Heliyon 9 (2023) e16981

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Screening for biomarkers in age-related macular degeneration

Daoxin Han [a], Xiaoli He [b,*]

[a] Department of Ophthalmology, Nanshi Hospital of Nanyang, Henan Province, China
[b] The Second Affiliated Hospital of Harbin Medical University, Heilongjiang, China

ARTICLE INFO

Keywords:
Age-related macular degeneration
Weighted gene Co-expression network analysis
Machine learning
Immune cell infiltration

ABSTRACT

Objective: Age-related macular degeneration (AMD) is a significant cause of blindness, initially characterized by the accumulation of sub-Retinal pigment epithelium (RPE) deposits, leading to progressive retinal degeneration and, eventually, irreversible vision loss. This study aimed to elucidate the differential expression of transcriptomic information in AMD and normal human RPE choroidal donor eyes and to investigate whether it could be used as a biomarker for AMD.
Methods: RPE choroidal tissue samples (46 Normal samples, 38 AMD samples) were obtained from the GEO (GSE29801) database and screened for differentially expressed genes in normal and AMD patients using GEO2R and R to compare the degree of enrichment of differentially expressed genes in the GO, KEGG pathway. Firstly, we used machine learning models (LASSO, SVM algorithm) to screen disease signature genes and compare the differences between these signature genes in GSVA and immune cell infiltration. Secondly, we also performed a cluster analysis to classify AMD patients. We selected the best classification by weighted gene co-expression network analysis (WGCNA) to screen the key modules and modular genes with the strongest association with AMD. Based on the module genes, four machine models, RF, SVM, XGB, and GLM, were constructed to screen the predictive genes and further construct the AMD clinical prediction model. The accuracy of the column line graphs was evaluated using decision and calibration curves.
Results: Firstly, we identified 15 disease signature genes by lasso and SVM algorithms, which were associated with abnormal glucose metabolism and immune cell infiltration. Secondly, we identified 52 modular signature genes by WGCNA analysis. We found that SVM was the optimal machine learning model for AMD and constructed a clinical prediction model for AMD consisting of 5 predictive genes.
Conclusion: We constructed a disease signature genome model and an AMD clinical prediction model by LASSO, WGCNA, and four machine models. The disease signature genes are of great reference significance for AMD etiology research. At the same time, the AMD clinical prediction model provides a reference for early clinical detection of AMD and even becomes a future census tool. In conclusion, our discovery of disease signature genes and AMD clinical prediction models may become promising new targets for the targeted treatment of AMD.

1. Introduction

AMD is the leading cause of severe vision loss in people under 55 years of age in developed countries [1], accounting for 6–9% of

* Corresponding author.
E-mail address: hxl00814@sina.com (X. He).

https://doi.org/10.1016/j.heliyon.2023.e16981
Received 9 November 2022; Received in revised form 24 April 2023; Accepted 2 June 2023
Available online 17 June 2023

legal blindness worldwide[2,3]. AMD affects the complex of photoreceptors, retinal pigment epithelium (RPE), Bruch's membrane (BrM) and choroid. AMD is characterized by the accumulation of drusen, leading to progressive degeneration of photoreceptors and RPE, resulting in the loss of central vision. Progressive central vision loss due to AMD affects the patient's quality of life and psychosocial well-being and imposes a significant economic burden on the patient [4–8]. With the increasing number of older people, the high prevalence of AMD has become a serious public problem in China.

AMD is considered a multifactorial disease, with risk factors including genetic factors, aging, smoking, light exposure, and abnormal nutritional intake [9]. AMD has been divided into two main types: dry AMD and wet AMD [10]. Wet AMD is diagnosed when neovascularization is detected. The diagnosis of AMD is currently determined by clinical symptoms, fundus photography, optical coherence tomography, and fluorescein fundus angiography [11]. Since neither severe visual problems nor discomfort are present in cases of dry AMD, it is not easy to screen for early cases in the population. So far, no effective prevention methods have been identified from early to wet AMD. the clinical situation of AMD highlights the need to develop potential biomarkers to predict the incidence of AMD. The development of AMD involves multiple pathophysiological mechanisms, all of which focus on RPE dysfunction and degeneration. Therefore, we obtained samples from the GEO (GSE29801) database (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE29801) and analyzed normal and AMD transcriptome information to find predictive genes and models, The analysis process of this study is shown in the flow diagram 1and 2.

2. Methods and materials

2.1. Materials

The transcriptome information of RPE choroidal tissue samples came from the GEO (GSE29801) database, containing 46 normal and 38 AMD samples. Gene ID conversion was done by DAVID Online Tools (https://david.ncifcrf.gov/), and gene expression levels were taken as mean values if there were duplicate genes. The gene expression was normalized by taking log2 for gene expression. Finally, 19289 genes and 84 samples were obtained as expression matrices.

2.2. Screening for differential genes

Extramacular RPE-choroid, AMD (Ex RPE-AMD) was the Treat group; Extramacular RPE-choroid, normal (Ex RPE-Normal) was the Control group. Two methods were used to obtain differential genes. In the first one, based on GEO2R online line difference analysis, the endpoint was set as adjust. P. value < 0.05. In the second one, using library(limma)R package line difference analysis, the endpoint was set as P. value < 0.05. The different genes obtained by the two methods were intersected to be the final difference genes.

2.3. Differential gene pathway enrichment analysis

To better understand the biological significance of common differential genes, we used "clusterProfiler," "org.Hs.eg.db", "enrichplot, "ggplot2″ R package to perform GO (Gene Ontology) and KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway enrichment analysis.

2.4. Disease signature gene screening

2.4.1. LASSO (least absolute shrinkage and selection operator) regression analysis

The LASSO regression model was constructed using the "glmnet" R package, and the number of genes corresponding to the most minor cross-validation point was the optimal number of trait genes.

2.4.2. SVM (support vector machine) algorithm

The SVM model is constructed using the "e1071″, "kernlab," and "caret" R packages, and the number of genes corresponding to the most minor cross-validation point is The number of genes corresponding to the most minor cross-validation point is the number of optimal feature genes.

The number of genes corresponding to the most minor cross-validation point is the optimal number.

2.5. Accuracy validation of disease signature genes and pathway enrichment analysis

ROC (receiver operating characteristic) curves were constructed using the "glmnet" and "pROC" R packages. The larger the area under the ROC curve, the higher the accuracy of the gene or model for disease diagnosis. In order to further clarify the biological functions involved in the disease signature genes, GSVA(Gene set variation analysis) pathway enrichment analysis was performed using the "GSEABase" and "GSVA" R packages. pvalue$<$0.05 was considered statistically significant.

2.6. Immune cell infiltration

The samples were analyzed for immune cell infiltration using the 'preprocessCore,' 'parallel,' and 'e1071' R packages. Immune cells differentially expressed in the Treat and control groups were screened using the Wilcox test and the 'vioplot' R package. The cut-off

point was set at Pvalue<0.05. The samples were analyzed by using "limma," "reshape2″, "tidyverse," "ggplot2″, "tidyverse," and "ggplot2". "ggplot2″ R package was used to clarify the relationship between disease signature genes and immune cell infiltration.

