PLTP* (3.3×10^{-2})

*eQTL is significant after correction for multiple testing. **Retina-specific. ***Lead SNP not present in the dataset, and suitable proxy SNPs are not available. †Gene is target of causal variant identified by eCAVIAR. ‡Low TWAS model fit ($R^2 < 0.01$). All results are based on eQTL analysis using 406 donor retinas.

Supplementary References

1. Chomczynski, P. A reagent for the single-step simultaneous isolation of RNA, DNA and proteins from cell and tissue samples. *Biotechniques* **15**, 532-534, 536-537 (1993).
2. Bolger, A.M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120 (2014).
3. Zerbino, D.R.*, et al.* Ensembl 2018. *Nucleic Acids Res* **46**, D754-D761 (2018).
4. Dobin, A.*, et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).
5. Wang, L., Wang, S. & Li, W. RSeQC: quality control of RNA-seq experiments. *Bioinformatics* **28**, 2184-2185 (2012).
6. Eisenberg, E. & Levanon, E.Y. Human housekeeping genes, revisited. *Trends Genet* **29**, 569-574 (2013).
7. Li, B. & Dewey, C.N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
8. Robinson, M.D., McCarthy, D.J. & Smyth, G.K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139-140 (2010).
9. Robinson, M.D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol* **11**, R25 (2010).
10. Robinson, M.D. & Smyth, G.K. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics* **23**, 2881-2887 (2007).
11. Roberts, A., Pimentel, H., Trapnell, C. & Pachter, L. Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics* **27**, 2325-2329 (2011).
12. Haas, B.J.*, et al.* De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc* **8**, 1494-1512 (2013).
13. Wang, L.*, et al.* CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Res* **41**, e74 (2013).
14. Pruitt, K.D.*, et al.* The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res* **19**, 1316-1323 (2009).
15. Harrow, J.*, et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* **22**, 1760-1774 (2012).
16. Howie, B.N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* **5**, e1000529 (2009).
17. 1000 Genomes Project Consortium*, et al.* A global reference for human genetic variation. *Nature* **526**, 68-74 (2015).
18. Patterson, N., Price, A.L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet* **2**, e190 (2006).
19. Price, A.L.*, et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**, 904-909 (2006).
20. Fritsche, L.G.*, et al.* A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants. *Nat Genet* **48**, 134-143 (2016).
21. Mele, M.*, et al.* Human genomics. The human transcriptome across tissues and individuals. *Science* **348**, 660-665 (2015).
22. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B-Statistical Methodology* **67**, 301-320 (2005).
23. Tibshirani, R. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society Series B-Methodological* **58**, 267-288 (1996).

24. Gusev, A.*, et al.* Integrative approaches for large-scale transcriptome-wide association studies. *Nat Genet* **48**, 245-252 (2016).
25. Shabalin, A.A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**, 1353-1358 (2012).
26. Strunz, T.*, et al.* A mega-analysis of expression quantitative trait loci (eQTL) provides insight into the regulatory architecture of gene expression variation in liver. *Sci Rep* **8**, 5865 (2018).
27. Consortium, G. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648-660 (2015).
28. Consortium, G.T.*, et al.* Genetic effects on gene expression across human tissues. *Nature* **550**, 204-213 (2017).
29. Pruim, R.J.*, et al.* LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* **26**, 2336-2337 (2010).
30. Machiela, M.J. & Chanock, S.J. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics* **31**, 3555-3557 (2015).