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]*

Genome-wide association studies (GWAS) have identified genetic variants at 34 loci contributing to age-related macular degeneration (AMD)[1-3]. We generated transcriptional profiles of postmortem retinas from 453 controls and cases at distinct stages of AMD and integrated retinal transcriptomes, covering 13,662 protein-coding and 1,462 noncoding genes, with genotypes at more than 9 million common SNPs for expression quantitative trait loci (eQTL) analysis of a tissue not included in Genotype-Tissue Expression (GTEx) and other large datasets[4,5]. Cis-eQTL analysis identified 10,474 genes under genetic regulation, including 4,541 eQTLs detected only in the retina. Integrated analysis of AMD-GWAS with eQTLs ascertained likely target genes at six reported loci. Using transcriptome-wide association analysis (TWAS), we identified three additional genes, *RLBP1*, *HIC1* and *PARP12*, after Bonferroni correction. Our studies expand the genetic landscape of AMD and establish the Eye Genotype Expression (EyeGEx) database as a resource for post-GWAS interpretation of multifactorial ocular traits.

AMD, a leading cause of incurable vision impairment, results in progressive loss of photoreceptors, particularly in the macular region of the retina[1]. AMD-GWAS have identified strong and highly replicated association of 52 independent SNPs at 34 genetic loci accounting for more than 50% of the heritability[3]. To derive mechanistic insights and further advance AMD genetics, we initiated the EyeGEx project to elucidate genetic regulation of gene expression in the human retina. We characterized 523 postmortem retinas from 517 donors by using the Minnesota Grading System (MGS)[6], with criteria similar to the Age-related Eye Disease Study (AREDS)[7] (Supplementary Fig. 1 and Supplementary Data 1). MGS1 donor retinas demonstrated no AMD features and served as controls, whereas MGS2 to MGS4 samples represented progressively more severe disease stages.

RNA-seq of the donor retinas provided 32.5 million (median) uniquely mapped paired-end reads per sample, with a 94% mapping rate to Ensembl release GRCh38.p7 (Supplementary Fig. 2).

After RNA-seq quality control (Supplementary Note), 105 MGS1, 175 MGS2, 112 MGS3, and 61 MGS4 samples were selected for further analyses. The reference-transcriptome profile was generated from MGS1 control retinas (Fig. 1a and Supplementary Data 2) and included 67% of the protein-coding genes (13,662) and 6.7% of the noncoding genes (1,462) in Ensembl, in agreement with findings from a previous study[8]. High-abundance genes (186 genes showing ≥100 fragments per kilobase of transcript per million mapped reads (FPKM)) accounted for half of the Ensembl-annotated transcripts in our RNA-seq data and were enriched in visual perception, metabolic processes, and energy homeostasis (Supplementary Fig. 3a and Supplementary Data 2). Overall, 34% of the retinal transcripts were of mitochondrial origin (Fig. 1a and Supplementary Fig. 3b), thus reflecting the high concentration of mitochondria in photoreceptors[9], the predominant cell type in the human retina[10].

Genome-guided transcript assembly supplemented 410 putative novel long intergenic noncoding RNAs (lincRNAs) and 2,861 protein-coding isoforms of genes expressed in the retina (Supplementary Fig. 3c and Supplementary Data 2). The putative lincRNA isoforms were not enriched in any biological pathway. In contrast, predicted gene function and classification of novel protein-coding isoforms showed enrichment in Gene Ontology (GO) biological processes involving synapse structure or activity (adjusted P value = 1.37×10^{-2}), sensory perception (adjusted P value = 1.64×10^{-2}), regulation of membrane potential (adjusted P value = 3.45×10^{-2}), and photoreceptor maintenance (adjusted P value = 3.45×10^{-2}). The multidimensional scaling plot of the retina reference transcriptome against the GTEx v7 data distinguished tissue-specific clusters consistent with the defined biological replicates, whereas tissue hierarchical clustering on the mean gene expression levels revealed a high degree of similarity between the brain and retina (Fig. 1b, Supplementary Fig. 3d and Supplementary Fig. 4). We identified 247 genes with tenfold or higher expression in the retina than in at least 42 of the 53 GTEx (v7) tissues (Supplementary Data 2).

Mapping of cis-eQTLs (as defined by SNP–gene combination within ±1 Mb of the transcriptional start site of each gene; Methods)
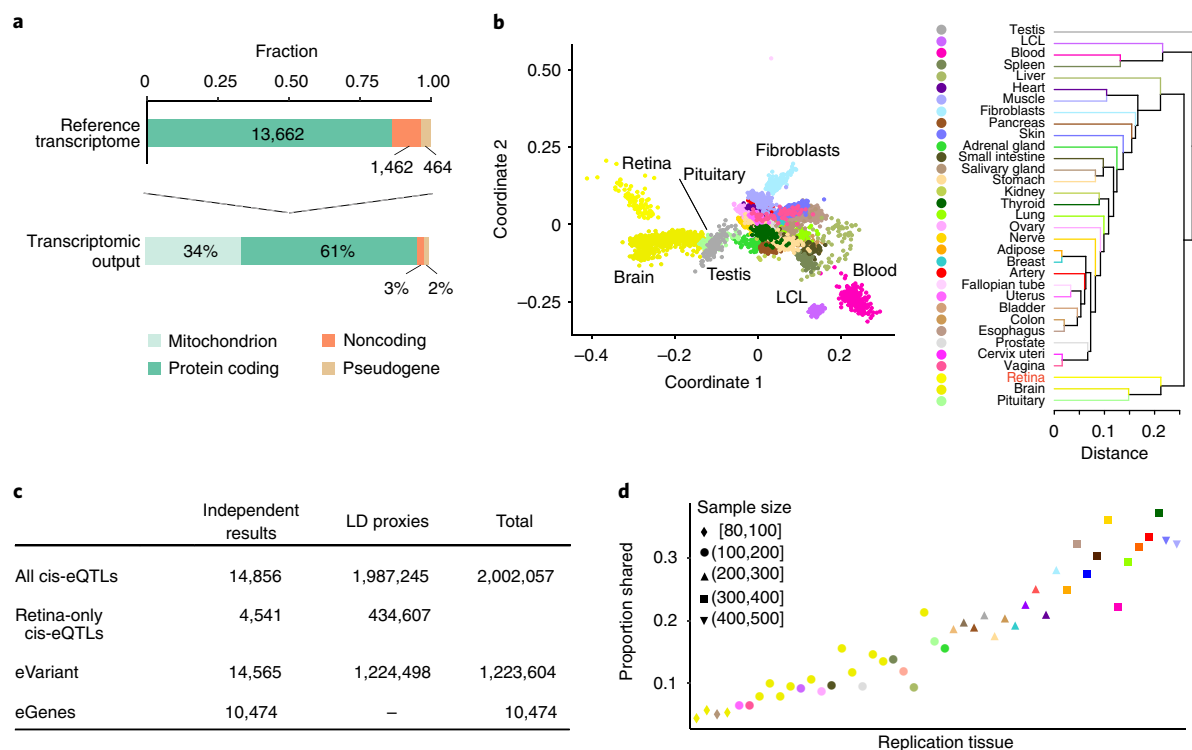
Fig. 1 | EyeGEx: retinal transcriptome and eQTL analyses. a, Reference-transcriptome output from 105 MGS1 control donor retinas. Top, fraction of expressed genes in Ensembl gene biotypes. Bottom, percentage of gene expression in distinct gene subtypes. **b**, Within-tissue sample similarity and transcriptome comparison across the retina ($n=105$ MGS1 retinas) and the GTEx tissues (v7) ($n=6,421$ samples across all body sites), on the basis of normalized gene expression levels. Each color represents a distinct tissue. Left, multidimensional scaling. Right, tissue hierarchical clustering. LCL, lymphoblastoid cell line. **c**, A summary of retinal cis-eQTLs, eGenes, and eVariants. 1.8% of the top eVariants (14,565) regulate more than one eGene. Variants in LD with the most significant eVariant are indicated as LD proxies. **d**, The proportion of cis-eQTLs in the retina (y axis) that are detected in GTEx (x axis), ordered by the sample size for each tissue. The color and shape of each point represent the tissue and sample size, respectively.