2.7. Clustering analysis and validation

Cluster analysis was performed using ConsensusClusterPlus, coalescent km clusters with 1-Pearson correlation distances, and resampling 80% of the samples 10 times with k values ranging from 2 to 9. The optimal number of fractions was determined using empirical cumulative distribution function plots. PCA analysis was performed using the "limma" and "ggplot2″ R packages to detect clustering effects.

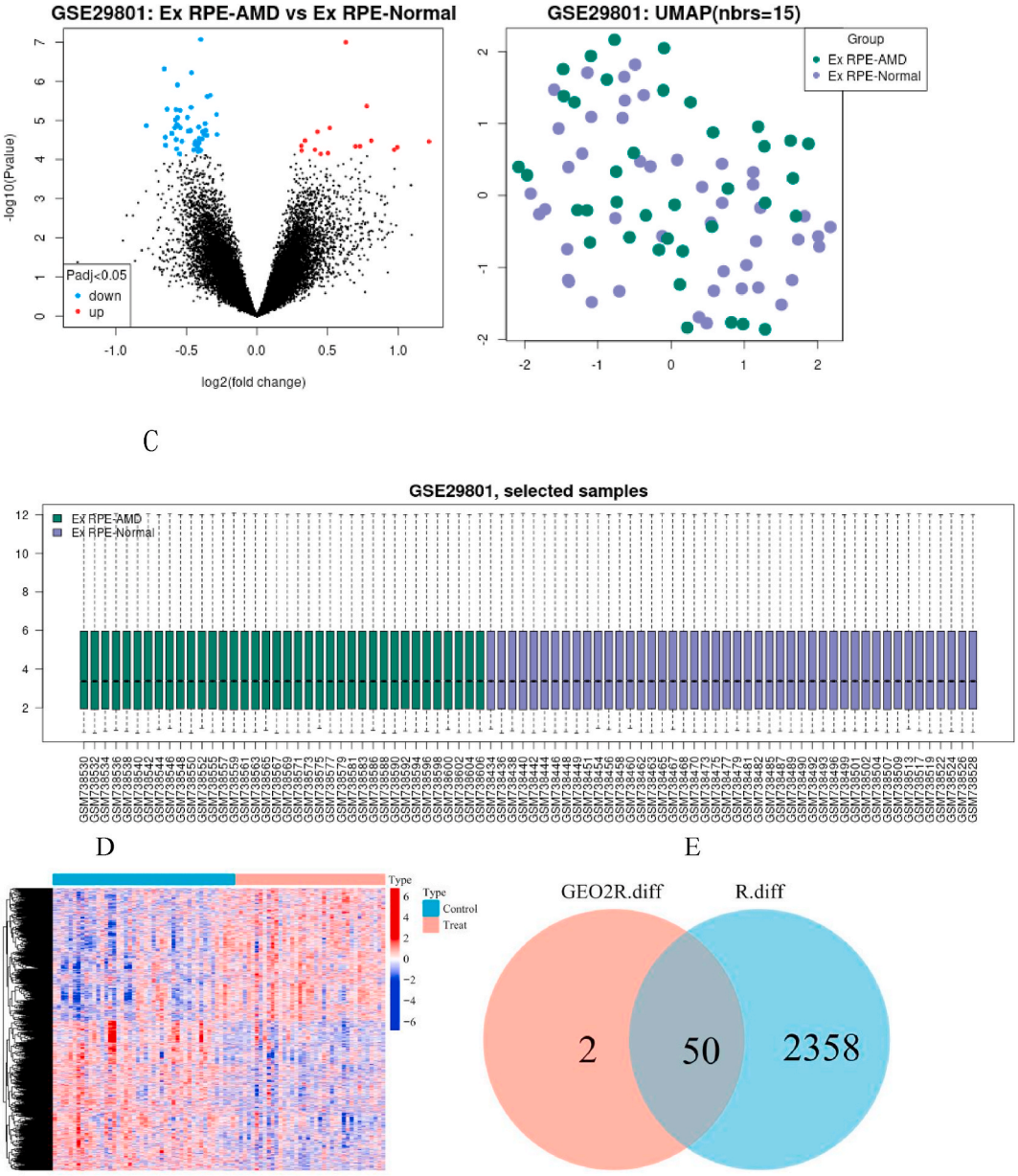

Fig. 1. Differential expressed genes analysis (A) Volcano plot of 52 differential genes. (B) Distribution of AMD and normal samples in GSE29801 in UMAP. (C) Gene expression levels of AMD and normal samples in GSE29801. (D) Heatmap of 2048 differential genes. (E) The two methods obtained the intersection of differential genes.

