npj | genomic medicine

Article

Published in partnership with CEGMR, King Abdulaziz University

https://doi.org/10.1038/s41525-025-00507-2

Integrating explainable machine learning and transcriptomics data reveals cell-type specific immune signatures underlying macular degeneration

Check for updates

Khang Ma[1], Hosei Nakajima[1], Nipa Basak[1], Arko Barman[2,3,4] & Rinki Ratnapriya[1,5]

Genome-wide association studies (GWAS) have established key role of immune dysfunction in Age-related Macular Degeneration (AMD), though the precise role of immune cells remains unclear. Here, we develop an explainable machine-learning pipeline (ML) using transcriptome data of 453 donor retinas, identifying 81 genes distinguishing AMD from controls (AUC-ROC of 0.80, CI 0.70–0.92). Most of these genes were enriched in their expression within retinal glial cells, particularly microglia and astrocytes. Their role in AMD was further strengthened by cellular deconvolution, which identified distinct differences in microglia and astrocytes between normal and AMD. We corroborated these findings using independent single-cell data, where several ML genes exhibited differential expression. Finally, the integration of AMD-GWAS data identified a regulatory variant, rs4133124 at *PLCG2*, as a novel AMD association. Collectively, our study provides molecular insights into the recurring theme of immune dysfunction in AMD and highlights the significance of glial cell differences in AMD progression.

Variation in gene expression has emerged as a significant source of phenotypic diversity among individuals and populations[1]. Additionally, human genetic studies have highlighted the critical role of gene expression dysregulation in both rare[2] and common[3] diseases. Understanding the dysregulation of gene expression in different diseases is essential for deciphering the underlying molecular mechanisms and identifying potential targets for therapeutic intervention. The cellular context has a profound influence on gene expression and regulation, emphasizing the importance of comprehensively studying transcriptome regulation in disease-relevant cells and tissues. However, the availability of disease-relevant tissues in a large number of individuals presents a significant challenge. Additionally, gene expression in humans is influenced by genetic variants, epigenetic changes, environmental factors, or a combination of these factors[4] making gene expression studies uniquely challenging to identify consistent disease-related patterns.

Age-related Macular Degeneration (AMD) is the leading cause of irreversible vision loss in people over 50 years of age[5]. It is a neurodegenerative disease that afflicts almost 10 million individuals in the United States alone and this number is expected to double by 2050[6]. AMD results from the deterioration of the photoreceptor support system, which includes the retinal pigment epithelium (RPE), Bruch's membrane (BrM), and the choroidal vasculature, leading to the death of photoreceptors primarily in the central region of the retina called macula[7]. It is a complex, multifactorial disease that is caused by the cumulative impact of genetic predisposition, environmental stress and aging[8]. Knowledge of genetic risk factors underlying AMD susceptibility has advanced rapidly with the advent of Genome-wide Association Studies (GWAS), which have successfully identified 52 independent genetic variants at 34 loci[9] establishing a strong genetic component of AMD that is mostly driven by common variants[10,11]. These findings have implicated immune, complement, cholesterol and lipid metabolism, extracellular/collagen matrix, and angiogenesis pathways in AMD pathogenesis[9]. Among them, variants identified in complement and immunoinflammatory genes such as *CFH, CFI* and *C3* have become the essential core for AMD genetics because of the high effect size associated with these genes[9–11]. Additionally, substantial clinical evidence underscores the significant involvement of immunologic processes such as the production of inflammatory molecules, recruitment of macrophages, complement activation and microglial activation in AMD pathology[12]. The majority of AMD-associated variants reside in the non-coding region of the genome mediating the disease risk through gene expression regulation in retina[13,14]

[1]Department of Ophthalmology, Baylor College of Medicine, Houston, TX, USA. [2]Data to Knowledge Lab, Rice University, Houston, TX, USA. [3]Department of Electrical & Computer Engineering, Rice University, Houston, TX, USA. [4]Department of Statistics, Rice University, Houston, TX, USA. [5]Department of Biochemistry and Molecular Pharmacology, Baylor College of Medicine, Houston, TX, USA. ✉e-mail: rpriya@bcm.edu

and RPE[15]. However, the molecular mechanisms underlying AMD, especially the cellular vulnerability, are poorly understood.

Recent advancements in genomics have transformed biomedical research into digitalized, data-intensive science that has broadened its application in biology and medicine. However, the scale, complexity and high information content are significant barriers in its application. These limitations have encouraged the application of machine learning (ML) methods to help make informed decisions to drive novel biological hypotheses and translate them into tangible therapeutics[16,17]. In particular, ML-based approaches have been frequently used to obtain insights related to regulatory regions of the genome and how they impact gene expression and phenotypic changes[18]. Within ophthalmology, ML has occupied a niche based on studies of the retinal fundus and optical coherence tomography (OCT) images and visual fields by achieving robust diagnosis performance in detecting various diseases including diabetic retinopathy, retinopathy of prematurity, glaucoma, macular edema and AMD[19].

Comparative transcriptome studies in disease-relevant tissues and cell-types hold great potential for identifying new genes as well as investigating mechanisms underlying the disease. However, small sample sizes and the high heterogeneity of study samples impose significant challenges in its interpretation. ML-based feature selection offers a great tool to address these limitations. Here we present the development of explainable ML models to classify the AMD based on their expression profiles of 453 samples. To the best of our knowledge, this is the first study to rigorously test gene expression data for their ability to accurately distinguish AMD from normal. We further analyzed the features selected for ML models using pathways and co-expression regulation networks. Finally, we integrate the data from AMD-GWAS and single-cell transcriptomics to identify the genes and cell types associated with AMD pathology.

Results

Feature selection and machine learning model reveals a core set of 81 AMD genes

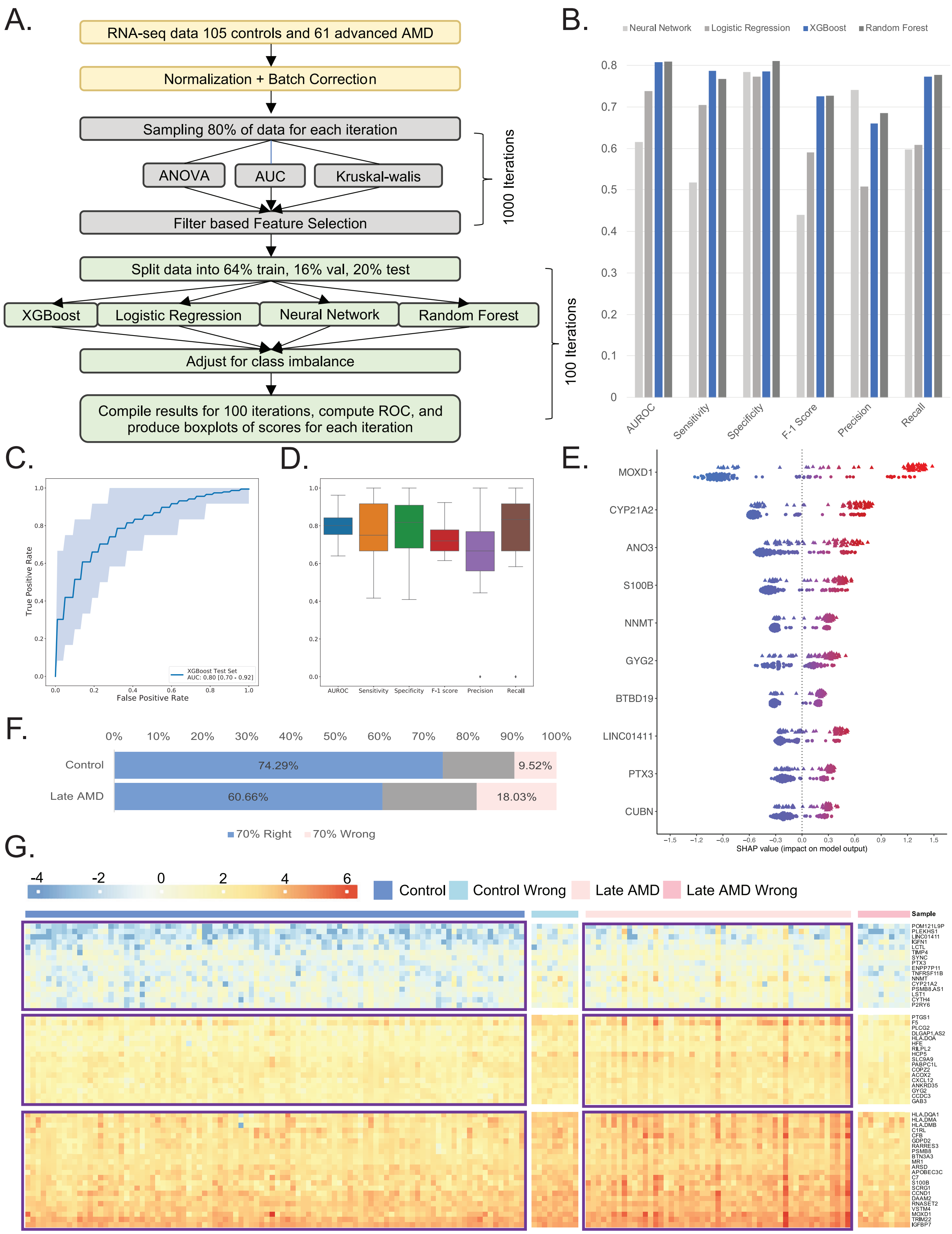
Transcriptome data often suffers from the "curse of dimensionality" as tens of thousands of genes can be profiled in a single RNA-seq experiment vs the limited number of subjects. Thus, we developed a pipeline (Fig. 1A) to reduce the dimension and improve the efficiency and interpretability of downstream analyses using 105 controls and 61 advanced AMD samples[13]. We implemented three feature selection methods, ANOVA (analysis of variance) F-test, AUC (area under the curve), and Kruskal-Wallis test to identify the most relevant features. We divided the dataset into an 80% training set and a 20% testing set. We used the training set to identify the most influential features within the training data and evaluated the model's performance on the separate 20% testing data, employing appropriate evaluation metrics. Comparing the features of top the 100 features identified across 1000 iterations selected by each method, we identified 81 genes (referred as ML-genes) that were common across three methods (Supplementary Fig. 1).

Next, we applied four ML-based models: neural network, logistic regression, eXtreme Gradient Boosting (XGB), and random forest for the classification model for AMD based on 81 ML-genes. We randomly portioned the data into 64% training (to learn potential underlying patterns), 16% validation (to tune the model's performance across different hyper-parameter choices) and 20% (to evaluate our model's prediction performance) external test sets. The optimal threshold for classification was determined by Youden's J statistic[20]. We evaluated classifier training and discrimination performance in 100 iterations of repeated randomized data splitting to ensure the robustness of the model and obtain confidence intervals. The AUC-ROC of other methods varied from 0.61 (CI 0.5–0.73) for Neural Network to 0.81 (CI 0.71–0.92) for random Forest (Fig. 1B) (Supplementary Fig. 2). XGB was found to perform the best with AUC-ROC statistic (0.80, CI 0.70–0.92) (Fig. 1C) and highest sensitivity (0.78) (Fig. 1D). Thus, we applied XGB for all further analyses.