identified 14,565 genetic variants (eVariants) controlling expression of 10,474 genes (eGenes) at a false-discovery rate (FDR) ≤0.05; these included 8,529 known protein-coding and 1,358 noncoding genes (Fig. 1c and Supplementary Data 3). The strength of association was contingent on the eVariant's distance from the transcriptional start site of its corresponding eGene (Supplementary Fig. 5). Most of the retinal cis-eQTLs were present in at least one GTEx tissue, and more retinal eQTLs replicated with an increase in GTEx tissue sample size (Fig. 1d). The proportion of GTEx cis-eQTLs replicated in the retina was larger for GTEx tissues with smaller sample sizes[5] (Supplementary Fig. 5f). Almost one-third of the retina-only eQTLs observed in our study, compared with those reported by GTEx for other tissues, were attributable to the relatively larger sample size (Supplementary Fig. 6a,b).

We examined the global roles of eQTLs in the genetics of AMD. Quantile–quantile plots identified cis-eQTL SNPs to be enriched for AMD associations, and more pronounced enrichment was observed in eVariants shared across several tissues[11,12]; this relationship remained relatively consistent across all other complex disease phenotypes examined (Supplementary Fig. 5g). We then integrated retina eQTL results with associations reported across loci identified by AMD GWAS (Supplementary Table 1). Nine lead SNPs at the GWAS loci were significant eQTLs in the retina for 19 SNP–gene associations. Similar analysis showed a comparable number of lead SNPs and eQTLs in several GTEx tissues (Supplementary Data 3). To ascertain the most likely causal variants, we applied eCAVIAR, which calculates the colocalization posterior probability to identify the variant responsible for both AMD-GWAS and retina eQTL signals, after accounting for local linkage disequilibrium (LD)

patterns. At the recommended threshold of 1% colocalization posterior probability[13], we discovered likely causal SNPs and underlying target genes at six AMD loci (Supplementary Table 1 and Fig. 2a). The lead GWAS signal at two loci (*B3GALTL* and *RDH5/CD63*) was identified as the most likely causal SNP, whereas the likely causal variant was distinct from the lead SNP at four other loci: *SLC16A8* (rs5756908), *ACAD10* (rs7398705), *TMEM97/VTN* (rs241777), and *APOE* (rs157580) (Supplementary Table 1).

We leveraged retinal eQTLs and the most recent GWAS data[3] to detect novel AMD-risk genes in a TWAS[14] using our retina transcriptome data. Gene expression was modeled with SNPs within a 1-Mb window by using mixed models, Least Absolute Shrinkage and Selection Operator (LASSO), and elastic net. The TWAS identified 61 transcriptome-wide-significant gene–AMD associations (FDR ≤0.05), which passed a gene expression model fit filter ($R^2 > 0.01$) (Supplementary Data 4). We detected 38 genes within 1 Mb of 13 AMD-GWAS loci, 28 of which passed genome-wide Bonferroni correction (Fig. 2b). TWAS analysis also identified 23 genes outside the GWAS loci (Fig. 2c); these genes were located in 16 separate regions (±1 Mb). Three of these—*RLBP1*, *PARP12*, and *HIC1*—were the only significant genes in the region and remained so even after Bonferroni correction; these genes were therefore considered the strongest new candidate AMD-associated genes (Fig. 2d). Conditional testing of the full 61 significant (FDR ≤0.05) candidates identified 47 independent signals (α=0.05). A permutation test (Methods) demonstrated two of the genes (*MTMR10* and *SH3BGR*) at least 1 Mb outside of any GWAS region, and the TWAS associations were significantly informed by eQTL data after Bonferroni correction for the number of genes permuted (α=0.05;
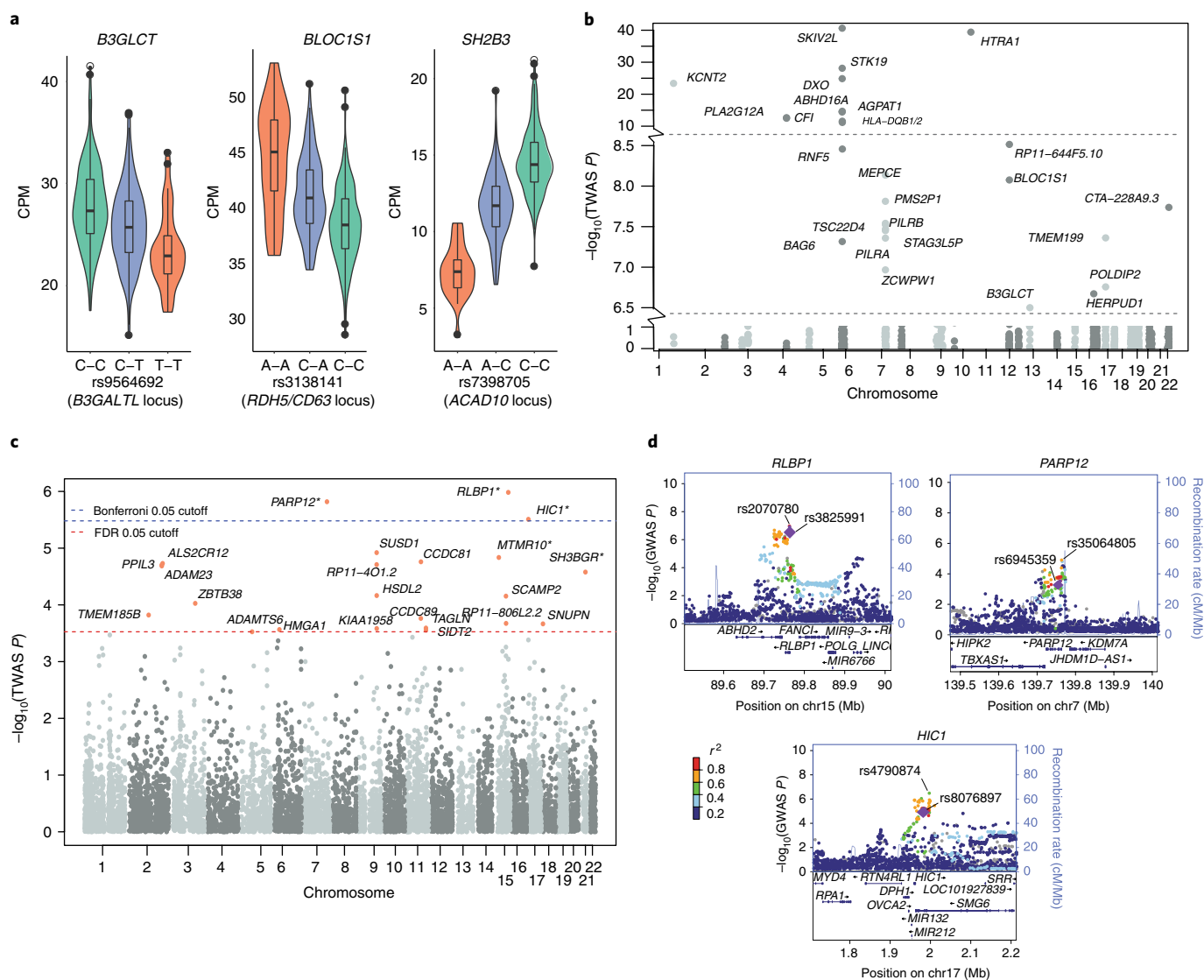
Fig. 2 | Genes and variants associated with AMD, on the basis of retina eQTL data (n = 406 retinas) and summary-level AMD-GWAS data (based on z scores of two-sided t tests on 33,976 individuals)[3]. **a,** Violin plots of the relationship between the variant at a GWAS locus and the target gene identified by eCAVIAR. At three loci, the target gene shown was the only one significantly associated (FDR ≤0.05) by TWAS. The y axis represents the distribution of expression levels (counts per million, CPM) of each gene, whereas the x axis shows the genotype (orange, homozygous minor allele; green, homozygous major allele; blue, heterozygous) for a given SNP. Box plots depict the median (thick black horizontal bar), the interquartile range, and minimum and maximum CPM values. **b,** TWAS results (n = 406 retinas) for genes passing Bonferroni-corrected significance, identified within 1 Mb on either side of the lead SNP at previously reported GWAS loci. *PLEKHA1* (TWAS P value = 7.91×10⁻¹¹⁹) was omitted for appropriate scaling, and the horizontal lines indicate a y-axis break. **c,** Manhattan plot of TWAS-identified genes outside the reported lead SNP (>1 Mb on either side) at the GWAS loci. Of the genes with expression model R^2 > 0.01, 23 genes met the FDR threshold of 0.05 (red line), and three of these (shown by asterisks) passed Bonferroni-corrected significance (cutoff shown as blue line). **d,** LocusZoom plots showing empirical GWAS association for the top three TWAS signals outside GWAS loci. The diamonds indicate the top eVariants for independent eQTL signals. The coloration of the points is determined by their LD with respect to the eQTL in purple. The top GWAS variant in the region is also labeled. The recombination rate is shown as a blue line. Chr, chromosome.