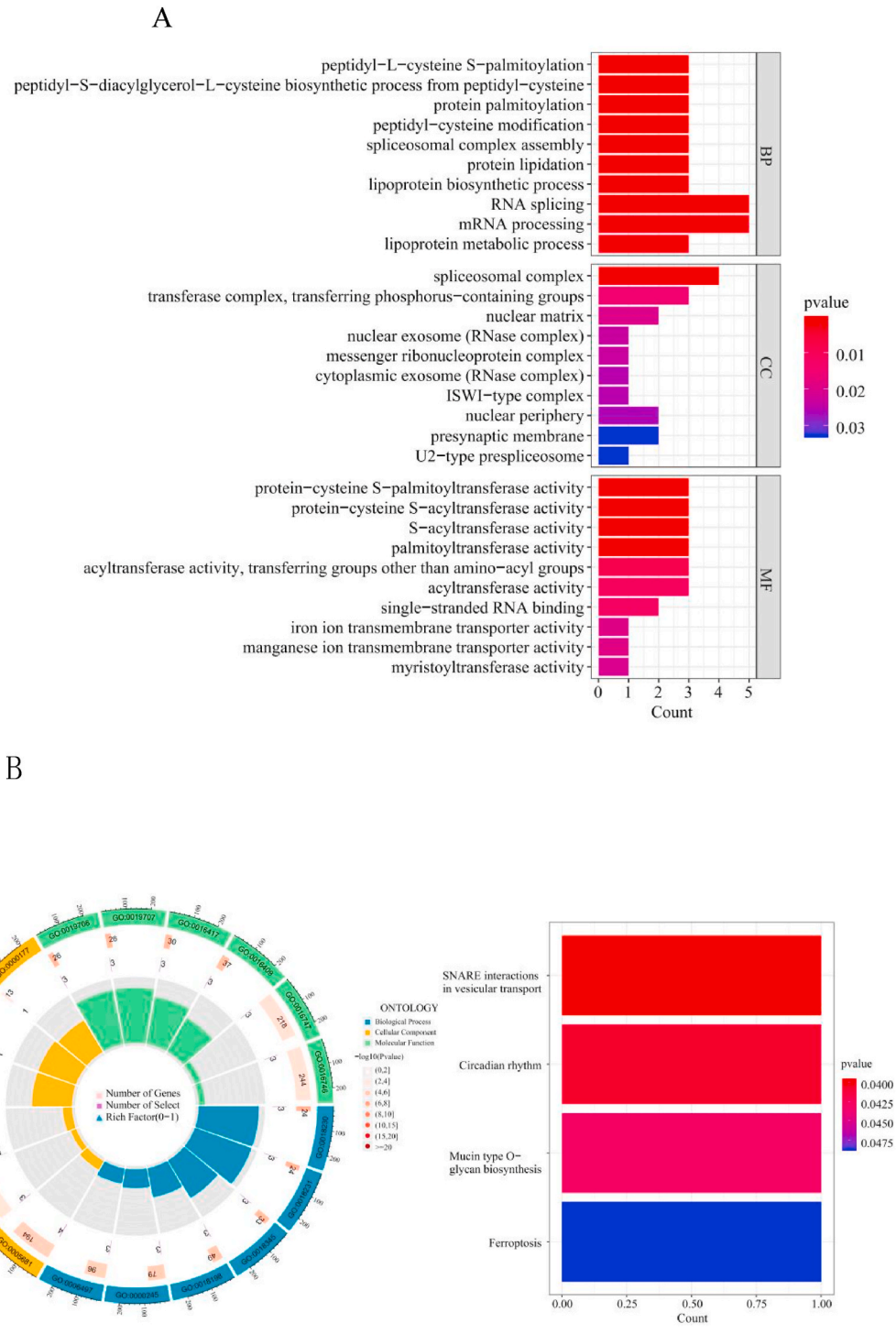

Fig. 2. (A) The GO function's actual name appears on the left vertical axis. , the right vertical axis is the taxonomic name of the function, and the horizontal axis is the number of differential genes released to the corresponding GO function. (B) Circle diagram of GO enrichment analysis. (C) KEGG enrichment analysis, the vertical axis is the specific name of the KEGG function, and the horizontal axis is the number of differential genes released to the corresponding function.

2.8. Immune cell infiltration

The immune cell distribution of the samples was calculated by the CIBERSORT algorithm using the R software to obtain the proportional fraction of immune cells for each sample in the data set. The "limma" R package was used to compare the immune cell infiltration for differences.

2.9. Cluster pathway enrichment analysis

In order to further clarify whether there are differences in the biological functions of different fractal clusters, "reshape2″, "ggpubr," "limma," "GSEABase," and "GSVA" R packages were used for GSVA pathway enrichment analysis. "GSEABase" and "GSVA" R packages were used for GSVA pathway enrichment analysis.

2.10. Cluster-wise WGCNA analysis

WGCNA is an R package that constructs gene co-expression networks from many genes and identifies co-expression modules [12]. Firstly, we used the "WGCNA" package in R to cluster the top 25% of the most fluctuating genes in the samples to assess whether there are significant outliers. Secondly, we identified gene co-expression modules based on the topological overlap, detected the modules using hierarchical clustering and dynamic tree-cutting functions, and calculated gene significance (GS) and module affiliation (MM). Modules were linked to clinical traits (sub-clusters). Correlations between these modules are calculated, and heat maps are executed to show the independence between these modules. Finally, we visualize the trait gene network.

2.11. Machine learning model construction

Using the above-obtained module genes, sample GSE29801 transcriptome expression levels and "caret," "DALEX," "ggplot2″, "randomForest," "kernlab," "xgboost," "pROC "R package to construct RF random forest tree model, SVM machine learning model, XGB model, and GLM model. The samples were divided into Treat and Test groups, with the Treat group accounting for 70% of the total samples for constructing the models and the Test group accounting for 30% of the total samples for verifying the accuracy of the models. The genes obtained from the four methods were analyzed for importance, and the top 5 genes with the highest importance scores were output as the predicted genes.

2.12. Construction of nomo plots

To predict the incidence of AMD, we constructed column line graphs using the importance genes obtained above and the "rms" and "rmda" R packages and evaluated the accuracy of the column line graphs using decision curves and calibration curves.

2.13. Methods

Online analysis of variance using GEO2R rows, adjpvalue$<$0.05, was considered statistically significant. Using R4.21 row-wise analysis of variance and co-expression analysis, pvalue$<$0.05 was considered statistically significant.

3. Results

3.1. Screening of differential genes

52 genes differentially expressed in AMD and normal samples were obtained using GEO2R online analysis (Fig. 1 A). The AMD and normal samples were evenly distributed in Uniform Manifold Approximation and Projection(UMAP) (Fig. 1B) with good consistency in gene expression levels (Fig. 1C), indicating that the differential genes obtained using GEO2R online analysis were reliable. Differential analysis using the limma R package line yielded 2048 genes (Fig. 1 D), of which 1275 genes were down-regulated, and 773 genes were up-regulated in AMD (Supplementary file 1). A total of 50 differential genes were obtained by taking the intersection of the two methods (Fig. 1E) (Supplementary file 2).

3.2. Differential gene GO and KEGG enrichment analysis

The results of GO enrichment analysis showed that the pathways enriched in Biological Process were lipoprotein metabolic process, mRNA processing, RNA splicing, lipoprotein biosynthetic process; The pathways enriched on Cell Component were transferase complex, transferring phosphorus-containing groups, spliceosomal complex Molecular Function, the enriched pathways are palmitoyltransferase activity, S-acyltransferase activity, protein-cysteine S-acyltransferase activity, S-acyltransferase activity, and S-acyltransferase activity; The enriched pathways on the Molecular Function were palmitoyltransferase activity, S-acyltransferase activity, protein-cysteine S-acyltransferase activity, and protein-cysteine S-palmitoyltransferase activity. KEGG enrichment analysis identified the above differential genes that were metabolically active in the Ferroptosis, Mucin type O-glycan biosynthesis, Circadian rhythm, and and SNARE interactions in vesicular transport pathways.

3.3. Disease signature gene screening and validation

LASSO obtained the minimum value of Log(λ) corresponding to the number of genes of 16 (Fig. 3A) (Supplementary file 3), and a vertical line was drawn at the value selected by 10-fold cross-validation. As the value of λ decreases, the compression of the model increases, and the selection of essential variables by the model increases (Fig. 3B). The minimum value of SVM corresponds to a number of genes of 40 (Fig. 3C) (Supplementary file 4). The 15 genes were obtained after taking the intersection (Fig. 3D) (Supplementary file 5). The heat map is shown in Fig. 3E. The area under the AUC of these 15 genes was more significant than 0.7 (Fig. 3F), indicating that these genes alone predicted AMD disease with high accuracy. The area under the AUC of the model was 0.989, with a 95% fluctuation range of 0.967–1 (Fig. 3G), indicating the excellent prediction performance of this model.