To test the robustness of the 81 ML-genes identified in our study, we conducted comparisons of model performance using four additional gene lists: (1) Genes within 500KB of the 34 AMD-GWAS loci[9] (2) High confidence AMD genes, comprising genes from the 34 loci with established connections to AMD through rare variant discovery or eQTL analysis (3) Genes deemed relevant to macular degeneration pathogenesis in the literature that emerged from extensive PubMed searched as previously described[13] and (4) 48 genes identified through 1000 iterations of label shuffling, with control and AMD labels randomized. The performance of 81 features was superior compared to the genes within GWAS loci (AUC-ROC = 0.72, CI 0.58–0.84), high-confidence genes (AUC-ROC = 0.64, CI 0.50–0.77), literature (AUC-ROC = 0.69, CI 0.56–0.84) and, shuffled, i.e., permutation testing (AUC-ROC = 0.60, CI 0.45–0.75) (Supplementary Fig. 3A–D). 48 genes identified in the permutation testing showed no overlap with the set of 81 genes and performed poorly on both the true and shuffled labels (Supplementary Fig. 4). These results further emphasize the specificity of the 81 genes associated with AMD.

We next used SHAP (Shapley Additive exPlanations)[21] to explain our best transcriptome-based AMD predictions by computing the contributions of each feature (gene) to that prediction (i.e., rank feature importance on classification). Shapley values indicate the contribution of every feature, i.e., gene expression value, towards the prediction for every individual sample, i.e., patient or control, vis-a-vis an average prediction. A positive Shapley value for a feature in each sample indicates that the feature value is favoring the prediction of the sample as the disease class with the magnitude of the Shapley value indicating the strength of how much it favors the prediction. A negative Shapley value for a feature in each sample can be considered vice versa. The results showed that high gene expression of *MOXD1* in AMD (red, triangle) and low gene expression of *MOXD1* in controls (blue, dots) contributed most to the model prediction. The trend was similar for the top 10 genes (Fig. 1E).

Gene expression variation within humans arises from a complex interplay of genetic, environmental, and epigenetic factors. Furthermore, AMD manifests with a wide array of clinical presentations, encompassing both dry and wet forms, each exhibiting varying rates of disease progression and degrees of visual impairment. Thus, we next set out to identify whether such heterogeneity existed at the molecular (transcriptome) level. To achieve this, we harnessed the predictive capacity of the XGB model, training it on our dataset through 100 iterations of repeated randomized data splits. We compared the predicted labels against the actual ground truth (disease vs. control status) to uncover patterns. We systematically identified samples for which the predicted labels consistently aligned or deviated with the true labels in over 70% of instances. We categorize these samples into two groups: the "70% right" group, comprising instances where predictions align with true labels, and the "70% wrong" group, encompassing instances where predictions deviate from true labels. Notably, our analysis revealed a distinct pattern: a higher proportion of control samples (74%) exhibited accurate labeling compared to AMD samples (60%) (Fig. 1F). This discrepancy was further highlighted by the observation that nearly twice as many AMD patients were subject to mislabeling (18%) compared to only 9% of control subjects (Fig. 1F). Additionally, the performance of XGB was significantly improved (AUC-ROC = 0.94, CI 0.86–0.91) when 70% wrong samples were excluded from the analysis (Supplementary Fig. 5). This suggests that while heterogeneity exists within both groups, its manifestation is notably more pronounced within the disease population. These differences could not be attributed to age as within the same age range, there were several normal and AMD patients that were predicted accurately (Supplementary Fig. 6A). Next, we compared the top 2 risk alleles for AMD in *CFH* (Y402H; rs1061170) and *ARMS2* (A69S;rs10490924) as well as the polygenic risk scores (PRS) based on 52 known common risk factors across 34 loci[9] (Supplementary Fig. 6C, D). We observed an expected, significant difference in *CFH* and *ARMS2* risk alleles and PRS in all samples, and 70% right group. This difference was notably absent in the 70% wrong group (Supplementary Fig. 6E). These results suggest a potential involvement of genetic risk factors in shaping the molecular landscape of the disease in AMD. It's important to note that the sample size remains small within the 70% wrong group (consisting of 10 controls and 11 AMD



cases), underscoring the need for validation within larger cohorts. A heatmap of 81 ML-genes when plotted in these four groups (70% right AMD and controls, and 70% wrong AMD and controls) highlights the distinct gene expression patterns with the gene expression profiles within the 70% wrong group aligning closely with their predicted labels (Fig. 1G).

Gene co-expression network-based analysis connect the ML-genes to AMD-relevant pathways

To gain further insight into the biological significance and relationships among the 81 genes, we utilized Weighted Gene Co-expression Network Analysis (WGCNA), known for its ability to associate gene co-expression modules with specific biological functions and pathways[22]. WGCNA

Fig. 1 | A flow-chart of ML-pipeline, models performance and results of late AMD classification. A Schematic representation of Machine Learning pipeline, consisting of three main parts: normalization and batch correction, feature selection, and model building. **B** Bar plot comparing each model's statistics when used to classify between AMD cases and controls. Different colors represent the models built using logistic regression, random forest, neural networks, and XGBoost separately. **C** An ROC plot showing the performance of the XGBoost model using default parameters. The closer the curve is to the top left corner, the more accurate the model will be at classifying cases and controls. The numbers presented at the bottom right represent the averages of 100 iterations of XGBoost models. **D** Boxplot showing the distribution of statistics (AUC, Sensitivity, Specificity, F-1 score Precision and Recall) generated across the 100 iterations. **E** Feature importance plot using SHAP analysis swarm plot showing the underlying weights of the top 10 genes for each sample. For each gene, the top swarm line shows the distribution of weights for AMD samples, while the bottom lines show the same information for Controls. The x-axis represents the SHAP value score each observation has within a gene. Observations are assigned colors corresponding to the range of gene expression, with dots (controls) and triangles (cases) closer to blue indicating lower gene expression values and those closer to red signifying higher gene expression values. **F** Bar plot showing the distribution of samples being classified correctly 70% or more of the time, or samples being classified wrongly 70% or more of the time. The sections are identified as blue and pink respectively. Grey sections are for samples not making the right and wrong predictions cutoff. **G** Heatmap showing the gene expression of selected ML-genes (as rows) and 166 samples (as columns) divided into 4 groups: Controls, Controls being predicted wrong 70% or more of the time, Cases, Cases being predicted wrong 70% or more of the time.

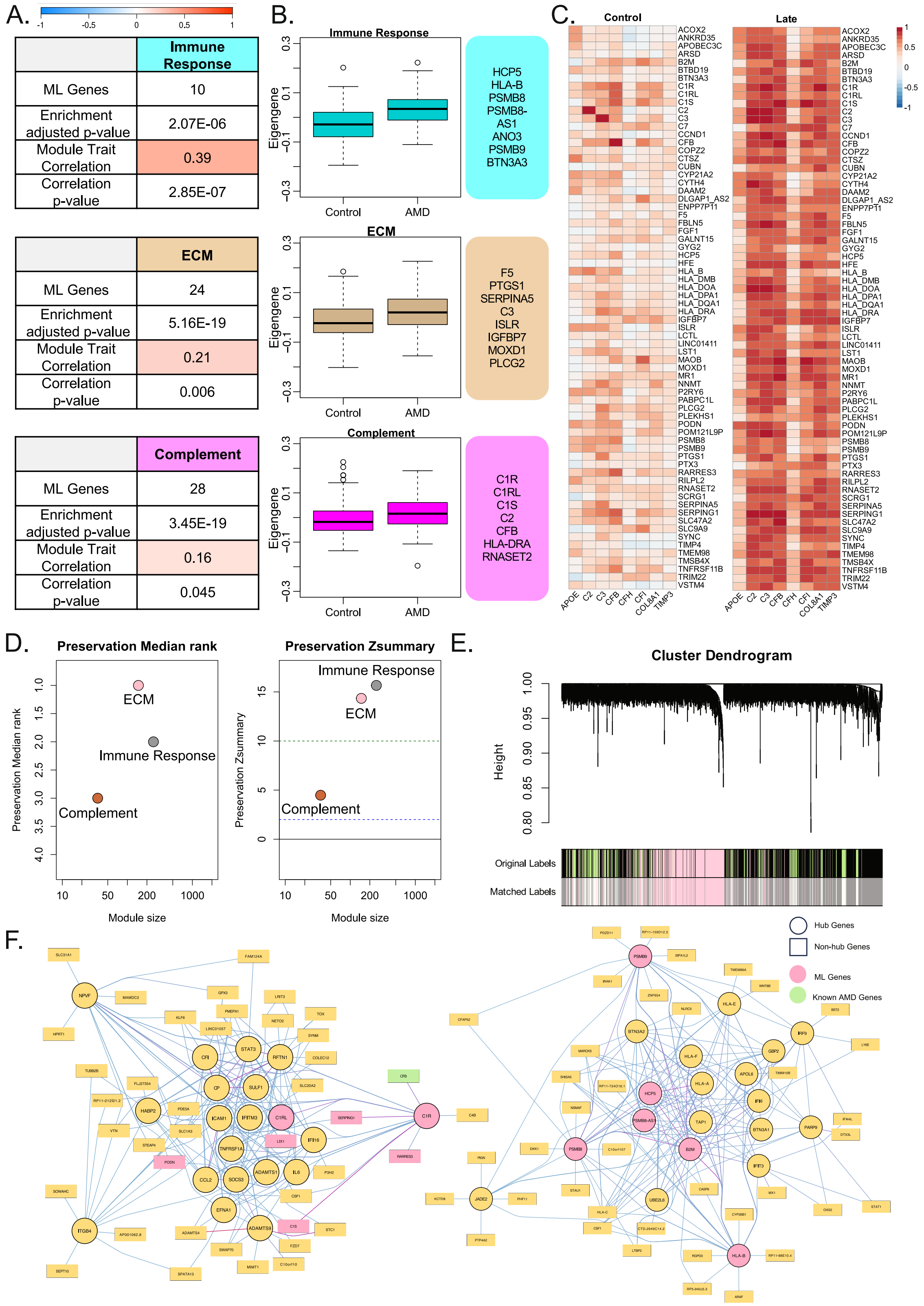
analysis was done using transcriptome data from 453 human retina and identified 44 modules and used GO analysis to identify the top biological pathways associated with these modules. We observed that majority of the (62/81) ML-genes were enriched within three modules associated with immune response (turquoise, *p-value* = 2.07×10^{-6}), extracellular matrix organization (ECM) (tan, *p-value* = 5.16×10^{-19}), and complement (magenta, *p-value* = 3.45×10^{-19}) pathways (Fig. 2A). These results are particularly interesting because of the putative role for these pathways in the pathogenesis of AMD[12,23]. Additionally, these modules also harbor three known AMD-GWAS genes, *C3* and *COL8A1* (tan) and *CFB* (magenta)[23]. *FBLN1*[24] and *MOXD1*[13] were particularly interesting for their implicated role in AMD. Additionally, we find several ML-genes involved in complement pathway such as *C7*, *C1S*, *C1R* and *C1RL* that have not been associated with AMD. Next, we assessed the module trait correlation across normal and AMD patients to identify the gene networks associated with the disease (Supplementary Fig. 7). Notably, all three ML-enriched modules exhibited a positive correlation with AMD and the eigengene for these modules demonstrated increased expression levels between normal and AMD (Fig. 2B).