Supplementary Data 4). However, we note that the test is overly con-servative in the presence of LD.

We compared the data from eQTL, eCAVIAR, and TWAS to highlight the most plausible target genes; *B3GLCT* and *BLOC1S1* were each identified as the only target gene at two AMD loci by all three methods, whereas *SH2B3*, *PLA2G12A*, *PILRB*, and *POLDIP2/TMEM199* were likely targets at four additional loci iden-tified by two methods (Table 1 and Supplementary Fig. 7). A com-parison of these findings with those reported in GTEx[5,15] showed that the contribution of these SNPs to gene regulation varied across different tissues (Supplementary Data 3 and Supplementary

Note 3.4). Specifically, no single nonretina tissue showed replica-tion of the retinal findings for all SNP–target gene combinations (Supplementary Data 3).

Differential expression analysis of retinal transcriptomes identi-fied 14 genes with and 161 genes without age correction in advanced AMD (FDR ≤0.20) (Supplementary Data 5 and Supplementary Fig. 8a). Thus, similarly to results for other complex diseases[16,17], our differential expression analysis did not detect many gene expres-sion changes, probably because of heterogeneity caused by aging, polygenic inheritance, and environmental factors. We then exam-ined biological pathways by gene set enrichment analysis (GSEA).

Table 1 | Significant target genes and variants for AMD susceptibility at GWAS loci after eQTL, eCAVIAR, and TWAS analyses

AMD locus	Lead GWAS SNP	Chromosome: position	GWAS P	eQTL P	Target gene(s)	Percentage variability explained	Significant TWAS gene at the locus (FDR)
B3GALTL	rs9564692	13: 31821240	3.31×10^{-10}	2.36×10^{-11a}	B3GLCT[b]	10.47	B3GLCT (1.34×10^{-4})
RDH5/CD63	rs3138141	12: 56115778	4.3×10^{-9}	5.69×10^{-19a}	BLOC1S1[b]	17.8	BLOC1S1 (7.06×10^{-6})
ACAD10	rs61941274	12: 112132610	1.07×10^{-9}	8.95×10^{-2}	SH2B3[b]	0.71	SH2B3 (0.0217)
CFI	rs10033900	4: 110659067	5.35×10^{-17}	3.98×10^{-7a}	PLA2G12A	6.17	CFI (3.01×10^{-10}), PLA2G12A (4.30×10^{-10})
PILRB/PILRA	rs7803454	7: 99991548	4.76×10^{-9}	3.57×10^{-77a}	PILRB, PILRA, ZCWPW1, TSC22D4	57.51	MEPCE (6.51×10^{-6}), PILRB (2.06×10^{-5})
TMEM97/VTN	rs11080055	17: 26649724	1.04×10^{-8}	8.37×10^{-19a}	POLDIP2, SLC13A2[c], TMEM199[b]	17.65	TMEM199 (2.55×10^{-5}), POLDIP2 (8.60×10^{-5})

[a]eQTL is significant after correction for multiple testing. [b]Target of causal variant identified by eCAVIAR. [c]Retina-specific eQTL. Only protein-coding genes are shown. *B3GLCT* is the new gene symbol for *B3GALTL*. *SH2B3* was identified by GWAS colocalization (eCAVIAR) and TWAS, two of the three criteria used to identify target genes in our study. Despite its high eQTL *P* value, *SH2B3* is an excellent biological candidate for AMD because of its association with inflammation. eQTL analysis was based on 406 postmortem donor retina samples.

Immune-regulation and cholesterol-metabolism pathways, previously implicated in GWAS[3], were upregulated in early and advanced AMD, whereas pathways associated with synapse development and function were largely and exclusively downregulated in intermediate AMD (Supplementary Data 5). We note that most of the genes within susceptibility loci for advanced AMD did not appear to be associated with intermediate AMD despite sufficient power[3]. Thus, intermediate AMD may not be a transitional stage between early and advanced AMD but a separate entity with unique and distinct genetic underpinnings that require further exploration. Furthermore, weighted gene-coexpression network analysis of all samples suggested that several of the pathways implicated in AMD operate through closely connected networks in the retina (Supplementary Fig. 8b,c and Supplementary Data 6).