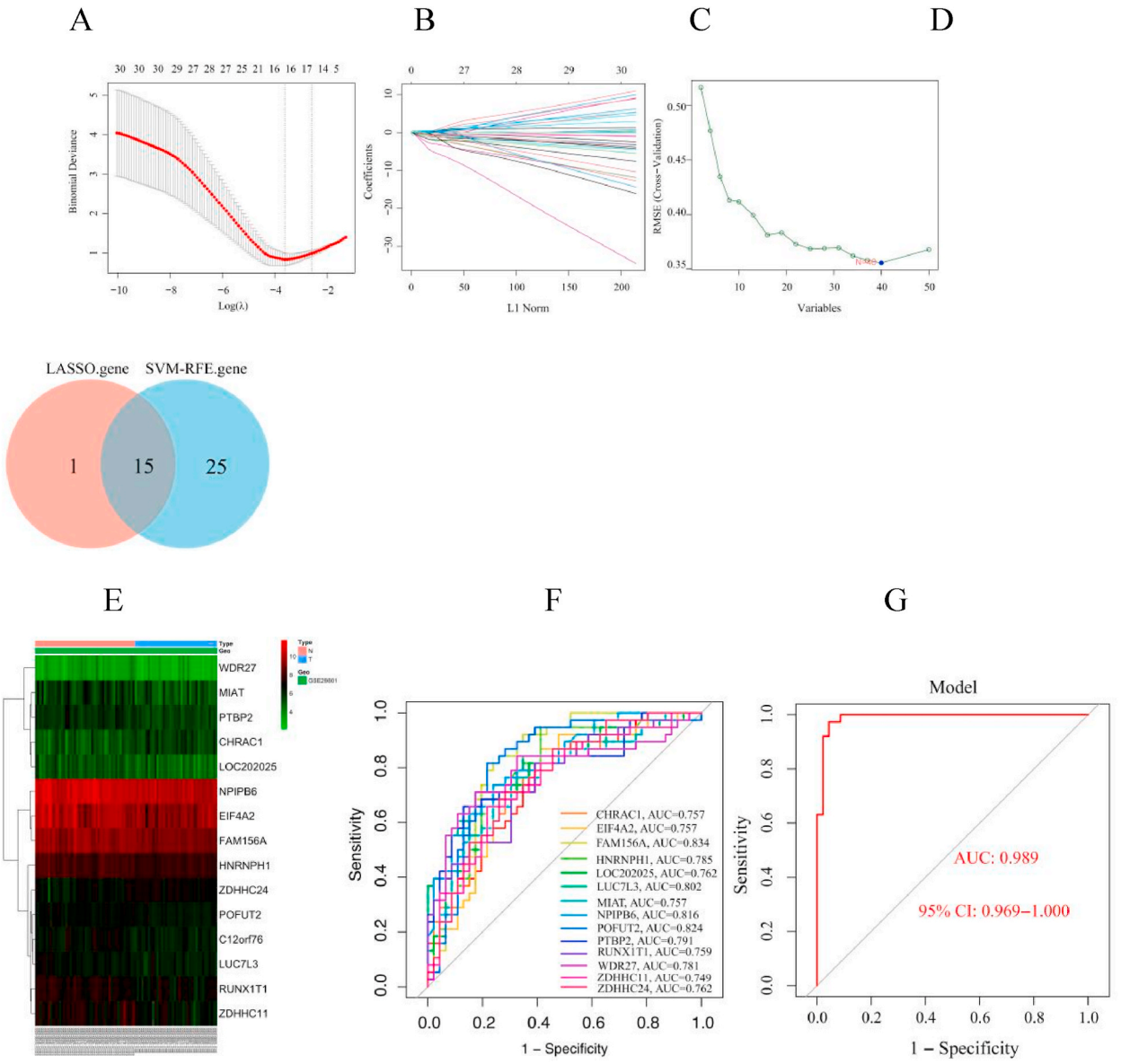

Figure 3. (A) Results of cross-validation. The values in the middle of the two dashed lines are the range of positive and negative standard deviations of log(λ). The dashed line on the left indicates the value of the harmonic parameter log(λ) when the model error is minimal. (B). Distribution of LASSO coefficients for 30 variables. (C). The performance of the feature subset selected by SVM on the dataset and the value of the horizontal coordinate corresponding to the lowest point of the curve indicates the optimal number of genes. (D) ROC curve of 15 genes, the larger area under the curve, indicates the better prediction performance of this gene. (E) ROC curves of the 15-gene model. The larger the area under the curve, the better the prediction performance of this gene.

3.4. Disease signature gene pathway enrichment analysis

GSVA enrichment analysis showed that 5 of the 15 differential genes were closely associated with the glucose metabolism pathway, such as C12orf76 was upregulated in the other glycan degradation pathway and downregulated in the Maturity onset diabetes of the young pathway (Fig. 4A). LUC7L3 was upregulated in the type I diabetes mellitus pathway (Fig. 4B). LOC202025 was downregulated in the taurine and hypotaurine metabolism pathway (Fig. 4C)·NPIPB6 was downregulated in the taurine and hypotaurine metabolism pathway (Fig. 4D). ZDHHC11 was downregulated in the other glycan degradation pathway (Fig. 4E). CHRAC1 (Fig. 4F) and EIF4A2