Expression correlations are often used to infer functionality and regulatory relationships within specific biological contexts. As we did not observe many known AMD-GWAS genes in 81 ML-genes, we further explored the functional relationship between known AMD-GWAS genes with the ML-genes identified in this study between controls and AMD patients' transcriptome data. We observed a strong correlation (r2 > 0.7) in cases compared to controls with eight known AMD-GWAS genes correlated with 70/81 ML-genes (Fig. 2C, Supplementary Fig. 8) in late AMD whereas in controls only 2 known AMD-GWAS genes were correlated with 7/81 ML-genes (Supplementary Fig. 9). Additionally, these correlations exhibited statistical differences between cases and controls for most ML-genes (Supplementary Fig. 8). In comparison, in random set of 81 genes, a much smaller number of genes showed correlation, and they were comparable in cases (20/81) and controls (25/81) (Supplementary Fig. 10**)**. Furthermore, and even fewer of these correlations were statistically different between cases and controls (Supplementary Fig. 11). Taken together, these findings demonstrate that there is an enhanced positive correlation and thus by extension functional relationship between the expression ML-genes with known AMD-GWAS genes in AMD transcriptomes.

Next, we analyzed the preservation of the three modules enriched for ML-genes in normal and AMD networks using the density and connectivity-based preservation statistics available within the modulePreservation in WGCNA. The overall measure of preservation was defined as $Z_{summary}$ (Fig. 2D, E). The two modules that functionally annotate to immune response and ECM were well preserved between controls and AMD ($Z_{summary}$ >10). However, the module enriched for complement pathway genes was found to be weakly preserved ($Z_{summary\ =}$ 4.47)[25]. Next, we identified the top 20 most connected genes (hub genes) and their top 10 connections within the complement module from controls and AMD using WGCNA. We identified two ML-genes (*C1R* and *C1RL*) as hub genes in controls, whereas AMD network has six ML-genes (*PSMB8*, *PSMB9*, *PSMB8-AS1*, *B2M*, *HPC5*, *HLA-B*) as hub genes (Fig. 2F). These findings suggest these genes within complement pathways are modulators of immune activity in the retina that play an important role in pathogenesis in AMD.

AMD disease progression has shared and unique gene signatures

AMD is a progressive disease with early, intermediate, and late stages of the disease. Early/intermediate AMD is the most common and asymptomatic form, characterized by pigmentary abnormalities in RPE of the macular region and accumulation of extracellular aggregates of proteins, lipids and cellular components (called drusen). Vision loss happens in the late stage, which is usually subdivided into dry (geographic atrophy, or GA) and wet (choroidal neovascularization, or CNV) forms[26]. The symptoms of AMD worsen over time, although the rate at which the disease progresses varies and not all patients with early/intermediate AMD develop late disease. In the United States alone, over 1.75 million people have late stages of AMD, and 7.3 million people are affected with intermediate stages which are at the risk of developing late AMD[6]. However, there is a paucity of studies on the early and intermediate stages of AMD, and as a result, there are no reliable biomarkers for predicting the disease progression. Thus, we next applied the ML pipeline developed for late AMD in early (n = 175) and intermediate (n = 112) AMD to identify the molecular events that lead to AMD. We identified a set of 57 genes for early AMD that provided AUC-ROC statistic of 0.62 (CI 0.51–0.74) (Fig. 3A), whereas a set of 62 genes gave AUC-ROC statistic of 0.71 (CI 0.59–0.83) for intermediate AMD (Fig. 3B). The relatively modest performance of these models can be attributed to subtle alternations in gene expression during these initial stages, where vision loss or cell death in early and intermediate stages is not yet prominent. Thus, we next tested the performance of the features identified in early and intermediate AMD in late AMD. This analysis showed notable enhancement in predictive power with the 57 early AMD-associated genes, leading to an AUC-ROC statistic of 0.74 (CI 0.58–0.86) (Fig. 3C). Conversely, the performance remained comparable for the intermediate stage (AUC-ROC = 0.72, CI 0.58–0.89) (Fig. 3D). For both stages, the performance of the features selected based on shuffled label did not perform well (Supplementary Fig. 12). These findings are also consistent with a lower sensitivity as well as a higher proportion of early (29%) (Fig. 3E) and intermediate AMD (27%) (Fig. 3F) deviating from their ground truth prediction. Thus, it is likely that intermediate AMD might have distinct molecular underpinning that does not represent a transitional stage between early and late AMD. This was also reflected in the expression correlation of the candidate genes with known AMD-GWAS genes. 81 ML-genes identified in the late AMD showed higher correlation in early AMD compared to the intermediate AMD (Supplementary Fig. 13). Similarly, early AMD 57 gene signatures also showed higher correlation with late AMD and not intermediate AMD (Supplementary Fig. 14). However, the gene identified in intermediate does not show correlation with known AMD genes in any stages (Supplementary Fig. 15). Importantly, genes identified across both early and intermediate stages were enriched within modules associated with immune response and ECM pathways

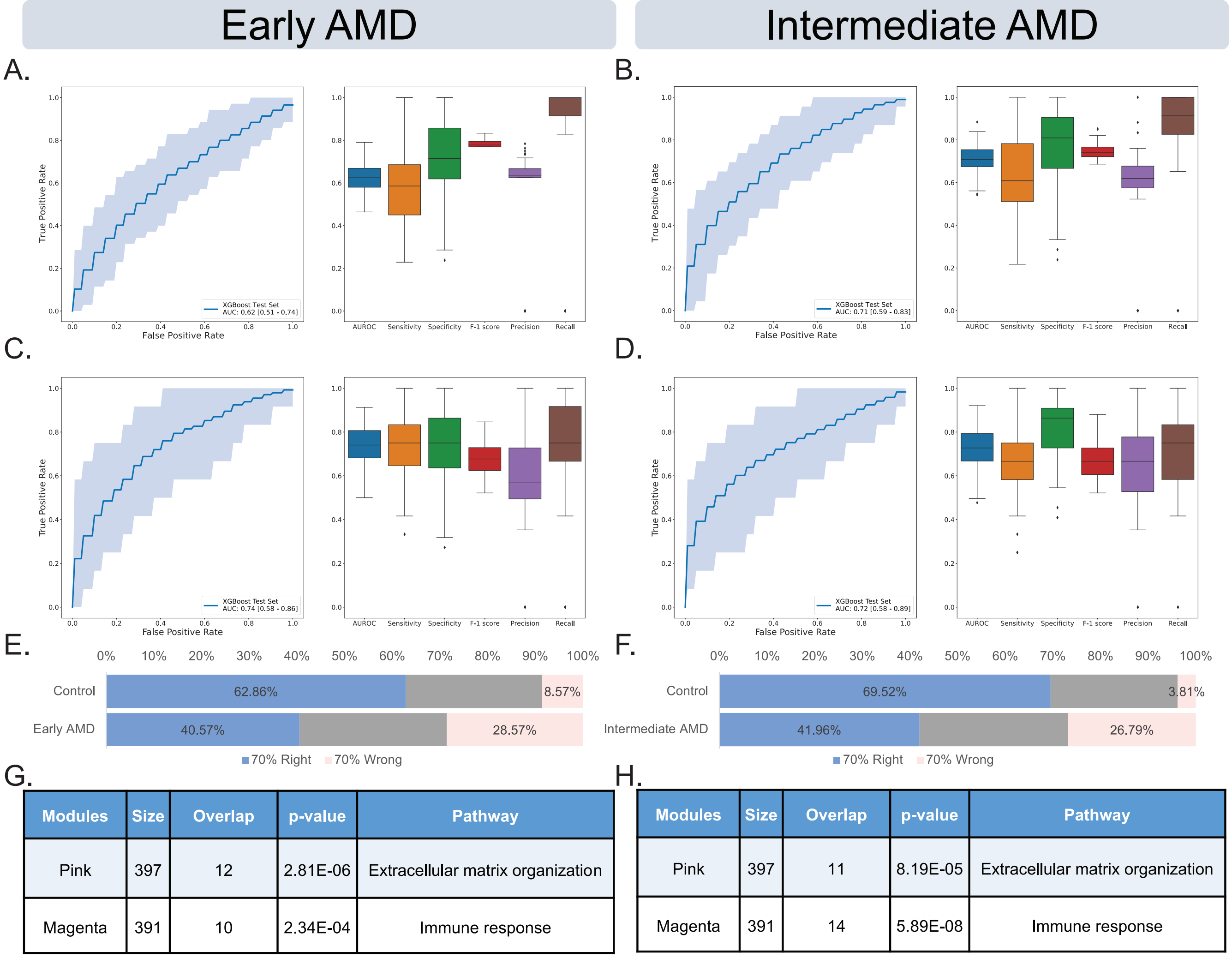


(Fig. 3G, H). While there exists a limited overlap in genes identified among the three disease stages of AMD, they manifest enrichment within identical modules associated with AMD-relevant pathways. This underscores that majority of AMD-progression response includes signatures reflecting immune response dysregulation, indicating a shared biological basis.