GWAS have successfully identified variants at hundreds of loci that contribute to health- and disease-associated traits, thereby defining their broad genetic architecture[18,19]. Interpretation of GWAS findings, however, remains a major challenge, because a large proportion of associated variants are not in protein-coding genomic regions, and their effects on specific phenotypes often individually appear to be small[20,21]. eQTL analysis in disease-relevant tissues appears to be a prominent tool for biological interpretation of GWAS loci[11,22]. Owing to the large sample size, we were able to identify 14,856 eQTLs that modulate retinal gene expression, a substantial proportion of which are not reported in GTEx v7 data. Moreover, we connected the lead GWAS signal to specific target genes at six known AMD-associated loci by at least two lines of evidence. Two of the target genes were validated by three independent methods: *B3GLCT* encodes a glucosyltransferase[23], and its loss of function leads to Peters plus syndrome[24]; *BLOC1S1* encodes a subunit of a multiprotein complex associated with the biogenesis of an organelle of the endosome–lysosome system[25], and its altered function can affect synaptic function[26]. Thus, altered expression of *B3GLCT* and *BLOC1S1* might affect extracellular-matrix stability or signaling and the degradation of unwanted/recycled proteins, respectively, thereby contributing to AMD pathogenesis. We attribute the lack of obvious target genes at the remaining AMD-GWAS loci to multiple factors, including LD structure, variants affecting expression in trans or in other AMD-relevant tissues (such as retinal pigment epithelium and choroid), and the power of this study. Interpretation of eVariants that regulate multiple genes at a particular locus requires further biological validation.

AMD is notable among complex traits because of its high heritability and large effect sizes for individual GWAS SNPs[3]. We show that variants associated with gene expression across many tissues as eQTLs, as opposed to those with only tissue-specific associations, are enriched in AMD associations despite high tissue specificity of the disease itself (Supplementary Data 3 and Supplementary Fig. 5g). We hypothesize that, at least in part, such associations reflect larger, more robust effects among the shared eQTLs. Not surprisingly, the retina is the only tissue for which we detected regulation consistently across all six identified SNPs (Supplementary Data 3). In addition, 36 of the 61 retina-identified TWAS candidates were significant (FDR ≤0.05) in at least one GTEx tissue. The remaining candidates could not be analyzed because they had no expression or heritability in the GTEx tissues, or they were not replicated in any other tissue. Our results corroborate findings from recent studies[12,27] and suggest that the best way to increase power for gene discovery through TWAS and similar approaches is to increase the diversity of tissues for greater resolution of the effects of regulatory variants. We emphasize, however, that eQTL effects detected only in a tissue without biological relevance, but not in a relevant tissue, would be difficult to interpret for disease-specific phenotypes. Although other tissues may show same eQTLs, the retinal effects of eQTLs are more likely to be directly relevant. We suggest that eQTL analyses of retinal pigment epithelium and choroid would further contribute to the understanding of genes involved in AMD pathobiology. AMD-associated genes uncovered by TWAS provide additional insights into the relevance of gene regulation to phenotypic consequences in this complex disease.

EyeGEx complements the GTEx project and provides a reference for the biological interpretation of genetic variants associated with common ocular traits, including glaucoma and diabetic retinopathy. Comparative analysis of retinal transcriptomes and eQTLs with the GTEx data should assist in exploring biological questions relating to visual function in syndromic and multifactorial traits.

URLs. 1000 Genomes Project reference panel, http://www.internationalgenome.org/; Retinal Information Network (RetNet), https://sph.uth.edu/retnet/; GTEx, https://www.gtexportal.org/home/; Gene Ontology structure, http://www.informatics.jax.org/vocab/gene_ontology/; HMMER, http://hmmer.org/; FastQC, http://www.bioinformatics.babraham.ac.uk/projects/fastqc/; precomputed TWAS weights, http://gusevlab.org/projects/fusion/; NEI Commons, https://neicommons.nei.nih.gov/#/; Biowulf Linux cluster: https://hpc.nih.gov/.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41588-019-0351-9.

References

1. Fritsche, L. G. et al. Age-related macular degeneration: genetics and biology coming together. *Annu. Rev. Genomics Hum. Genet.* **15**, 151–171 (2014).
2. Grassmann, F., Ach, T., Brandl, C., Heid, I. M. & Weber, B. H. F. What does genetics tell us about age-related macular degeneration? *Annu. Rev. Vis. Sci.* **1**, 73–96 (2015).
3. 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).
4. Small, K. S. et al. Identification of an imprinted master trans regulator at the *KLF14* locus related to multiple metabolic phenotypes. *Nat. Genet.* **43**, 561–564 (2011).
5. Battle, A., Brown, C. D., Engelhardt, B. E. & Montgomery, S. B. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
6. Olsen, T. W. & Feng, X. The Minnesota Grading System of eye bank eyes for age-related macular degeneration. *Invest. Ophthalmol. Vis. Sci.* **45**, 4484–4490 (2004).
7. Ferris, F. L. et al. A simplified severity scale for age-related macular degeneration: AREDS Report No. 18. *Arch. Ophthalmol.* **123**, 1570–1574 (2005).
8. Pinelli, M. et al. An atlas of gene expression and gene co-regulation in the human retina. *Nucleic Acids Res.* **44**, 5773–5784 (2016).
9. Hoang, Q. V., Linsenmeier, R. A., Chung, C. K. & Curcio, C. A. Photoreceptor inner segments in monkey and human retina: mitochondrial density, optics, and regional variation. *Vis. Neurosci.* **19**, 395–407 (2002).
10. Curcio, C. A., Sloan, K. R., Kalina, R. E. & Hendrickson, A. E. Human photoreceptor topography. *J. Comp. Neurol.* **292**, 497–523 (1990).
11. Finucane, H. K. et al. Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat. Genet.* **50**, 621–629 (2018).
12. Gamazon, E. R. et al. Using an atlas of gene regulation across 44 human tissues to inform complex disease- and trait-associated variation. *Nat. Genet.* **50**, 956–967 (2018).
13. Hormozdiari, F. et al. Colocalization of GWAS and eQTL signals detects target genes. *Am. J. Hum. Genet.* **99**, 1245–1260 (2016).
14. Gusev, A. et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* **48**, 245–252 (2016).
15. 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).
16. Fromer, M. et al. Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nat. Neurosci.* **19**, 1442–1453 (2016).
17. Gandal, M. J. et al. Shared molecular neuropathology across major psychiatric disorders parallels polygenic overlap. *Science* **359**, 693–697 (2018).
18. Beck, T., Hastings, R. K., Gollapudi, S., Free, R. C. & Brookes, A. J. GWAS Central: a comprehensive resource for the comparison and interrogation of genome-wide association studies. *Eur. J. Hum. Genet.* **22**, 949–952 (2014).
19. MacArthur, J. et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45**, D896–D901 (2017).
20. Chakravarti, A., Clark, A. G. & Mootha, V. K. Distilling pathophysiology from complex disease genetics. *Cell* **155**, 21–26 (2013).
21. Gallagher, M. D. & Chen-Plotkin, A. S. The post-GWAS era: from association to function. *Am. J. Hum. Genet.* **102**, 717–730 (2018).
22. Brown, C. D., Mangravite, L. M. & Engelhardt, B. E. Integrative modeling of eQTLs and cis-regulatory elements suggests mechanisms underlying cell type specificity of eQTLs. *PLoS Genet.* **9**, e1003649 (2013).
23. Kozma, K. et al. Identification and characterization of aβ1,3-glucosyltransferase that synthesizes the Glc-β1,3-Fuc disaccharide on thrombospondin type 1 repeats. *J. Biol. Chem.* **281**, 36742–36751 (2006).
24. Lesnik Oberstein, S. A. et al. Peters Plus syndrome is caused by mutations in B3GALTL, a putative glycosyltransferase. *Am. J. Hum. Genet.* **79**, 562–566 (2006).
25. Langemeyer, L. & Ungermann, C. BORC and BLOC-1: shared subunits in trafficking complexes. *Dev. Cell* **33**, 121–122 (2015).
26. Mullin, A. P. et al. Gene dosage in the dysbindin schizophrenia susceptibility network differentially affect synaptic function and plasticity. *J. Neurosci.* **35**, 325–338 (2015).
27. Hormozdiari, F. et al. Leveraging molecular quantitative trait loci to understand the genetic architecture of diseases and complex traits. *Nat. Genet.* **50**, 1041–1047 (2018).