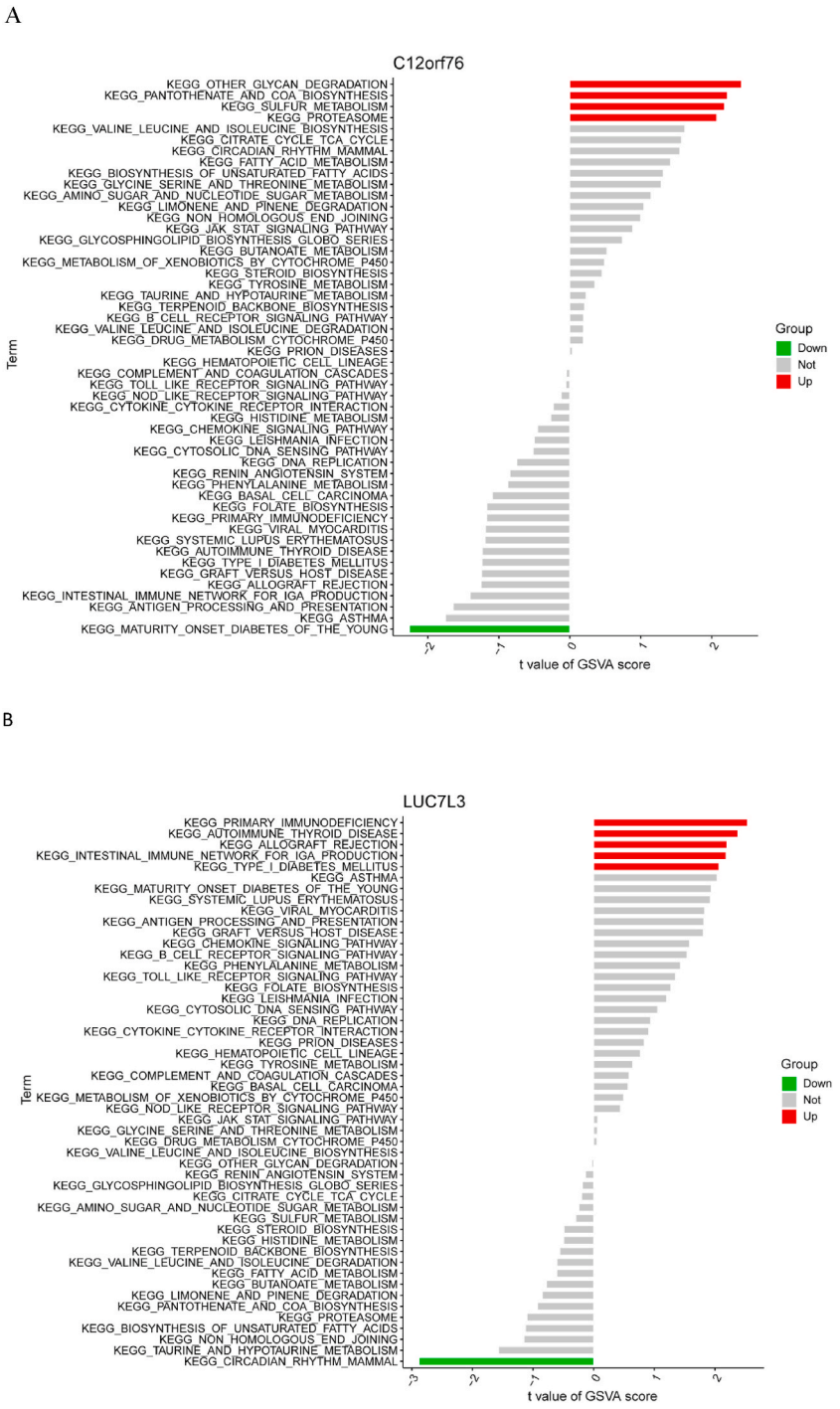

Fig. 4. GSVA enrichment of 7 differential genes. c12orf76 (A), LUC7L3 (B), LOC202025 (C), NPIPB6 (D), ZDHHC11 (E), CHRAC1(F), EIF4A2(G).

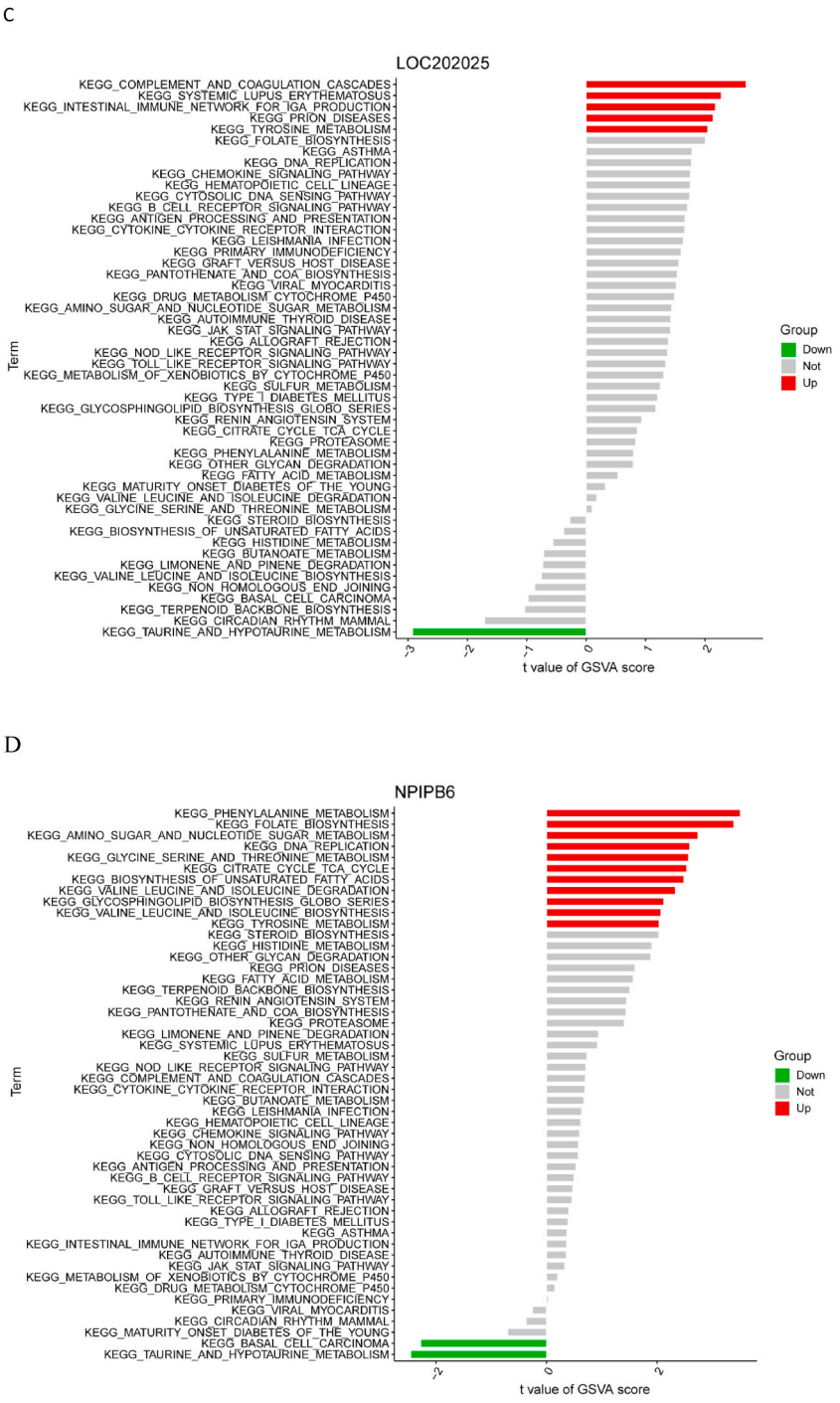

Fig. 4. (*continued*).

(Fig. 4G) are associated with the fatty acid metabolic pathway.

FAM156A (Supplementary F. 4A), MIAT (Supplementary F. 4B), RUNX1T1 (Supplementary F. 4C), WDR27 (Supplementary F. 4D) and ZDHHC24 (Supplementary F. 4E) were associated with Pantothenate and coa biosynthe, cytokine receptor interaction and limonene and pinene degradation pathways. PTBP2 (Supplementary F. 4F), HNRNPH1 (Supplementary F. 4G), and POFUT2 (Supplementary F. 4H) were not found to be significantly enriched in the biological pathway.