ML-genes are expressed within specific cell types in retina that are impacted in AMD

Our RNA-seq data was performed at the tissue level and yielded an average of gene transcript abundance that reflects the average signal from mixtures of cell-type-specific gene expression levels. This is particularly relevant for tissues characterized by a highly heterogeneous cell type composition, such

Fig. 2 | Gene co-expression network analysis to connect ML-genes to disease pathways. **A** Table representing the number of ML genes included in the modules, the p-value of the enrichment, the correlation value between the module's eigengene expression profile and the AMD status of the samples, and the correlation p-value. The color in each correlation cell corresponds to the correlation value on the scale provided at the top of the table. Notable, immune Response ranks highest in correlation to a patient being diagnosed with AMD. **B** Boxplot showing the eigengene value of each module between cases and controls. Next to the boxplots are the subset of 81 ML-genes that were identified as part of the top 10 hub genes within those modules. **C** Heat maps illustrating the correlation values between 81 ML-genes and known AMD-GWAS genes, categorized into control and AMD groups. ML-genes with a correlation value ≥ 0.7 with any known AMD genes in the AMD sample group are included. That list of genes was used to generate the heatmap for the Control group. **D** Module Preservation plot showing the preservation of gene composition in immune response and ECM modules from the control network in the late AMD network. The module enriched for complement pathways was weakly preserved in the AMD network. **E** Hierarchical cluster dendrogram of control and AMD co-expression networks. Each black branch (vertical line) corresponds to one gene. The color rows below the dendrogram indicate module membership showing the labels of genes in the Immune Response (Black), ECM (pink), and Complement (light-green) from the AMD network when matching with their labels in control network. **F** A Cytoscape visualization of the top 20 hub genes and their top 10 connections to other genes. Round nodes are the top 20 hub genes while rectangular nodes depict connected genes. ML-genes are highlighted in pink, and known AMD-GWAS genes are highlighted in green. Connections between genes are color-coded, with purple indicating stronger connections, while blue represents weaker connections.



G.

Modules	Size	Overlap	p-value	Pathway
Pink	397	12	2.81E-06	Extracellular matrix organization
Magenta	391	10	2.34E-04	Immune response

H.

Modules	Size	Overlap	p-value	Pathway
Pink	397	11	8.19E-05	Extracellular matrix organization
Magenta	391	14	5.89E-08	Immune response

Fig. 3 | Performance ML methods in classifying early and intermediate-stage AMD. **A** ROC and box plots illustrating the outcomes of classifying control and early AMD samples across 100 iterations using a set of 57 genes generated by the XGBoost model. **B** ROC and box plots display the outcomes of classifying control and Intermediate AMD samples across 100 iterations using a set of 62 genes generated by the XGBoost model. **C** ROC and box plots illustrate the notable enhancement in predicting late AMD outcomes by employing the XGBoost model across 100 iterations with 57 genes selected for Early AMD and control samples but applied to distinguish late AMD samples from the controls. **D** ROC and box plots demonstrate that utilizing the XGBoost model across 100 iterations with 62 genes associated with intermediate AMD did not improve performance in predicting late AMD outcomes. **E** A bar plot displays the distribution of early AMD and controls classified correctly 70% or more of the time (blue), or samples classified wrongly 70% or more of the time (pink). Grey sections represent samples that do not meet the criteria for either correct or wrong predictions. **F** Bar plots distribution for the intermediate AMD. Only ~40% of the intermediate AMD samples are predicted right compared to the ~70% of the controls. **G** Table demonstrating the enrichment of 57 early AMD genes within co-expression network modules associated with immune response and extracellular matrix (ECM) pathways, as determined by userListEnrichment within WGCNA. **H** A table depicting the enrichment of the same pathways for the 62 intermediate AMD genes.

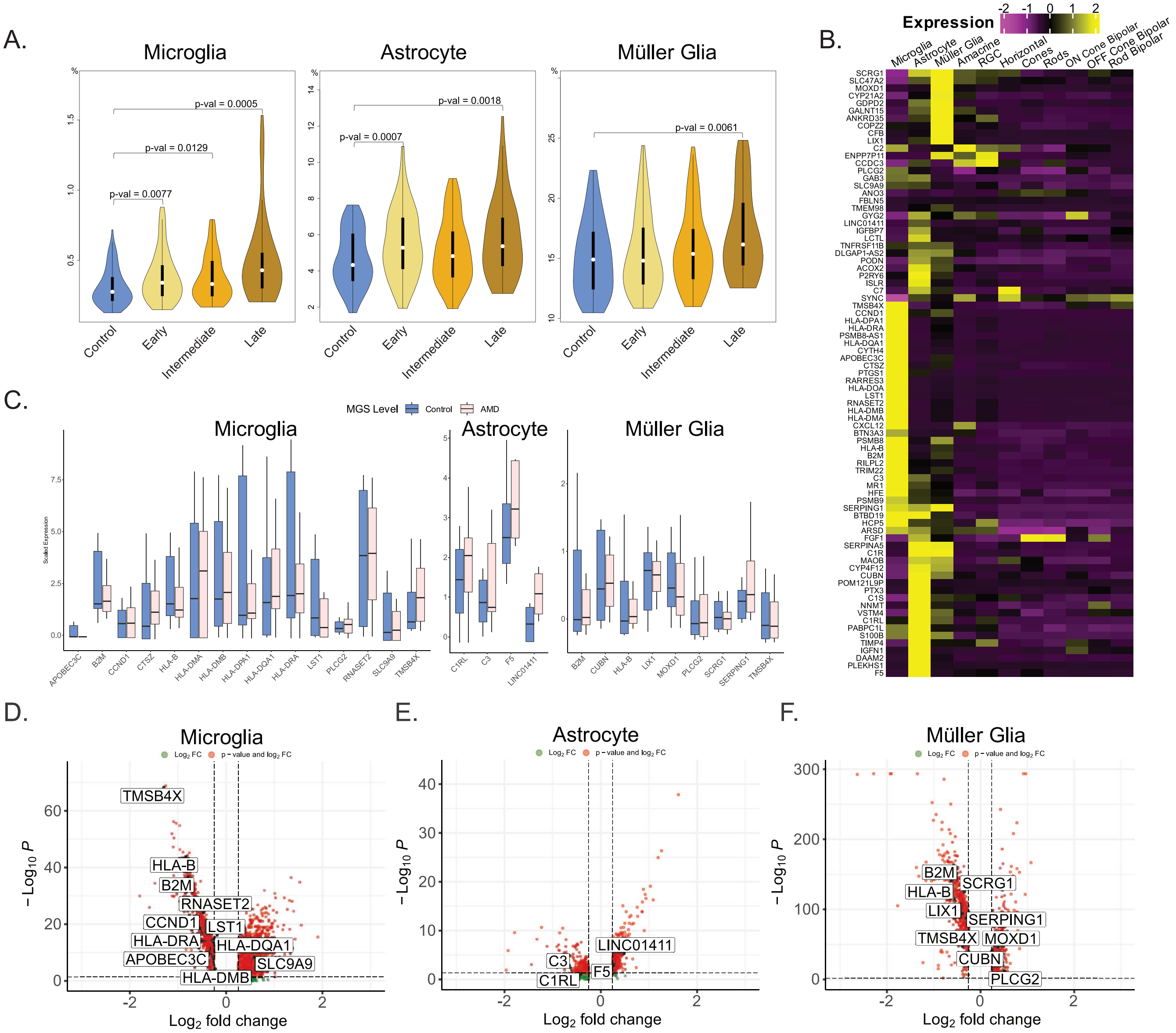


Fig. 4 | Expression of ML-genes across various retinal cell types and their alterations in AMD. A Violin plots illustrating the cell fraction of various cell types in deconvolution results from 453 bulk RNA samples, utilizing a single-cell RNA dataset as a reference. P-values are annotated to indicate significant differences in cell fraction ranges between Control and different AMD stages. P-values are omitted when they exceed 0.05, indicating a lack of statistical significance. **B** A heatmap illustrating the average gene expression of 81 ML-genes (as rows) across 11 retinal cell types (as columns). The color gradient indicates whether genes are predominantly expressed (positive value, yellow) or minimally expressed (negative value, purple) in a particular cell type. Most ML-genes are expressed in astrocytes, microglia and Müller glia. **C** Box plots showing the differences in ML-gene expression between 20 normal and 20 AMD single nuclei data across astrocytes, microglia, and Müller glia. The selected genes are those that have successfully passed Differential Expression analysis utilizing DESeq2 with a false discovery rate (FDR) threshold of 5%. **D** Volcano plot showing differentially expressed ML-genes within microglia, with eleven genes (one upregulated, ten downregulated) passing the DE threshold of Log2 fold change of 0.25 or higher and FDR 5%. **E** Volcano plot showing differentially expressed genes within astrocytes. Two upregulated and two down-regulated passing the DE threshold of Log2 fold change of 0.25 or higher and FDR 5%. **F** Volcano plot displaying significant ML-genes within Müller Glia (three upregulated and six downregulated) using the same threshold as mentioned above.

as the retina, which is made of six different cell types[27]. To understand the role of AMD-relevant cell type, we built a reference for the average expression of retinal cell types using cell-type specific markers[28] from six human retinas across three different studies[29–31] (Supplementary Table 1). Next, we implemented three distinct methods- CIBERSORTx[32], dTangle[33], and BayesPrism[34]—to deconvolute the cellular composition of both control and AMD samples. Subsequently, we applied student t-test to identify the cell types exhibiting significant changes associated with the disease, revealing astrocyte, microglia, Müller glia, and rods proportion to be significantly different between normal and late AMD (Supplementary Fig. 16). Microglia, astrocyte, and Müller glia proportion increase in the disease whereas the proportion of rods decreases (Fig. 4A). The decrease in the rods is observed only in the late stage, which could be the results of aging[35,36] as well as disease-related photoreceptor degeneration[37]. Notably, microglia were the only cell type that significantly changed in cell proportion across all stages of AMD, while alterations in astrocyte proportion were confined to early and late AMD stages. (Fig. 4A). The three tools used differ in their underlying algorithms, input requirements, and output formats. However, all of them point to the involvement of microglia in AMD, suggesting microglial activation and increased immune activity begin in early AMD much before the onset of photoreceptor loss in late AMD.