Author contributions

Overall conceptualization: R.R. and A.S.; clinical and tissue resources: D.A.F., R.J.K., S.R.M., and E.Y.C.; transcriptome data: R.R., M.R.S., L.G., A.W., and A.P.; genotyping data: L.G.F. and G.R.A.; bioinformatic analysis: M.R.S., R.R., M.K., and A.W.; eQTL analysis: O.A.S., N.C., M.A., and A.B.; statistical supervision: N.C.; data curation: M.R.S.; writing original draft: R.R., M.R.S., O.A.S., M.K., N.C., and A.S.; writing, review, and editing: all authors; funding: D.A.F. and A.S.; supervision and project administration: A.S.

Competing interests

G.R.A. is now employed by Regeneron Pharmaceuticals. The other authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-019-0351-9.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to D.A.F., N.C. or A.S.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Study subjects. Postmortem human donor eyes were procured by the Minnesota Lions Eye Bank after informed consent from the donor or next of kin was obtained, in accordance with the tenets of the Declaration of Helsinki. These studies were approved by the institutional review boards of the University of Minnesota and National Eye Institute, National Institutes of Health. Exclusion criteria for donors included a history of diabetes or glaucoma. Donors were also excluded from this study if, upon examination of donor macular images, there were clinical symptoms of diabetic retinopathy, advanced glaucoma, myopic degeneration, or the presence of atypical debris in the eyes. Donor eyes were enucleated within 4 h of death and stored in a moist chamber at 4 °C until retinal dissection was performed. Dissection and classification of donor retinas for AMD were carried out according to the four-step MGS as previously described[6,28]. Tissue sections were flash frozen in liquid nitrogen and stored at −80 °C until further processing. Samples with ambiguous or no MGS levels were excluded from downstream analysis. Details of donor characteristics are described in the Supplementary Note.

GTEx data. RNA-seq and genotyping data from GTEx release v7 were downloaded from the Database of Genotypes and Phenotypes (dbGaP) under accession phs000424.v7.p2 and from the GTEx portal (see URLs), respectively.

RNA-seq, genotyping, and quality control. Details of RNA-seq, genotyping, and quality control are provided in the Supplementary Note.

Batch correction. Surrogate variables were identified and estimated for known batch effects as well as latent factors by using the supervised SVA (SSVA version 3.28.0/3.24.4) method[29–31] based on the following model:

$$\text{gene expression} \sim \text{MGS} + \text{sex} + \text{age}$$

Negative-control genes for SSVA were selected from a reported list of 3,804 housekeeping genes that are uniformly expressed across 16 human tissues[32]. The Pearson method was used to determine correlations between all significant surrogate variables identified by SSVA and possible sources of variation, including biological and technical factors. Known batch effects were assessed with Principal Variance Component Analysis (PVCA) (version 1.23.0) before and after batch correction[33]. All surrogate variables identified by SSVA were used for batch correction. Additional details are described in the Supplementary Note.

Reference transcriptome. We generated the control human retina transcriptome profile from 105 MGS1 retinas by applying two criteria for gene expression: the first was to remove weakly expressed genes across all MGS stages (i.e., ≥1 CPM in ≥10% of all 453 samples), and the second was to describe the transcriptomic landscape in the retina with greater confidence (i.e., ≥2 CPM in ≥50% of all 105 MGS1 samples). We calculated the cumulative transcriptional output as previously defined[34] by converting CPM into FPKM values to take gene length into account. Similarities in transcriptomes between the retina and 53 GTEx tissues were observed with a gene filter of ≥1 CPM in ≥10% of all samples across all tissues, whereas a different gene filter, namely ≥1 CPM in ≥10% of samples within each tissue, was applied to identify genes whose expression was at least tenfold higher in the retina than in other tissues. Pathway enrichment analysis was performed with GO biological-process terms[35,36] within clusterProfiler version 3.4.4 (ref. [37]) by using a Benjamini–Hochberg-adjusted P value ≤0.05 as the significance threshold. The analysis and classification of potentially novel isoforms of known genes and unknown intergenic transcripts were performed with the Cufflinks suite, version 2.21 (refs. [38,39]). Further details are provided in the Supplementary Note.

Comparison of transcriptomes across retina and GTEx tissues. Raw GTEx v7 RNA-seq data were analyzed through our bioinformatics pipeline as described above for the retina. Effects due to differences in bioinformatics pipelines between our analysis and that of GTEx were compared, as described in the Supplementary Note.

cis-eQTL mapping. The analysis included 406 individuals for whom genotype and retina gene expression data were available, 17,389 genes expressed at ≥1 CPM in at least 10% of the retina samples, and 8,924,684 genotyped and imputed common variants. Cis-eQTL analysis was conducted with QTLtools version 1.0 (ref. [40]), with a linear model to adjust for disease status (MGS level), age, sex, population stratification (ten principal components), and batch effects (21 surrogate variables). In the first step of the analysis, the variant most associated with each gene was selected, and then permutation was used to determine the distribution of its test statistic under the null. This procedure was subsequently used to obtain the P value for each gene. These P values were adjusted for multiple testing with the q-value approach[41] at the desired type I–error level. The second step of the analysis involved the identification of all eVariants with independent effects on a given eGene (significant gene from the first stage). This step was done by using the gene-level thresholds derived from the first stage and then identifying which variants exhibited nominal P values below these thresholds, on the basis of the forward–backward stepwise regression algorithm.

GTEx comparison. To calculate π_1, we compared our cis-eQTL discoveries by using the following definition:

$\pi_1 = P(\text{cis-eQTL in discovery tissue is significant in replication tissue} \mid \text{cis-eQTL in discovery tissue was also analyzed in the replication tissue})$

Thus, for each cis-eQTL (gene-variant combination) we required that the combination be analyzed in both tissues being compared.