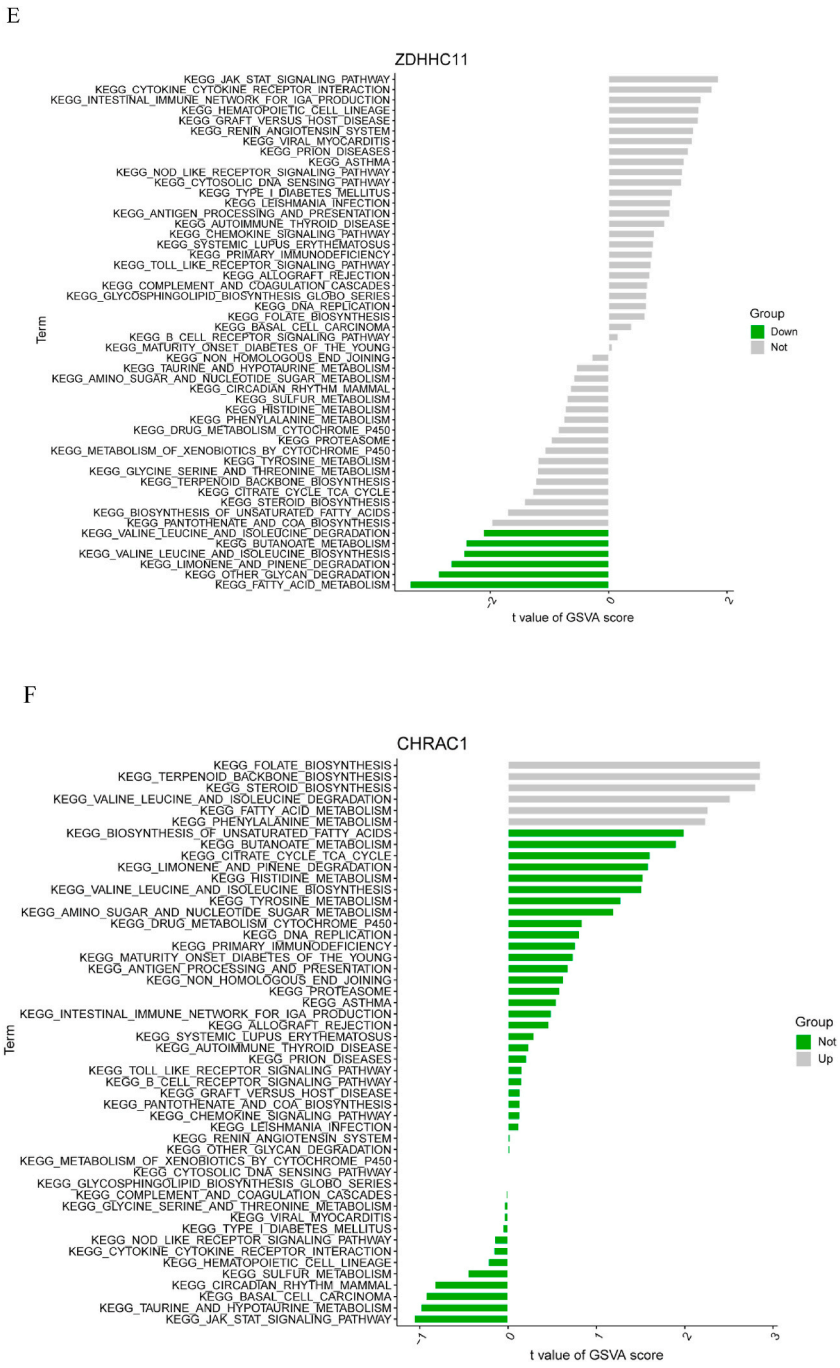

Fig. 4. (*continued*).

3.5. Immune cell infiltration

Immune cell infiltration was obtained for each sample using the R package (Supplementary file 6). Immune cell Neutrophils were highly expressed in the control group (normal samples) with pvalue = 0.0497 (Fig. 5A), and the difference was statistically significant. 15 disease signature genes were strongly associated with immune cell infiltration, such as LUC7L3 positively correlated with M0 macrophages ($p < 0.05$), HNRNPH1 negatively correlated with M2 macrophages ($p < 0.05$), RUNX1T1 and WDR27 were positively correlated with NK cells activated ($P < 0.05$) (Fig. 5B).

G

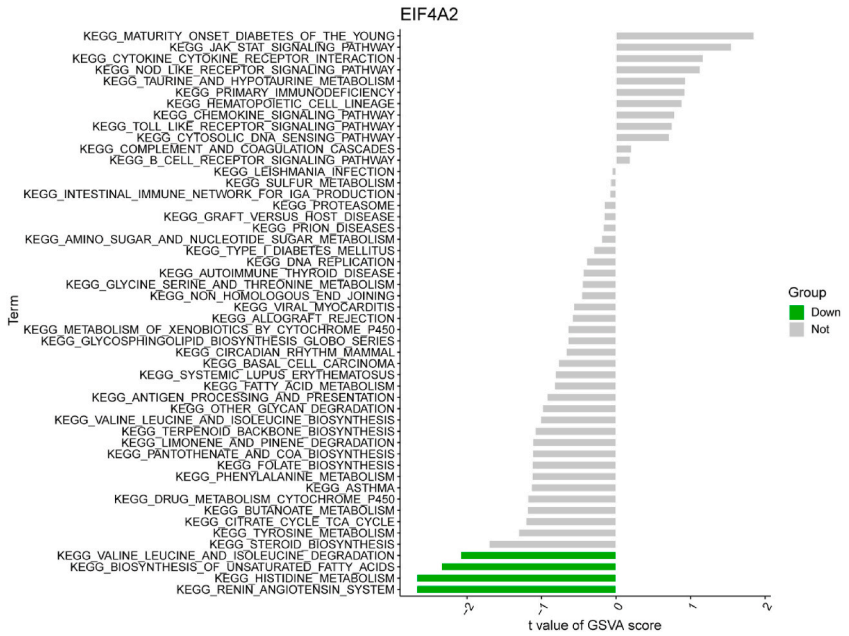

Fig. 4. (*continued*).

3.6. Cluster analysis and detection

After a comprehensive evaluation, the number of clusters with the highest average consistency within the group was selected as K = 2 (Fig. 6A and B). Fig. 6C and D shows that CHADL, HNRNPH1, SCGB1D2, SLC39A8, and ST6GALNAC2 were highly expressed in C2, and PCA analysis showed that this cluster could well distinguish C1 from C2 (Fig. 6E).

3.7. Inter-cluster immune cell infiltration and gsva pathway enrichment analysis

In Fig. 7A, we obtained little difference in immune cell infiltration between C1 and C2 clusters. From Fig. 7B, we obtained that the immune cells with higher enrichment in C2 are T cells CD4 Memoryactivated and B cells naive. The immune cells with higher enrichment in C1 are Monocytes and Neutrophils. GSVA analysis showed C2 clusters in the NEUROTROPHIN signaling pathway, natural killer cell-mediated cytotoxicity, regulation of actin cytoskeleton, autoimmune thyroid disease, tubule-like receptor signaling pathway, FC gamma R-mediated phagocytosis, leishmaniasis infection point-like receptor signaling pathway, chronic myeloid leukemia, viral myocarditis, systemic lupus erythematosus enriched in the lupus erythematosus pathway. The C1 cluster was enriched in the peroxisome, lysine degradation, β-alanine metabolism, oxidative phosphorylation, valine leucine, and isoleucine degradation, terpene skeleton biosynthesis, aspartate and glutamate metabolism, nitrogen metabolism, unsaturated fatty acid biosynthesis, and butyric acid metabolic pathways (Fig. 7c).

3.8. Weighted gene co-expression network

A total of 6 different gene co-expression modules were generated by WGCNA raw letter analysis. Correlations between each module and clinical features (C1 and C2) were tested. The signature gene dendrogram and heat map indicated that the MEturquoise modules were highly correlated with C1 and C2 (Fig. 8A–C). The results of the eigengene dendrogram and heat map plots showed that the MEturquoise module was positively correlated with C2 and negatively correlated with C1 (Fig. 8D). Gene salience and module membership were plotted for the MEturquoise module, indicating that the module was significantly positively correlated with C2 (Fig. 8E). Finally, 52 genes for the MEturquoise module were obtained.