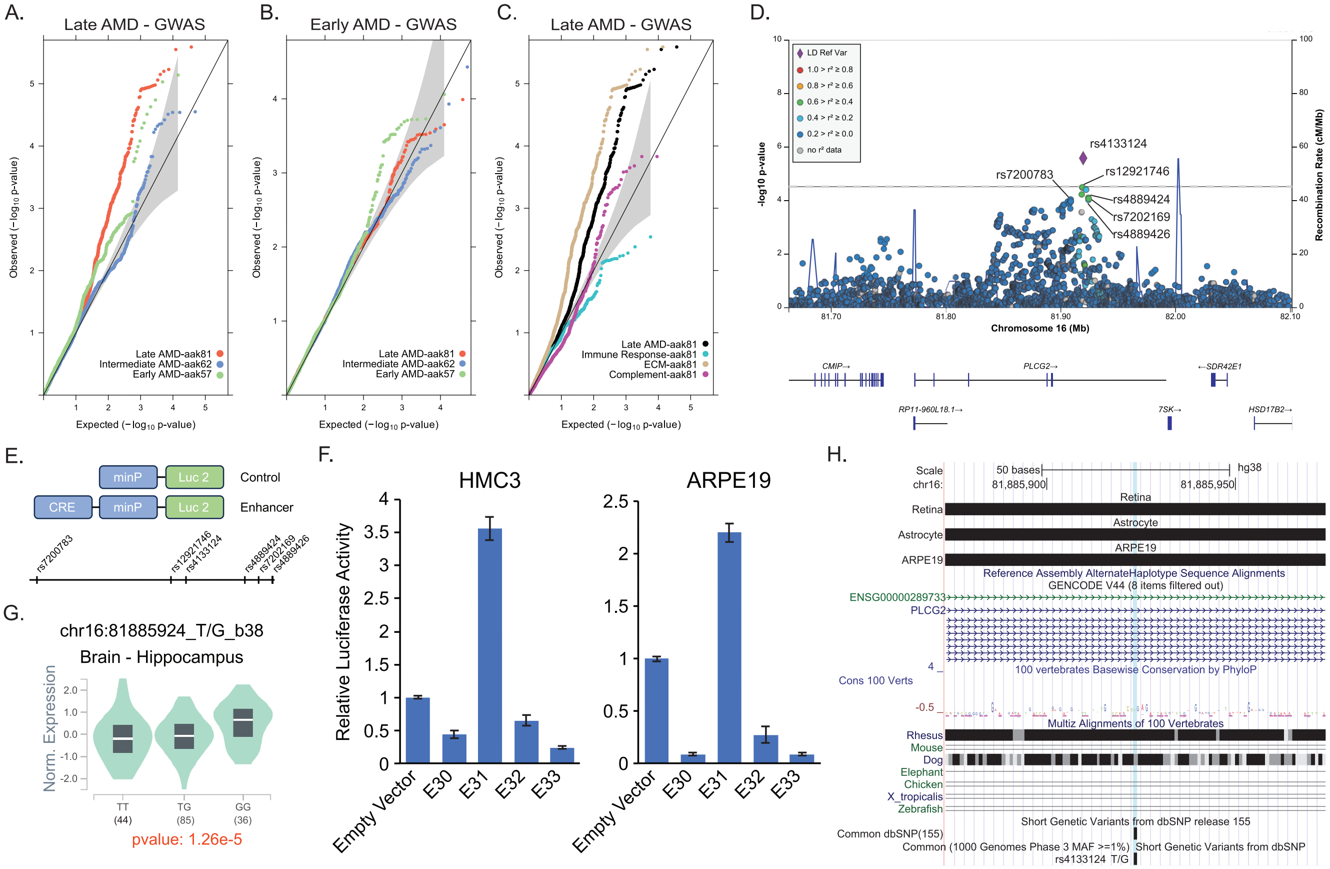


Fig. 5 | ML-genes are enriched for AMD-associated variants. **A** Quantile-quantile (Q-Q) plot using summary statistics data from late AMD demonstrates a greater deviation from the null distribution (solid black line) for ML-genes identified in early (green dots) and late AMD (red dots) compared to intermediate AMD (blue dots). **B** Q-Q plot using the early AMD data reveals similar but less pronounced trends across early, intermediate, and late AMD. **C** Q-Q plot for the 81 ML-genes, further segregated into 3 groups based on the modules identified in the WGCNA network, also demonstrates deviation for the complement (purple dots) and ECM pathways (tan dots), but not for the immune response (turquoise dots). **D** Regional association plot generated using LocusZoom plots displays the most strongly associated SNP, rs4133124 (purple diamond), along with other suggestively associated SNPs (p-value < 5×10^{-5}) within the intron of *PLCG2*. **E** A schematic representation of the luciferase assay, and the relative locations of the six SNPs around *PLCG2* that were tested in the assays. **F** The luciferase assay results for four constructs (E30-E33) indicate that construct E31, which contains rs4133124, exhibits a 3.5-fold increase in luciferase activity compared to the empty vector in HMC3 and a 2.2-fold increase in ARPE19. Error bars represent the standard error of the mean (SEM) calculated from three independent experiments. **G** eQTL violin plots sourced from GTEx to illustrate the correlation between the SNP rs4133124 and *PLCG2* gene expression specifically within the hippocampus region of the brain. **H** The UCSC Genome Browser graph displaying custom tracks from human retina, astrocytes, and ARPE19 cells shows the overlap of rs4133124 with open chromatin regions of AMD-relevant tissues and cell types. Additionally, the genomic region spanning rs4133124 shows conservation in primates as highlighted through multi-species alignment.

To validate the results of the deconvolution, we analyzed single-nuclei data from 20 controls and 20 late AMD patients from four published studies (Zenodo 7532115[38], GSE221042[39], GSE208434[40] and GSE203499) (Supplementary Table 2, Supplementary Fig. 17). Microglia cluster was further subdivided into microglia, perivascular macrophages, monocyte-derived macrophages and monocytes based on the top differentially expressed genes identified in human donor retina data from a recently published study[40], as well as cell-type specific markers genes[40–44] (Supplementary Fig. 18). A heatmap of 81 ML-genes across retinal cell types showed that majority of them were enriched in their expression in microglia, astrocytes, Müller glia (Fig. 4B). In addition, identified several ML-genes (21/81) that were differentially expressed across microglia, astrocyte and Müller glia (Fig. 4C). In microglia, 11 genes were differentially expressed (one upregulated, ten downregulated (Fig. 4D). In astrocytes, four gene were differentially expressed (two upregulated and two downregulated) (Fig. 4E). For Müller glia, three genes were upregulated and six were downregulated (Fig. 4F). In addition, several known AMD genes including *APOE* and *VEGFA* were also found to be differentially expressed within glial population of normal and AMD patients (Supplementary Table 3). These findings collectively provide support for the consistency and validity of the genes identified using the ML approach and cell types identified using the deconvolution method in an independent dataset, reinforcing the relevance of the identified gene expression alterations in the context of AMD.

AMD signature genes are enriched for AMD associated variants

Comparing transcript levels between healthy and diseased individuals cannot separate the cause vs consequences of the disease under scrutiny. Thus, we resorted to the published AMD-GWAS data on late AMD, comprising 16,144 patients and 17,832 controls[9] as well as early AMD data consisting of 14,034 cases and 91,214 controls[45] to access the potential association of genetic variants within ML-genes with AMD. The Quantile-Quantile (Q-Q) plot in late AMD-GWAS data (Fig. 5A) showed the largest deviation from the null p-value of the ML-genes identified in late AMD (red line) followed by early AMD (green line) suggesting that a subset of the ML-genes had genetic variants associated with AMD. In early AMD data, the gene identified in early AMD showed the largest deviation (green line) succeeded by late AMD (Fig. 5B). Interestingly, neither dataset exhibited apparent deviation for intermediate AMD (indicated by the blue line) (Fig. 5A, B). Furthermore, the ML-genes within the WGCNA modules enriched for complement and ECM organization individually also showed enrichment within late AMD-GWAS (Fig. 5C). By applying a suggestive association threshold (p-value < 5×10^{-5}), we identified two candidates, *PLCG2*;

rs4133124, p-value = 2.59×10^{-6} (Fig. 5D) and *IGFBP7;* rs1718877, p-value = 2.83×10^{-6} (Supplementary Fig. 19A) for late AMD and *USP7;* rs1471435, p-value = 7.27×10^{-6} (Supplementary Fig. 19B) and *NEIL1*; rs11634109, p-value = 4.27×10^{-5} (Supplementary Fig. 19C) for early AMD.

Next, we accessed the functional relevance of the suggestive associated SNPs, we selected four genomic regions spanning six SNPs around *PLCG2* and six genomics regions spanning seven SNPs around *IGFBP7* (Supplementary Table 4). We cloned these elements upstream of a minimal promoter-driven firefly luciferase gene in pGL4.23 (Fig. 5E) and tested for enhancer activity of the elements in the human microglia cell line, HMC3 and human RPE cell line, ARPE19. We identified that one element spanning rs4133124 within *PLCG2* showed 3.5- and 2.2-fold higher luciferase activity compared to the empty vector in HMC3 and ARPE19, respectively (Fig. 5F). We also tested the effect of reference T allele with the alternative G allele but found no change in enhancer activity in the rs4133124 region (data not shown). Additionally, this variant has been identified as an eQTL for *PLCG2* in hippocampus in the GTEx data (Fig. 5G)[4]. We did not find this eQTL in the retina (data not shown), which could be attributed to small proportion of glial cell in the bulk retina data[13]. The variant, rs4133124 reside in the intronic region, which is highly conserved in primates, but not in mice (Fig. 5H). It is noteworthy that macular degeneration is also caused by the degeneration of photoreceptors and underlying RPE in the central region called macula, which is a primate specific structure[46]. Additionally, this variant resides within the open chromatin region in retina, ARPE19 and astrocyte shown as custom track, suggesting a regulatory role (Fig. 5H). These results suggest that including the biological context of the genes can reveal additional genetic association within current GWAS datasets.

Discussion

For most complex diseases, including AMD, we have not exhausted the search for the disease genes as a significant proportion of heritability remains unexplained[9,23]. Identification of additional loci warrants large case-control cohorts, which can be cost-prohibitive and limited by sample availability. Most AMD-GWAS variants reside in the non-coding region and mediate their effects through gene expression regulation[13,15,47]. Consequently, gene expression profiling in normal and disease samples provides valuable resource for studying disease mechanisms and discovering additional causal genes. Gene expression data exhibits high heterogeneity, with significant natural variation present within and among human populations[1], a phenomenon exacerbated in diseases. Moreover, non-linear behavior is common in human systems due to their complex dynamics. Consequently, relying solely on a simple linear model, as often employed in the most common methods of differential gene expression analysis[48], harbors inherent limitations and pitfalls. Additionally, arbitrary cutoffs of fold-change and statistical thresholds does not necessarily reflect biological relevance[49]. In contrast, our approach can detect and learn from non-linear data patterns to identify a robust molecular classifier through a series of rigorous feature recognition and dimensionality reduction.

Integration of prior knowledge from AMD biology with molecular networks be leveraged to understand the functional relevance of novel genes. The interconnected nature of gene regulatory networks implies that the expression of all genes in disease-relevant cells has the potential to influence the functions of core disease-related genes[50]. Co-expression networks are particularly useful for this purpose because when constructed using disease-relevant expression profiles, they can capture the tissue and cell-type-specific nature of the disease[51,52]. It was notable that 76% (62/81) of ML-genes were involved in AMD-relevant immune response, complement and ECM pathways. Additionally, these genes had a much stronger pair-wise correlation with known AMD genes in cases compared to controls suggesting that during the disease process, they work closely with known AMD genes. Network preservation analysis further identified the modules involved in complement pathways to be not well preserved in the disease network affirming the well-established role of complement dysfunction in AMD[53]. The central regulatory hub genes of the complement modules were different in disease and controls. Among several ML-genes that were identified as hubs in the late AMD network, genes related to proteasome complex (*PSMB8*, *PSMB9*) are particularly interesting because of their role in immune system regulation[54]. It is therefore conceivable that dysregulation of proteasome activity-led immune response may contribute to the pathogenesis of AMD[55]. These results highlight the benefits of integrative analytical approaches to regain the holistic view of the AMD that is lost in experimentally tested reductionist approaches or hard statistical cut-offs.