GWAS lead-variant analysis. Forty-one lead variants from AMD-GWAS[3] were analyzed. Those not found either were not in the reference dataset used for imputation (six variants) or did not pass our MAF threshold of <1% (five variants). Matrix eQTL version 2.1.1 (ref. [42]) was then used to obtain the marginal associations by using the same cis criteria, which were then corrected for multiple testing only for the number of variants tested, by using the Bonferroni method with a type I error rate of 5%.

Enrichment. In general, we processed quantile–quantile plots for each GWAS dataset by removing all SNPs within ±1 Mb of the known GWAS signals and subsetting to variants with a MAF of at least 5%, after removing variants in the major histocompatibility region. The remaining variants were then grouped according to eQTL characteristics. Details can be found in the Supplementary Note.

Colocalization. Likely colocalizing variants between the eQTL and the GWAS data were identified with eCAVIAR version 2.0 (ref. [13]) (Supplementary Note) on the basis of marginal statistics from the cis-eQTL analysis and from AMD GWAS[3].

TWAS. To perform the TWAS, the log-transformed, SSVA-corrected expression data from the 406 samples in our dataset that both passed RNA-seq and genotyping quality control were inverse-normal transformed (rank offset = 3/8)[42] to moderate the influence of potential outliers. Expression was then controlled for sex, age, and the ten population-structure variables determined by Eigenstrat version 7.2.1 (refs. [44,45]). For each gene, we took the subset of SNPs within 1 Mb of its start or end site that had GWAS statistics[3] by using VCFtools version 0.1.15 (ref. [45]). TWAS implementation was performed according to Gusev et al.[14], heritability was calculated with GCTA version 1.21 (ref. [46]), and genetic control of expression was modeled with mixed models, LASSO, or elastic net ($\alpha = 0.5$), depending on which of the three methods produced the highest fivefold cross-validation R^2.

The effect sizes from these models acted as weights. Weighted z scores were summed for each gene, and this gene–trait association statistic was divided by its standard deviation while LD was accounted for between GWAS statistics. Standardized gene-level scores were tested against the standard normal distribution on both sides. The FDR was calculated to account for multiple testing across genes with calculated P values; genes that had an FDR <0.05 were considered significant. We also determined whether genes passed a 0.05 significance threshold after Bonferroni correction. Genes were then filtered according to their model expression fit; genes with a genetic model $R^2 < 0.01$ were discarded.

We also performed a permutation test to determine the role that the eQTL data played in the associations: for genes with a TWAS P value <0.001, weights were randomly assigned to SNPs, and the gene-level z scores were recomputed for an adaptive number of iterations to generate a null distribution against which the original TWAS statistic was tested[14]. Details on the methods used for the conditional TWAS test can be found in the Supplementary Note.

Differential expression. Differential expression was assessed with the limma package in R version 3.34.2 (ref. [47]) with a significance threshold of FDR ≤0.20. MGS was treated as an ordinal variable in pairwise comparisons between controls and each AMD stage. Differential expression was performed with adjustments for sex and batch effects (22 surrogate variables), with or without age as a covariate. Age was the most significant nongenetic risk factor for AMD, and age-related gene expression changes would probably be relevant to AMD. We therefore also performed differential expression analysis without correcting for age to generate a comprehensive list of candidate genes that require further investigation to ascertain their contribution to AMD pathogenesis. Additional differential expression analyses, performed after the removal of samples with conditions such as hypertension, high cholesterol, and cardiovascular disease, were consistent across all comparisons made (data not shown).

Gene set enrichment analysis and leading-edge analysis. GSEA was performed by preranking genes by significance and the direction of fold change from differential expression analysis, and then testing for association with the GO biological-process gene set deposited in the GSEA MSigDB resource version 2.2.4 (ref. [48]). Leading-edge analysis was performed on gene sets reaching a significance threshold of FDR ≤0.25 and absolute normalized enrichment score of ≥2.0. Significant gene sets were further classified into common functional categories by visualization of the GO structure as described in the Supplementary Note (see URLs).

Weighted gene-correlation network analysis. Weighted gene-correlation network analysis[49] was performed on all 453 samples that passed RNA-seq quality control,

to group genes by expression profile, with the associated software WGCNA version 1.51. The log-transformed expression values were corrected for age, sex, and batch effects (determined by SSVA[29–31]). Adjacency was calculated with Spearman correlation, and the power by which we raised the absolute values of the correlation to obtain the adjacency matrix was $k = 3$. Through hypergeometric testing, at a significance threshold of 0.05 α after Bonferroni correction for multiple testing, modules were assessed for the enrichment of the following types of genes: (i) genes deemed relevant to macular-degeneration pathogenesis in the literature, (i) genes within 500 kb of the 34 AMD loci identified through GWAS[3], and (iii) genes identified as leading edge by GSEA[48]. A list of genes relevant to AMD was obtained from a previous published study[50] and was updated through extensive PubMed searching (through December 2017) with one of several search terms (Supplementary Note). Pathway analysis was performed on each module with GO biological-process terms[35,36] through clusterProfiler version 3.4.4 (ref. [37]). The connections between genes in modules were visualized with Cytoscape version 3.5.1 (ref. [51]).

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability
The sequencing data are available at Gene Expression Omnibus (GEO) under accession code GSE115828 and NEI Commons (see URLs). The GTEx data used here were obtained from the GTEx Portal on 26 March 2018 and/or dbGaP accession number phs000424.v7.p2.

References
28. Decanini, A., Nordgaard, C. L., Feng, X., Ferrington, D. A. & Olsen, T. W. Changes in select redox proteins of the retinal pigment epithelium in age-related macular degeneration. *Am. J. Ophthalmol.* **143**, 607–615 (2007).
29. Gagnon-Bartsch, J. A. & Speed, T. P. Using control genes to correct for unwanted variation in microarray data. *Biostatistics* **13**, 539–552 (2012).
30. Leek, J. T. svaseq: removing batch effects and other unwanted noise from sequencing data. *Nucleic Acids Res.* **42**, e161 (2014).
31. Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E. & Storey, J. D. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28**, 882–883 (2012).
32. Eisenberg, E. & Levanon, E. Y. Human housekeeping genes, revisited. *Trends Genet.* **29**, 569–574 (2013).
33. Scherer, A. (ed.) *Batch Effects and Noise in Microarray Experiments: Sources and Solutions* (Wiley, West Sussex, UK, 2009).
34. Melé, M. et al. The human transcriptome across tissues and individuals. *Science* **348**, 660–665 (2015).
35. Ashburner, M. et al. Gene ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
36. The Gene Ontology Consortium. Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Res.* **45**, D331–D338 (2017).
37. Yu, G., Wang, L. G., Han, Y. & He, Q. Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**, 284–287 (2012).
38. Trapnell, C. et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).
39. Roberts, A., Pimentel, H., Trapnell, C. & Pachter, L. Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics* **27**, 2325–2329 (2011).
40. Delaneau, O. et al. A complete tool set for molecular QTL discovery and analysis. *Nat. Commun.* **8**, 15452 (2017).
41. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA* **100**, 9440–9445 (2003).
42. Shabalin, A. A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**, 1353–1358 (2012).
43. Beasley, T. M., Erickson, S. & Allison, D. B. Rank-based inverse normal transformations are increasingly used, but are they merited? *Behav. Genet.* **39**, 580–595 (2009).
44. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2**, e190 (2006).
45. Price, A. L. et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
46. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
47. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
48. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
49. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
50. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).
51. Newman, A. M. et al. Systems-level analysis of age-related macular degeneration reveals global biomarkers and phenotype-specific functional networks. *Genome Med.* **4**, 16 (2012).
52. Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).

nature research

Corresponding author(s): Anand Swaroop,
Nilanjan Chatterjee,
Deborah Ferrington.