3.9. Machine learning model and clinical prediction model construction and validation

Through the construction of R, SVM, XGB, and GLM machine learning models, we can obtain that the predictive performance of GLM and SVM models is better in Boxplots of |residual|. ROC curve, Reverse cumulative distribution of |residual| both obtained better predictive performance of RF and SVM models, and the SVM model was selected for the following analysis in a comprehensive consideration. Feature Importance shows 10 genes, and we selected the five genes with the highest scores for constructing the nomo

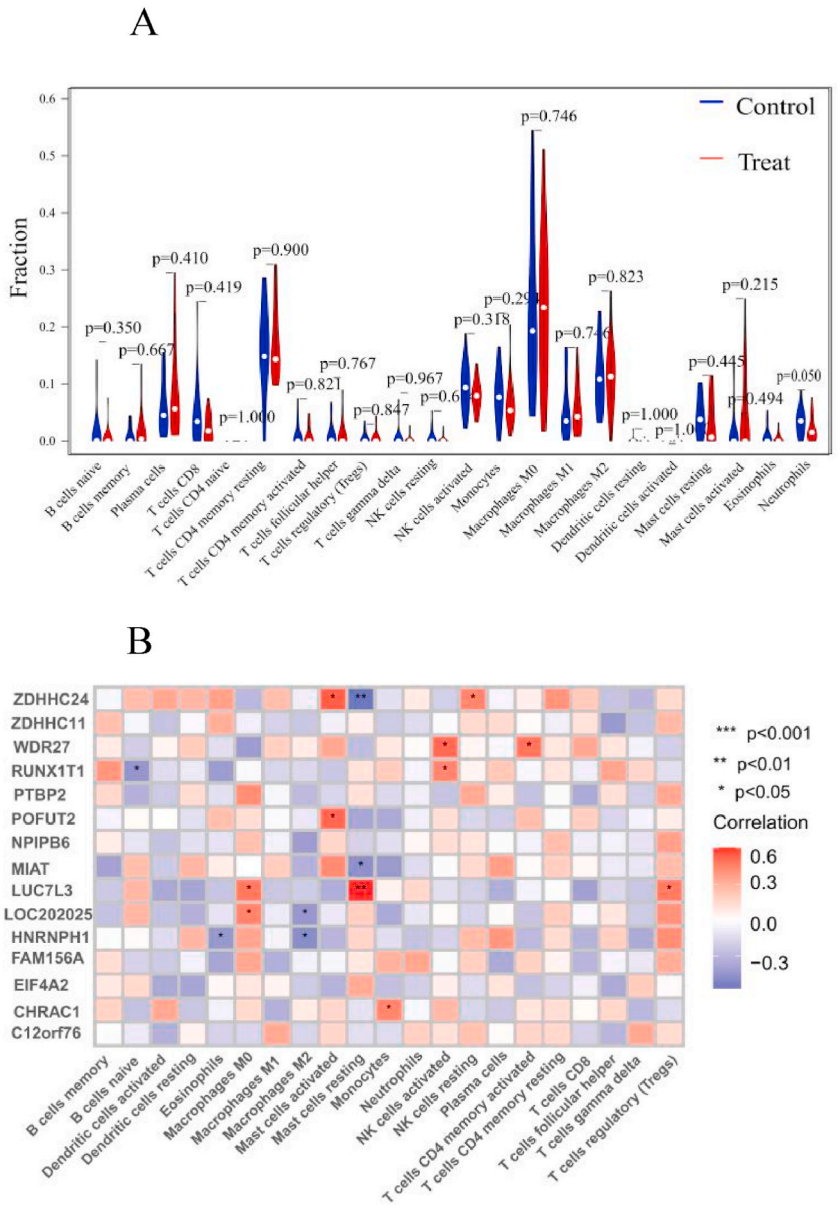

Fig. 5. (A) Immune cell infiltration in control and Treat groups; (B) Co-expression of 15 disease signature genes with immune cells.

plot of the clinical prediction model. From the nomo plot, we can see those different gene expressions correspond to different points, and the sum of 5 gene expressions corresponding to different points in Total points, and Total different points correspond to different Risks of Disease. The Risk of Disease is the probability that we presume an individual patient has AMD. The calibration curve shows that the solid and dashed lines are very close to each other, which proves that the model is more accurate. The red curve of the DCA decision curve is far from the All curve, which proves that the model is more accurate. In conclusion, the accuracy of our constructed model is high.

4. Discussion

AMD is a major cause of blindness, beginning with the formation of the outer blood retinal barrier (oBRB) by the retinal pigment epithelium (RPE), Bruch's membrane and choriocapillaris. The risk of AMD development remains poorly understood due to the lack of relevant predictive models for the human retinal pigment epithelium (RPE).

We obtained 84 RPE choroidal tissue samples (46 normal samples and 38 AMD samples) from the GEO (GSE29801) database. Gene expression in normal and AMD samples was first analyzed online using GEO2R and 52 differentially significant genes were obtained;