Progression of early and intermediate to late AMD is observed frequently are attributed to multiple risk factors[56,57]. However, the role of known genetic risk factors doesn't seem to contribute significantly to the progression of intermediate to late AMD[58]. Additionally, in a sample of 6657 cases of intermediate AMD, 10 out of 34 late AMD loci did not show association, despite having adequate statistical power[9]. This is further strengthened by our findings of intermediate AMD that showed modest improvement in model performance in late AMD. Similarly, digital deconvolution analyses revealed the changes in the astrocyte in the early and late AMD stage but not in the intermediate stage. However, the changes in the microglia proportion were observed across all three stages. Finally, integration of the AMD-GWAS data from early and late AMD does not show deviation for the genes associated with intermediate AMD. Taken together, our results suggest that AMD progression may not be linear and involve both shared and stage-specific genes, pathways, and cellular perturbations.

The dysregulated immune response is a hallmark of normal aging[59] and a prominent feature in many neurodegenerative diseases[60] including AMD[61]. However, molecular, and cellular mechanisms underlying immune dysregulation-mediated neurodegeneration are multifaceted and have not been completely resolved in AMD. In the human retina, immune responses are orchestrated by three distinct glial cell types: Müller cells, astrocytes, and microglia[62]. Microglia, akin to macrophages, serve as the resident immune cells and clear cellular debris through complement pathay[63]. They then maintain immune surveillance in the retina, supporting neuroprotection and homeostasis[63] and can have different function based on anatomical location[42]. However, in disease conditions, they can get activated, migrate to the site of degeneration, and undergo morphological transformation leading to excessive release of inflammatory mediators and exacerbation of neurodegeneration[64]. Ocular sections from AMD samples shows the presence of activated microglia near the disease site that are thought to promote degeneration[65], and activated glial populations are enriched in AMD and related neurodegenerative diseases[39]. In contrast, a neuroprotective role of microglia has also been described in neurodegeneration[40,66]. Our data further emphasizes the role of microglia at a molecular and cellular level in a large cohort (105 controls and 348 AMD). We show that genes associated with AMD have abundant expression in microglia and astrocytes. Secondly, we detected distinct differences in the cellular composition of microglia between normal and diseased individuals based on digital deconvolution of transcriptome profiles. These results suggest that gene relevant to AMD pathology modulate the retinal glial function that are driving force in disease progression and photoreceptor degeneration in AMD.

Reaching significant association signals ($p < 5 \times 10^{-8}$) in traditional GWAS requires increasingly larger sample sizes to overcome statistical correction for multiple testing. Our approach of integrating the biological context, specifically genes exhibiting altered expression in disease-relevant cell type, with the GWAS data revealed novel genes within suggestive association signals ($p < 5 \times 10^{-5}$). *PLCG2* encodes for an enzyme that catalyzes the hydrolysis of phospholipids and releases critical signaling messengers involved in diverse cellular functions[67]. Genetic variants in *PLCG2* have been associated with autoinflammation, antibody deficiency, and immune dysregulation syndrome[68]. Recently, the identification of *PLCG2* rare variants in Alzheimer's patients has brought the focus on its role in neurodegenerative disease[69], that is likely to cause the disease through microglia-mediated innate immune response[70]. A pathway-based analysis implicated the role of *PLCG2* in AMD[71], however, our study presents the first convincing genetic and molecular evidence including the identification of rs4133124 as a single variant that is associated in both GWAS and eQTL analyses. The lack of observed differences in enhancer activity between the two alleles of rs4133124 could be due to the limitations of the luciferase construct used in the assay, potentially lacking crucial genomic elements necessary for detecting allele-specific effects. *IGFBP7* represents another such example which was identified as an AMD locus in the Japanese population[72], but was not replicated in Caucasian-dominant AMD-GWAS[23]. Taken together, our study shows that the integration of gene expression data from normal and disease individuals with existing GWAS data provides a powerful approach for gaining systems-level insights into AMD pathogenesis.

Methods

Cohort and data processing

This study was performed in accordance with the ethical standards of the Declaration of Helsinki and informed consent were taken from all the participants in original study[13]. The study was approved by the institutional review board of Baylor College of Medicine. We used RNA-seq data from 453 post-mortem donor retina that were evaluated to determine the level of AMD based on the Minnesota Grading System (MGS)[73], with criteria similar to the Age-related Eye Disease Study (AREDS)[26]. MGS1 donor retina had no AMD features and served as controls, whereas MGS2-4 represented early, intermediate and late stages of AMD, respectively. A total of 105 controls, 175 early AMD, 112 intermediate AMD, and 61 late AMD were included in the dataset[13], which were primarily of Caucasian origin. Transcriptome analysis of donor retina was performed using RNA-Seq after enriching for poly-adenylated RNA. Raw RNA-Seq reads were processed as described earlier[13]. Briefly, trimmed reads were aligned to the Ensembl release 85 (GRCh38.p7) human genome using STAR version 2.5.2a[74] RSEM[75] was used to obtain estimated gene expression levels. Gene expression matrix was normalized using Trimmed Mean of M-values (TMM) in Counts per Million (CPM) using edgeR[76] and genes were filtered by setting a threshold of 1 CPM in 10% of all samples. After initial quality control, 105 normal, 175 early, 112 intermediate and 61 late AMD samples were used in subsequent analyses.

Feature selection

Normalized RNA-seq data was used to select the best features to be used in machine learning models. We applied three different feature selection methods ANOVA, AUC and Kruskal using mlr3filters (version 0.7.1) in R. All three methods are filter-based on scoring all the available features and then selecting features with the highest scores. ANOVA calculates the F-score to test the significance of each gene based on the analysis of variance. AUC computes a score called Area Under the Curve, also known as classification accuracy. The Kruskal method estimates the score for each gene utilizing the Kruskal-Wallis rank sum test, which is non-parametric compared to ANOVA.

We applied the feature selection 1000 times. For each iteration, we randomly sampled 80% of the data, used the three methods to obtain a score for each gene, and then filtered for the top five hundred genes. At the end of one thousand iterations, we obtained the list of genes and calculated how many times they were in the top five hundred in each iteration for each method. Finally, we took the top one hundred genes from each method and proceeded with the genes that appeared in all three methods.

Machine learning models

We employed four machine learning models: neural network, logistic regression, eXtreme Gradient Boosting (XGB), and random forest. All machine learning models are available in Python as separate libraries. To train the model, we first randomly split the data into an 80% train and validation set and a 20% test set. Then we further split the train-validate set into eighty percent training set and twenty percent validation set. In other words, we randomly divided into 64% training set, 16% validation set, and 20% testing set. Subsequently, we fed the data into the models to train and generate predictions for each sample, validating the results. Once the models were trained, we evaluated their performance on the test set using metrics such as sensitivity, specificity, recall, precision, F-1 score, and AUROC. To handle the class imbalance between cases and controls, we applied the Youden J[20] method to adjust the threshold for each iteration. The *p*-value cutoff was set to 0.05, and the best threshold was calculated using Youden's J statistic[20,77]. The maximum distance to the diagonal line was considered as the optimal cutoff point value. We then used the binary predictions to determine the frequency of correct predictions for each sample. Subsequently, we categorized the samples into two groups: those predicted correctly at least 70% of the time and those predicted incorrectly at least 70% of the time.

The following gene lists were used to compare the performance of chosen feature selection methods: **(1)** Genes within 500KB of the 34 AMD-GWAS loci[9] **(2)** High confidence AMD genes, including genes from the 34 loci for which the connection with AMD has been establish either through rare variant discovery, QTL analyses or functional validation **(3)** Genes deemed relevant to macular degeneration pathogenesis in the literature that emerged from extensive PubMed searched described before[13] and **(4)** Features obtained by randomly swapping the labels of our dataset and rerun feature selection to obtain another set of genes.

We then tested both the original set of 81 genes and the later set of 48 genes obtained using shuffled labels by running the model using both sets. In one iteration, we ran the model with each set of genes using the true label. In another iteration, we shuffled all the labels in the training, validating, and testing set, and then ran the model. Finally, we ran the model with the training and validation set having shuffled labels, while keeping the labels of the testing set true.

SHAP (SHapley Additive exPlanations) was used for interpreting the output of the models by attributing the contribution of each feature to the final prediction. We used the model parameters built using the training data and the original gene expression as inputs for the SHAP library in Python to compute SHAP values for each instance[78]. The analysis shows a view of how each gene, or variable, will affect the model and alter the prediction.

Weighted gene-correlation network analysis

Weighted co-expression networks were constructed using the WGCNA[22] using the Bioconductor R package. Briefly, a similarity matrix between each gene was obtained and the adjacency was calculated using Spearman correlation. We then used hypergeometric testing at a significance threshold of 0.05 alpha-level after Bonferroni correction accessing enrichment of genes for enrichment across identified modules. Pathway analysis was performed on each module using Gene Ontology biological process terms.

For module preservation, we first built the co-expression networks separately on controls and AMD samples. Next, we used the preservation statistics available within the modulePreservation function in the WGCNA[22] in R. We then employed a permutation test (number of permutations = 500), which randomly permutes the module assignment in the control and AMD networks to assess if the observed value of preservation statistic is higher than what is expected by chance and assigns a permutation test *p*-value. The observed preservation values were then standardized with regard to the mean and variance and a significance Z score was defined for each preservation statistic. In order to compare the degree of preservation of modules between the normal and AMD networks, differential module preservation"—$\Delta Z_{summary}$, which is the arithmetic difference between the two preservation scores was calculated.

Hub genes were identified using the signedKME function, which calculated the KME values between a gene and all the modules in the network. With the corOptions parameter: "use = 'p', method = 'spearman'". This calculates the correlation between the expression patterns of each gene and the module eigengene. Genes with the largest kME are considered 'hub' genes within the modules. We selected the genes from their respective modules and sorted in descending order the KME value for that module. Genes with the largest kME we assigned 'hub' genes within the modules.

Polygenic risk score

We obtained the beta coefficient of common, independently associated AMD-risk variants at 34 loci from published AMD-GWAS data[9]. 42 common, independently associated risk variants out of 52 were found in our data as the remaining were rare variants. To compute the polygenic risk score for each individual, we multiplied the genotype (coded as 0, 1, or 2) by its corresponding beta coefficient and sum up the weighted beta coefficient values across all variants. To test the difference between the PRS of Control and AMD, we used the Mann-Whitney U Test through the wilcox.test() function in R.

Heat map

We created the gene expression heatmaps using pheatmap() function in R. First, we extracted the genes of interest's expressions from the normalized bulk data and performed log 2-based transformation. The function pheatmap takes in a matrix. We defined the columns as the samples and the rows as the genes. Samples from control and AMD groups are separated to investigate the difference in gene expression between the two groups.