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☒	☐	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	TRIzol® was used to isolate the RNA. TruSeq® Stranded mRNA Library Preparation Kit (Illumina) was used to prepare the library. Paired-end RNAseq reads s were obtained using the HiSeq 2500 platform.
Data analysis	Trimmomatic (version 0.36), FastQC (version 0.11.5), STAR (version 2.5.2a), RSeQC (version 2.6.4), RSEM (version 1.13.1), edgeR (version 3.18.1), Cufflinks suite (version 2.21), TransDecoder (version 2.0.1), CPAT (version 1.2.2), HMMER (version 3.1.b), QTLtools (version 1.0),limma (version 3.34.2), sva (version 3.28.0/3.24.4), FUSION/TWASGCTA (version 1.26.0), WGCNA (version 1.51), Cytoscape version (3.5.1), eCAVIAR (version 2.0), VCFtools (version 0.1.15), GSEA JAVA program (version 2.2.4), clusterProfiler (version 3.4.4), EIGENSOFT (version 7.2.1), IMPUTE2 (version 2.3.1), Matrix eQTL (version 2.1.1), PVCA (version 1.23.0). In-house python and R scripts will be provided upon request.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The raw sequencing data has been submitted to Gene Expression Omnibus (GEO) (accession number: GSE115828). All raw and analyzed data will also be available at NEI Commons (https://neicommons.nei.nih.gov/#/) and at GTEx portal after the manuscript is accepted.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences   ☐ Behavioural & social sciences   ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We used 453 postmortem human retina samples for transcriptome and eQTL analyses. Previous eQTL studies have shown that this size is sufficient to detect eQTLs.
Data exclusions	We excluded some data based on pre-established quality control cretria. Details are provided in online methods and supplementary material.
Replication	We successfully replicated and extended the reference transcriptome of the retina with another published report. Additionally, we replicated eQTLs identified in GTEx in our data. All replications were successful.
Randomization	We randomized cases and controls to make sure that they were distributed evenly across library preparation dates.
Blinding	We did not perform any blinding. The information on post-mortem retina samples was essential for the analysis. For example, we used only MGS1 control samples to generate reference transcriptome. Similarly, we needed to know AMD status of samples before doing any analysis of what is affected by disease.

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Briefly describe the study type including whether data are quantitative, qualitative, or mixed-methods (e.g. qualitative cross-sectional, quantitative experimental, mixed-methods case study).
Research sample	State the research sample (e.g. Harvard university undergraduates, villagers in rural India) and provide relevant demographic information (e.g. age, sex) and indicate whether the sample is representative. Provide a rationale for the study sample chosen. For studies involving existing datasets, please describe the dataset and source.
Sampling strategy	Describe the sampling procedure (e.g. random, snowball, stratified, convenience). Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient. For qualitative data, please indicate whether data saturation was considered, and what criteria were used to decide that no further sampling was needed.
Data collection	Provide details about the data collection procedure, including the instruments or devices used to record the data (e.g. pen and paper, computer, eye tracker, video or audio equipment) whether anyone was present besides the participant(s) and the researcher, and whether the researcher was blind to experimental condition and/or the study hypothesis during data collection.
Timing	Indicate the start and stop dates of data collection. If there is a gap between collection periods, state the dates for each sample cohort.
Data exclusions	If no data were excluded from the analyses, state so OR if data were excluded, provide the exact number of exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.
Non-participation	State how many participants dropped out/declined participation and the reason(s) given OR provide response rate OR state that no participants dropped out/declined participation.

Randomization	*If participants were not allocated into experimental groups, state so OR describe how participants were allocated to groups, and if allocation was not random, describe how covariates were controlled.*

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	*Briefly describe the study. For quantitative data include treatment factors and interactions, design structure (e.g. factorial, nested, hierarchical), nature and number of experimental units and replicates.*
Research sample	*Describe the research sample (e.g. a group of tagged Passer domesticus, all Stenocereus thurberi within Organ Pipe Cactus National Monument), and provide a rationale for the sample choice. When relevant, describe the organism taxa, source, sex, age range and any manipulations. State what population the sample is meant to represent when applicable. For studies involving existing datasets, describe the data and its source.*
Sampling strategy	*Note the sampling procedure. Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient.*
Data collection	*Describe the data collection procedure, including who recorded the data and how.*
Timing and spatial scale	*Indicate the start and stop dates of data collection, noting the frequency and periodicity of sampling and providing a rationale for these choices. If there is a gap between collection periods, state the dates for each sample cohort. Specify the spatial scale from which the data are taken*
Data exclusions	*If no data were excluded from the analyses, state so OR if data were excluded, describe the exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.*
Reproducibility	*Describe the measures taken to verify the reproducibility of experimental findings. For each experiment, note whether any attempts to repeat the experiment failed OR state that all attempts to repeat the experiment were successful.*
Randomization	*Describe how samples/organisms/participants were allocated into groups. If allocation was not random, describe how covariates were controlled. If this is not relevant to your study, explain why.*
Blinding	*Describe the extent of blinding used during data acquisition and analysis. If blinding was not possible, describe why OR explain why blinding was not relevant to your study.*

Did the study involve field work? ☐ Yes ☐ No

Field work, collection and transport

Field conditions	*Describe the study conditions for field work, providing relevant parameters (e.g. temperature, rainfall).*
Location	*State the location of the sampling or experiment, providing relevant parameters (e.g. latitude and longitude, elevation, water depth).*
Access and import/export	*Describe the efforts you have made to access habitats and to collect and import/export your samples in a responsible manner and in compliance with local, national and international laws, noting any permits that were obtained (give the name of the issuing authority, the date of issue, and any identifying information).*
Disturbance	*Describe any disturbance caused by the study and how it was minimized.*

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☐	☒ Unique biological materials
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Unique biological materials

Policy information about availability of materials

Obtaining unique materials | Post-mortem human donor eyes were procured by the Minnesota Lions Eye Bank after informed consent from the donor or next of kin and in accordance with the tenets of the Declaration of Helsinki. The unidentified retina samples were obtained for genotyping and RNA analyses. Human retina tissue is difficult to obtain. We obtained small punches of postmortem retina from controls and cases from University of Minnesota (Dr. Deb Ferrington). These samples were converted into total RNA and DNA. Almost all of the RNA has been used for obtaining the transcriptome data. Similarly, the DNA used for genotyping was relatively small and has been used for other experiments.

Antibodies

Antibodies used | *Describe all antibodies used in the study; as applicable, provide supplier name, catalog number, clone name, and lot number.*

Validation | *Describe the validation of each primary antibody for the species and application, noting any validation statements on the manufacturer's website, relevant citations, antibody profiles in online databases, or data provided in the manuscript.*

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s) | *State the source of each cell line used.*

Authentication | *Describe the authentication procedures for each cell line used OR declare that none of the cell lines used were authenticated.*

Mycoplasma contamination | *Confirm that all cell lines tested negative for mycoplasma contamination OR describe the results of the testing for mycoplasma contamination OR declare that the cell lines were not tested for mycoplasma contamination.*

Commonly misidentified lines (See ICLAC register) | *Name any commonly misidentified cell lines used in the study and provide a rationale for their use.*

Palaeontology

Specimen provenance | *Provide provenance information for specimens and describe permits that were obtained for the work (including the name of the issuing authority, the date of issue, and any identifying information).*

Specimen deposition | *Indicate where the specimens have been deposited to permit free access by other researchers.*

Dating methods | *If new dates are provided, describe how they were obtained (e.g. collection, storage, sample pretreatment and measurement), where they were obtained (i.e. lab name), the calibration program and the protocol for quality assurance OR state that no new dates are provided.*

☐ Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals | *For laboratory animals, report species, strain, sex and age OR state that the study did not involve laboratory animals.*

Wild animals | *Provide details on animals observed in or captured in the field; report species, sex and age where possible. Describe how animals were caught and transported and what happened to captive animals after the study (if killed, explain why and describe method; if released, say where and when) OR state that the study did not involve wild animals.*

Field-collected samples | *For laboratory work with field-collected samples, describe all relevant parameters such as housing, maintenance, temperature, photoperiod and end-of-experiment protocol OR state that the study did not involve samples collected from the field.*

ChIP-seq

Data deposition

☐ Confirm that both raw and final processed data have been deposited in a public database such as [GEO](GEO).

☐ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links
May remain private before publication. | *For "Initial submission" or "Revised version" documents, provide reviewer access links. For your "Final submission" document, provide a link to the deposited data.*

Files in database submission	*Provide a list of all files available in the database submission.*
Genome browser session (e.g. <u>UCSC</u>)	*Provide a link to an anonymized genome browser session for "Initial submission" and "Revised version" documents only, to enable peer review. Write "no longer applicable" for "Final submission" documents.*

Methodology

Replicates	*Describe the experimental replicates, specifying number, type and replicate agreement.*
Sequencing depth	*Describe the sequencing depth for each experiment, providing the total number of reads, uniquely mapped reads, length of reads and whether they were paired- or single-end.*
Antibodies	*Describe the antibodies used for the ChIP-seq experiments; as applicable, provide supplier name, catalog number, clone name, and lot number.*
Peak calling parameters	*Specify the command line program and parameters used for read mapping and peak calling, including the ChIP, control and index files used.*
Data quality	*Describe the methods used to ensure data quality in full detail, including how many peaks are at FDR 5% and above 5-fold enrichment.*
Software	*Describe the software used to collect and analyze the ChIP-seq data. For custom code that has been deposited into a community repository, provide accession details.*

Flow Cytometry

Plots

Confirm that:

☐ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☐ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☐ All plots are contour plots with outliers or pseudocolor plots.

☐ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	*Describe the sample preparation, detailing the biological source of the cells and any tissue processing steps used.*
Instrument	*Identify the instrument used for data collection, specifying make and model number.*
Software	*Describe the software used to collect and analyze the flow cytometry data. For custom code that has been deposited into a community repository, provide accession details.*
Cell population abundance	*Describe the abundance of the relevant cell populations within post-sort fractions, providing details on the purity of the samples and how it was determined.*
Gating strategy	*Describe the gating strategy used for all relevant experiments, specifying the preliminary FSC/SSC gates of the starting cell population, indicating where boundaries between "positive" and "negative" staining cell populations are defined.*

☐ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Magnetic resonance imaging

Experimental design

Design type	*Indicate task or resting state; event-related or block design.*
Design specifications	*Specify the number of blocks, trials or experimental units per session and/or subject, and specify the length of each trial or block (if trials are blocked) and interval between trials.*
Behavioral performance measures	*State number and/or type of variables recorded (e.g. correct button press, response time) and what statistics were used to establish that the subjects were performing the task as expected (e.g. mean, range, and/or standard deviation across subjects).*

Acquisition

Imaging type(s)
Specify: functional, structural, diffusion, perfusion.

Field strength
Specify in Tesla

Sequence & imaging parameters
Specify the pulse sequence type (gradient echo, spin echo, etc.), imaging type (EPI, spiral, etc.), field of view, matrix size, slice thickness, orientation and TE/TR/flip angle.

Area of acquisition
State whether a whole brain scan was used OR define the area of acquisition, describing how the region was determined.

Diffusion MRI ☐ Used ☐ Not used

Preprocessing

Preprocessing software
Provide detail on software version and revision number and on specific parameters (model/functions, brain extraction, segmentation, smoothing kernel size, etc.).

Normalization
If data were normalized/standardized, describe the approach(es): specify linear or non-linear and define image types used for transformation OR indicate that data were not normalized and explain rationale for lack of normalization.

Normalization template
Describe the template used for normalization/transformation, specifying subject space or group standardized space (e.g. original Talairach, MNI305, ICBM152) OR indicate that the data were not normalized.

Noise and artifact removal
Describe your procedure(s) for artifact and structured noise removal, specifying motion parameters, tissue signals and physiological signals (heart rate, respiration).

Volume censoring
Define your software and/or method and criteria for volume censoring, and state the extent of such censoring.

Statistical modeling & inference

Model type and settings
Specify type (mass univariate, multivariate, RSA, predictive, etc.) and describe essential details of the model at the first and second levels (e.g. fixed, random or mixed effects; drift or auto-correlation).

Effect(s) tested
Define precise effect in terms of the task or stimulus conditions instead of psychological concepts and indicate whether ANOVA or factorial designs were used.

Specify type of analysis: ☐ Whole brain ☐ ROI-based ☐ Both

Statistic type for inference
(See Eklund et al. 2016)
Specify voxel-wise or cluster-wise and report all relevant parameters for cluster-wise methods.

Correction
Describe the type of correction and how it is obtained for multiple comparisons (e.g. FWE, FDR, permutation or Monte Carlo).

Models & analysis

n/a	Involved in the study
☐	☐ Functional and/or effective connectivity
☐	☐ Graph analysis
☐	☐ Multivariate modeling or predictive analysis

Functional and/or effective connectivity
Report the measures of dependence used and the model details (e.g. Pearson correlation, partial correlation, mutual information).

Graph analysis
Report the dependent variable and connectivity measure, specifying weighted graph or binarized graph, subject- or group-level, and the global and/or node summaries used (e.g. clustering coefficient, efficiency, etc.).

Multivariate modeling and predictive analysis
Specify independent variables, features extraction and dimension reduction, model, training and evaluation metrics.