Correlation with known AMD genes

We utilized the CPM normalized counts matrix to generate two distinct matrices: one comprising the 81 genes and the other containing the known AMD genes, separately for cases and controls. Subsequently, we employed the "cor" function in R to establish the correlation matrix between these two matrices. A threshold of 0.7 was applied, indicating that correlation coefficients greater than 0.7 or smaller than −0.7 would be considered significant. The "pheatmap" function in R was utilized to generate a correlation heat map between 81 ML-genes and known AMD genes between the control and AMD group. We used the "cocor" package in R to test the difference between the correlation of each gene pair in Control vs AMD. The "cocor" analysis was performed using a formula in the form "Gene1 + Gene2 | Gene1 + Gene2", where two independent datasets, the controls group, and the cases group were specified. cocor automatically selected Fisher's test to determine the significance of differences between correlation coefficients. The resulting p-values were used to determine whether the correlation coefficients between the two groups were significantly different.

Integration of GWAS data and Q-Q plot

We first removed of variants in the major histocompatibility complex region, and within +/− 1 Mb of the known GWAS signals[9,45]. The data used for generating the Q-Q plot consists of a matrix derived from SNPs located within the gene bodies of the genes identified through feature selection (57 for early AMD, 62 for intermediate AMD and 81 genes for late AMD). Each entry in the matrix represents the negative base-10 logarithm of the quantiles for the corresponding SNP's p-value. G-G Plot was used to plot the quantile on the x-axis and the minus log 10 p-value on the y-axis.

Single cell RNA-seq analysis

FASTQ files were downloaded from the GEO databases from three published studies: 2 samples from GSE202747[30], 1 sample from GSE130636[31] and 3 samples from the UK BioStudies database[29]. Subsequently, sequencing reads were mapped to the available hg38 genome using CellRanger (version 6.1.2). The gene expression matrices generated by Cell Ranger were filtered to remove cells with unique molecular identifier (UMIs) less than 200 or more than 6000 or with more than twenty percent mitochondrial reads. Data was normalized and transformed using SCTransform V2 from Seurat. To annotate the data with a UMAP visualization, we initially conducted dimensionality reduction using RunPCA (seed.use = 1) and corrected batch effects with Harmony (group.by.vars = "samples"), followed by applying RunUMAP to create the visual matrix, identifying nearest neighbors with FindNeighbors (reduction = "umap", dims = 1:2) and group into clusters with FindClusters (resolution = 0.5, random.seed = 1, algorithm = 1). Cells were annotated using the gene expression heatmap of curated cell-type-specific marker genes[27,79,80] (Supplementary Table 5). The gene expression heatmap was made with DoHeatmap function from Seurat.

Deconvolution

The bulk RNA-seq data containing 453 samples consisting of 105 control, 175 early AMD, 112 intermediate AMD, and 61 late AMD were used as mixture dataset[13]. We used Seurat-generated reference to implement three deconvolution methods: CIBERSORTx[32], dTangle[33], and BayesPrism[34]. For CIBERSORTx, the normalized count matrix of 3,956 differentially expressed genes across cell types was uploaded to the CIBERSORTx web application to generate the custom signature matrix of 2002. Normalized RNA-seq CPM counts data of the bulk data was used to perform the deconvolution. For dTangle on R, we used the CPM normalized counts RNA-seq counts data that was log2 normalized. The average expression of genes in the Seurat object was log2 normalized and used for dTangle. 14,705 genes that were present in both the bulk and single cell data were used in the deconvolution. BayesPrism analysis employed raw counts from our bulk RNA-seq dataset and SCT-normalized cell type-specific data from single-cell analysis. Genes on sex chromosomes and ribosomal genes were excluded, resulting in 13,716 genes used in the analysis. All three methods output cell fractions for each sample. Afterward, we performed a t-test on the cell fractions between the AMD samples and the control samples, specifically, samples from each stage of AMD against the samples from the control. From the cell fraction output of the deconvolution, we plotted the violin plots using "vioplot" package in R to visualize the difference in cell fractions across disease levels. We removed outliers in each disease stage level by setting a quantile range of 0.05 to 0.95.

Single nuclei RNA-seq analysis

Single nuclei data containing 20 control and 20 AMD samples from four datasets was used for this analysis (Zenodo 7532115[38], GSE221042[39], GSE208434[40], GSE203499). Counts data provided by the study or as the output of Cell Ranger (v6.1.2) was loaded into a Seurat object for each sample. They were then merged based on which dataset they belonged and performed quality control to remove bad quality cells and doublets based on counts not less than 200 and more than 7500, not more than ten percent mitochondrial reads per cell. We implemented the fast integration pipeline from Seurat 4.4 by performing normalization with SCTransform V2, used RunPCA (seed.use = 1) on each dataset. We feed the list of Seurat objects for all datasets to SelectIntegrationFeatures (nFeatures = 3000), PrepSCTIntegration, and FindIntegrationAnchors (normalization.method = "SCT", dims = 1:30, reduction = "rpca", k.anchors = 20)[81]. With the list of anchors, we integrate using IntegrateData (normalization.method = "SCT", dims = 1:30). After the integration, cell embeddings for projections were calculated using RunPCA (seed.use = 1) and RunUMAP(reduction = "pca", dims = 1:30, min.dist = 0.2, seed.use = 1). Unsupervised clustering and cluster identification were done using FindNeighbors (reduction = "umap", dims = 1:30) and FindClusters (algorithm = 1, resolution = 0.25, random.seed = 1). To annotate the cell-types, we used addModuleScore function from Seurat to generate the scores for each cell type for each cell with the list of cell-type-specific marker genes curated from literature[27,79,80] (Supplementary Table 5).

To further refine the specificity of glial cell annotations, we isolated clusters with the highest microglia module scores and re-clustered the data to annotate microglia, monocyte-derived macrophages, perivascular macrophages, and monocytes. This process followed the same steps for generating cell embeddings, and unsupervised clustering as described above. Additionally, we implemented a two-step approach to annotate the cell types. First, we calculated the gene module scores for homeostatic microglia, perivascular macrophages, and monocyte-derived macrophages based on the top differentially expressed genes (DEGs) ranked by fold change, as identified in human donor retina data from a recently published single cell atlas of human retinal mononuclear phagocytes in AMD and age match controls[40]. Secondly, we utilized curated marker genes[40–44] to generate the module scores (Supplementary Table 5). Data visualization was performed using dot plots, and cell types were assigned based on the consensus of module scores derived from both methods.

Differential expression analysis between normal and AMD samples was conducted using the FindMarkers function (logfc_threshold = 0.25, min.pct = 0.1, test.use = 'wilcox', assay = 'RNA') in Seurat, comparing control and AMD groups for microglia, astrocyte, and Müller glia cell types. The resulting differentially expressed genes (DEGs) were visualized using the EnhancedVolcano package. An expression heatmap of 81 genes was generated based on the control group across all 11 retinal cell types using the DoHeatmap function from Seurat.

Luciferase assay

We cloned 10 genomic regions spanning 13 SNPs (6 in *PLCG2*, and 7 in *IGFBP7*) upstream of a minimal promoter-driven firefly luciferase gene in pGL4.23 (Promega). HMC3, and ARPE19 cells were plated in 24-well plates (28 K cells/well) and were transiently transfected after 24 h with test luciferase constructs (500 ng) and *Renilla* luciferase vector (10 ng for transfection normalization) in duplicates in three independent experiments using 2 μL of FuGENE HD transfection reagent (Roche Diagnostic) in 100 μL of Opti-MEM medium (Invitrogen). Cells were grown for 48 h and luminescence was measured using a dual luciferase reporter assay system on a Texan Spark Multimode Microplate Reader per the manufacturer's instructions.

Data availability

The transcriptome data from 453 human donor retina used in this study are available in GEO (accession code GSE115828). Summary Statistics of advanced AMD is available at http://amdgenetics.org/ and early AMD is available for the download from www.genepi-regensburg.de/earlyamd. Single-cell retina data are available from GSE202747 and GSE130636 and single nuclei data was available under the accession code GSE196235, GSE203499, GSE208434, GSE221042 and from Zenodo data repository number 7532115 (https://zenodo.org/records/7532115).

Code availability

No customized code was used in this manuscript.

Received: 11 June 2024; Accepted: 22 May 2025;
Published online: 14 June 2025

References

1. Storey, J. D. et al. Gene-expression variation within and among human populations. *Am. J. Hum. Genet* **80**, 502–509 (2007).
2. Montgomery, S. B., Bernstein, J. A. & Wheeler, M. T. Toward transcriptomics as a primary tool for rare disease investigation. *Cold Spring Harb. Mol. Case Stud.* **8**, a006198 (2022).
3. Barbeira, A. N. et al. Exploiting the GTEx resources to decipher the mechanisms at GWAS loci. *Genome Biol.* **22**, 49 (2021).
4. Consortium, G. T. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
5. Wong, W. L. et al. Global prevalence of age-related macular degeneration and disease burden projection for 2020 and 2040: a systematic review and meta-analysis. *Lancet Glob. Health* **2**, e106–e116 (2014).
6. Friedman, D. S. et al. Prevalence of age-related macular degeneration in the United States. *Arch. Ophthalmol.* **122**, 564–572 (2004).
7. Young, R. W. Pathophysiology of age-related macular degeneration. *Surv. Ophthalmol.* **31**, 291–306 (1987).
8. Ratnapriya, R. & Chew, E. Y. Age-related macular degeneration-clinical review and genetics update. *Clin. Genet* **84**, 160–166 (2013).
9. 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).
10. Singh, N., Swaroop, A. & Ratnapriya, R. Making Biological Sense of Genetic Studies of Age-Related Macular Degeneration. *Adv. Exp. Med Biol.* **1256**, 201–219 (2021).
11. Priya, R. R., Chew, E. Y. & Swaroop, A. Genetic studies of age-related macular degeneration: lessons, challenges, and opportunities for disease management. *Ophthalmology* **119**, 2526–2536 (2012).
12. Parmeggiani, F. et al. Mechanism of inflammation in age-related macular degeneration. *Mediators Inflamm.* **2012**, 546786 (2012).
13. Ratnapriya, R. et al. Retinal transcriptome and eQTL analyses identify genes associated with age-related macular degeneration. *Nat. Genet* **51**, 606–610 (2019).
14. Strunz, T. et al. A mega-analysis of expression quantitative trait loci in retinal tissue. *PLoS Genet* **16**, e1008934 (2020).
15. Orozco, L. D. et al. Integration of eQTL and a Single-Cell Atlas in the Human Eye Identifies Causal Genes for Age-Related Macular Degeneration. *Cell Rep.* **30**, 1246–1259 e1246 (2020).
16. Angermueller, C., Parnamaa, T., Parts, L. & Stegle, O. Deep learning for computational biology. *Mol. Syst. Biol.* **12**, 878 (2016).
17. Libbrecht, M. W. & Noble, W. S. Machine learning applications in genetics and genomics. *Nat. Rev. Genet* **16**, 321–332 (2015).
18. Beer, M. A., Shigaki, D. & Huangfu, D. Enhancer Predictions and Genome-Wide Regulatory Circuits. *Annu Rev. Genomics Hum. Genet* **21**, 37–54 (2020).
19. Ting, D. S. W. et al. Artificial intelligence and deep learning in ophthalmology. *Br. J. Ophthalmol.* **103**, 167–175 (2019).
20. Youden, W. J. Index for rating diagnostic tests. *Cancer* **3**, 32–35 (1950).
21. Scott, M. Lundberg S-IL. A unified approach to interpreting model predictions. *Adv. Neural Inf.* **30**, 4765–4774 (2017).
22. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinforma.* **9**, 559 (2008).
23. Fritsche, L. G. et al. Age-Related Macular Degeneration: Genetics and Biology Coming Together. *Annu Rev Genomics Hum Genet.* **15**, 151–171 (2014).
24. Lotery, A. J. et al. Reduced secretion of fibulin 5 in age-related macular degeneration and cutis laxa. *Hum. Mutat.* **27**, 568–574 (2006).
25. Langfelder, P., Luo, R., Oldham, M. C. & Horvath, S. Is my network module preserved and reproducible?. *PLoS Comput Biol.* **7**, e1001057 (2011).
26. Ferris, F. L. et al. A simplified severity scale for age-related macular degeneration: AREDS Report No. 18. *Arch. Ophthalmol.* **123**, 1570–1574 (2005).
27. Yan, W. et al. Cell Atlas of The Human Fovea and Peripheral Retina. *Sci. Rep.* **10**, 9802 (2020).
28. Hoshino, A. et al. Molecular Anatomy of the Developing Human Retina. *Dev. Cell* **43**, 763–779.e4 (2017).
29. Lukowski, S. W. et al. A single-cell transcriptome atlas of the adult human retina. *EMBO J.* **38**, e100811 (2019).
30. Mullin, N. K. et al. Multimodal single-cell analysis of nonrandom heteroplasmy distribution in humanretinal mitochondrial disease. *JCI Insight* **8**, e165937 (2023).
31. Voigt, A. P. et al. Molecular characterization of foveal versus peripheral human retina by single-cell RNA sequencing. *Exp. Eye Res* **184**, 234–242 (2019).
32. Newman, A. M. et al. Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat. Biotechnol.* **37**, 773–782 (2019).
33. Hunt, G. J., Freytag, S., Bahlo, M. & Gagnon-Bartsch, J. A. dtangle: accurate and robust cell type deconvolution. *Bioinformatics* **35**, 2093–2099 (2019).
34. Chu, T., Wang, Z., Pe'er, D. & Danko, C. G. Cell type and gene expression deconvolution with BayesPrism enables Bayesian integrative analysis across bulk and single-cell RNA sequencing in oncology. *Nat. Cancer* **3**, 505–517 (2022).

35. Panda-Jonas, S., Jonas, J. B. & Jakobczyk-Zmija, M. Retinal photoreceptor density decreases with age. *Ophthalmology* **102**, 1853–1859 (1995).
36. Curcio, C. A., Millican, C. L., Allen, K. A. & Kalina, R. E. Aging of the human photoreceptor mosaic: evidence for selective vulnerability of rods in central retina. *Invest Ophthalmol. Vis. Sci.* **34**, 3278–3296 (1993).
37. Curcio, C. A., Medeiros, N. E. & Millican, C. L. Photoreceptor loss in age-related macular degeneration. *Invest Ophthalmol. Vis. Sci.* **37**, 1236–1249 (1996).
38. Orozco, L. D. et al. A systems biology approach uncovers novel disease mechanisms in age-related macular degeneration. *Cell Genom.* **3**, 100302 (2023).
39. Kuchroo, M. et al. Single-cell analysis reveals inflammatory interactions driving macular degeneration. *Nat. Commun.* **14**, 2589 (2023).
40. Yu, C. Microglia at sites of atrophy restrict the progression of retinal degeneration via galectin-3 and Trem2. *J. Exp. Med.* **221**, e20231011 (2024).
41. Butovsky, O. & Weiner, H. L. Microglial signatures and their role in health and disease. *Nat. Rev. Neurosci.* **19**, 622–635 (2018).
42. O'Koren, E. G. et al. Microglial Function Is Distinct in Different Anatomical Locations during Retinal Homeostasis and Degeneration. *Immunity* **50**, 723–737 e727 (2019).
43. Bloomfield, C. L. et al. Retinal microglia express more MHC class I and promote greater T-cell-driven inflammation than brain microglia. *Front Immunol.* **15**, 1399989 (2024).
44. Hao, Y. et al. Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587 e3529 (2021).
45. Winkler, T. W. et al. Genome-wide association meta-analysis for early age-related macular degeneration highlights novel loci and insights for advanced disease. *BMC Med Genomics* **13**, 120 (2020).
46. Masland, R. H. The fundamental plan of the retina. *Nat. Neurosci.* **4**, 877–886 (2001).
47. Advani, J. et al. QTL mapping of human retina DNA methylation identifies 87 gene-epigenome interactions in age-related macular degeneration. *Nat. Commun.* **15**, 1972 (2024).
48. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010).
49. Greenland, S. et al. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *Eur. J. Epidemiol.* **31**, 337–350 (2016).
50. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* **169**, 1177–1186 (2017).
51. Voineagu, I. et al. Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature* **474**, 380–384 (2011).
52. Zhang, B. & Horvath, S. et al. A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* **4**, Article17 (2005).
53. Geerlings, M. J., de Jong, E. K. & den Hollander, A. I. The complement system in age-related macular degeneration: A review of rare genetic variants and implications for personalized treatment. *Mol. Immunol.* **84**, 65–76 (2017).
54. Kammerl, I. E. & Meiners, S. Proteasome function shapes innate and adaptive immune responses. *Am. J. Physiol. Lung Cell Mol. Physiol.* **311**, L328–L336 (2016).
55. Ethen, C. M. et al. Transformation of the proteasome with age-related macular degeneration. *FEBS Lett.* **581**, 885–890 (2007).
56. Chakravarthy, U. et al. Progression from Early/Intermediate to Advanced Forms of Age-Related Macular Degeneration in a Large UK Cohort: Rates and Risk Factors. *Ophthalmol. Retin.* **4**, 662–672 (2020).
57. Yan, Q. et al. Genome-wide analysis of disease progression in age-related macular degeneration. *Hum. Mol. Genet* **27**, 929–940 (2018).
58. Sardell, R. J. et al. Progression Rate From Intermediate to Advanced Age-Related Macular Degeneration Is Correlated With the Number of Risk Alleles at the CFH Locus. *Invest Ophthalmol. Vis. Sci.* **57**, 6107–6115 (2016).
59. Shaw, A. C., Goldstein, D. R. & Montgomery, R. R. Age-dependent dysregulation of innate immunity. *Nat. Rev. Immunol.* **13**, 875–887 (2013).
60. Hammond, T. R., Marsh, S. E. & Stevens, B. Immune Signaling in Neurodegeneration. *Immunity* **50**, 955–974 (2019).
61. Ambati, J., Atkinson, J. P. & Gelfand, B. D. Immunology of age-related macular degeneration. *Nat. Rev. Immunol.* **13**, 438–451 (2013).
62. Murakami, Y., Ishikawa, K., Nakao, S. & Sonoda, K. H. Innate immune response in retinal homeostasis and inflammatory disorders. *Prog. Retin Eye Res* **74**, 100778 (2020).
63. Li, F., Jiang, D. & Samuel, M. A. Microglia in the developing retina. *Neural Dev.* **14**, 12 (2019).
64. Rashid, K., Akhtar-Schaefer, I. & Langmann, T. Microglia in Retinal Degeneration. *Front Immunol.* **10**, 1975 (2019).
65. Gupta, N., Brown, K. E. & Milam, A. H. Activated microglia in human retinitis pigmentosa, late-onset retinal degeneration, and age-related macular degeneration. *Exp. Eye Res* **76**, 463–471 (2003).
66. Peng, B. et al. Suppression of microglial activation is neuroprotective in a mouse model of human retinitis pigmentosa. *J. Neurosci.* **34**, 8139–8150 (2014).
67. Jackson, J. T., Mulazzani, E., Nutt, S. L. & Masters, S. L. The role of PLCgamma2 in immunological disorders, cancer, and neurodegeneration. *J. Biol. Chem.* **297**, 100905 (2021).
68. Zhou, Q. et al. A hypermorphic missense mutation in PLCG2, encoding phospholipase Cgamma2, causes a dominantly inherited autoinflammatory disease with immunodeficiency. *Am. J. Hum. Genet* **91**, 713–720 (2012).
69. Sims, R. et al. Rare coding variants in PLCG2, ABI3, and TREM2 implicate microglial-mediated innate immunity in Alzheimer's disease. *Nat. Genet* **49**, 1373–1384 (2017).
70. Tsai, A. P. et al. PLCG2 is associated with the inflammatory response and is induced by amyloid plaques in Alzheimer's disease. *Genome Med* **14**, 17 (2022).
71. Waksmunski, A. R. et al. Pathway Analysis Integrating Genome-Wide and Functional Data Identifies PLCG2 as a Candidate Gene for Age-Related Macular Degeneration. *Invest Ophthalmol. Vis. Sci.* **60**, 4041–4051 (2019).
72. Arakawa, S. et al. Genome-wide association study identifies two susceptibility loci for exudative age-related macular degeneration in the Japanese population. *Nat. Genet* **43**, 1001–1004 (2011).
73. 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).
74. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
75. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinforma.* **12**, 323 (2011).
76. 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).
77. Robin, X. et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinforma.* **12**, 77 (2011).
78. Lundberg, S. & Lee, S. Advances in Neural Information Processing Systems, NIPS'17. (2017).
79. Menon, M. et al. Single-cell transcriptomic atlas of the human retina identifies cell types associated with age-related macular degeneration. *Nat. Commun.* **10**, 4902 (2019).
80. Shekhar, K. & Sanes, J. R. Generating and Using Transcriptomically Based Retinal Cell Atlases. *Annu Rev. Vis. Sci.* **7**, 43–72 (2021).

81. Stuart, T. et al. Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888–1902 e1821 (2019).

Acknowledgements

This study is supported by Career Development Award from Research to Prevent Blindness (RPB) and New Investigator Award by BrightFocus Foundation to RR. RR is also supported by an Unrestricted grant from RPB to Baylor College of Medicine.

Author contributions

R.R. conceived and designed the study. K.M., H.N. A.B., and R.R. conducted the computational analysis and analyzed the data. N.B. performed the luciferase assays. K.M. and R.R. wrote the manuscript. All authors were involved in manuscript revision.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41525-025-00507-2.

Correspondence and requests for materials should be addressed to Rinki Ratnapriya.

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