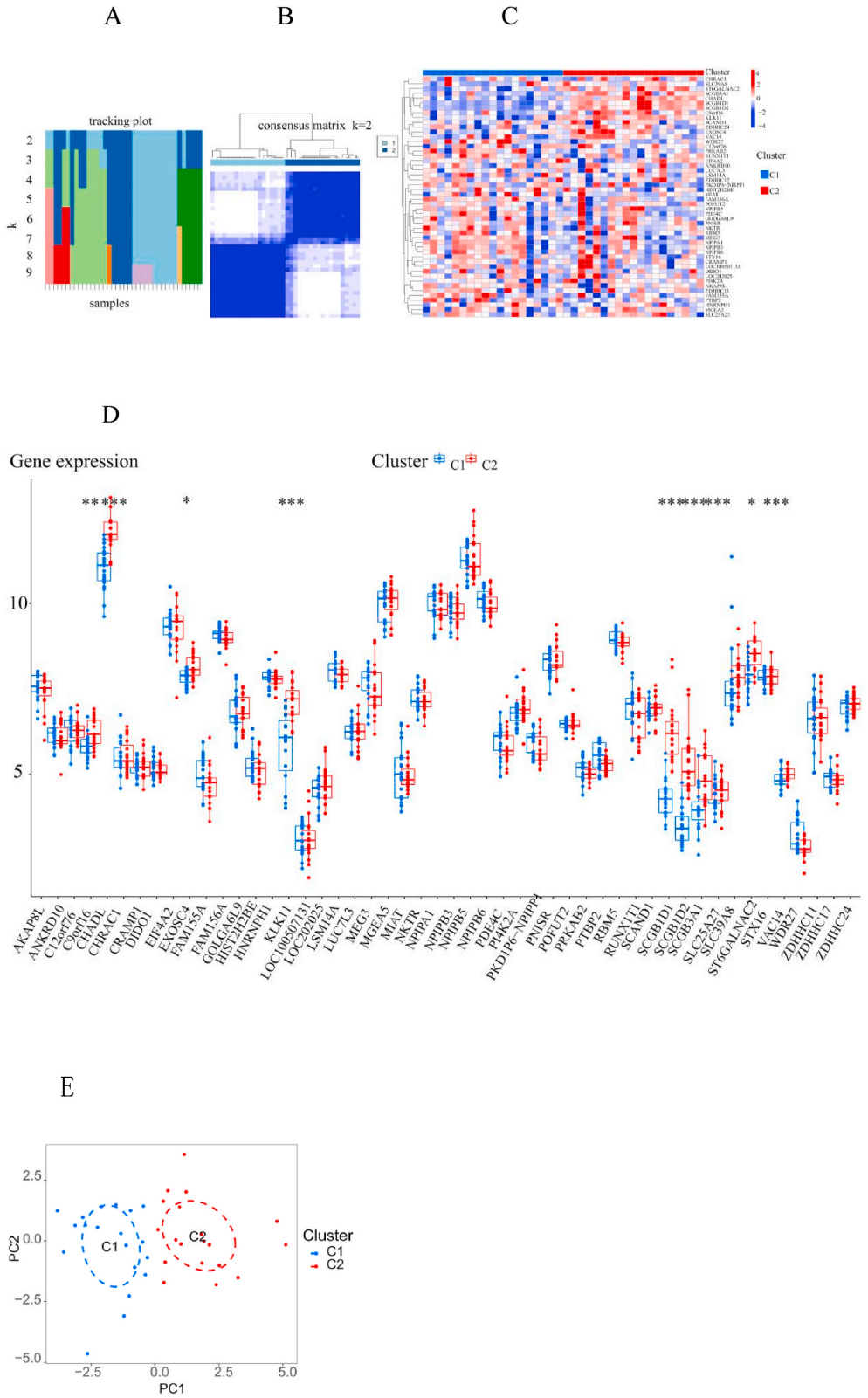

Fig. 6. (A) Different colors in the horizontal axis indicate different clustering, and other colors within the color indicate the presence of impurities, the more impurities contained, the worse the clustering effect, and the vertical axis indicates different K values. (B) The clustering situation when K = 2. (C–D) Expression of 15 disease signature genes in clusters C1 and C2, heatmap (C), boxplot (D). (E) PCA analysis.

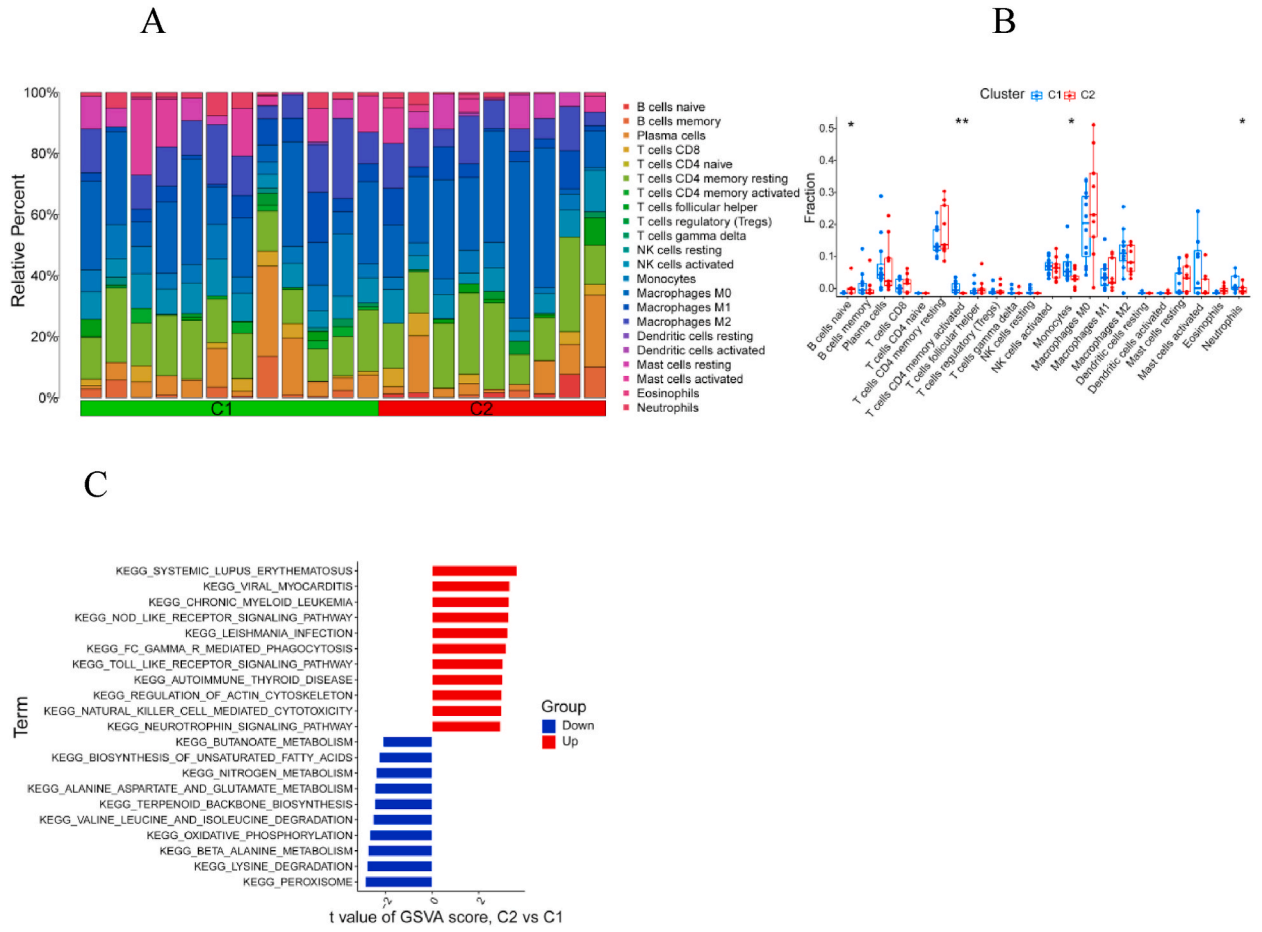

Fig. 7. (A) Infiltration of immune cells in C1 and C2 clusters. (B) Differential analysis of immune cell infiltration in C1 and C2 clusters. (C) GSVA pathway enrichment analysis of immune cells in C1 and C2 clusters.

2408 genes were also screened using the R package. The genes obtained by these two methods were taken to intersect to obtain 50 differential genes. To clarify which metabolic processes and pathways these differential genes are mainly involved in, we performed GO enrichment analysis and concluded that these differential genes are mainly involved in protein synthesis and RNA editing. To further explore which metabolic pathways these genes are involved in, we performed KEGG pathway enrichment and found that these differential genes were significantly enriched in the Ferroptosis metabolic pathway. Thus, we hypothesized that protein synthesis and RNA editing in the Ferroptosis metabolic pathway might be involved in the development of AMD. The results of several scholarly studies confirm the validity of our research and speculations, for example, Wei, Ting-Ting suggests that AMD induces Ferroptosis in retinal pigment epithelial cells and that inhibition of Ferroptosis may be a potential target for the treatment of AMD [13], Sun, Yun scholars concluded that Glutathione depletion induces ferroptosis in retinal pigment epithelial cells [14]. In conclusion, the differential genes we screened accurately respond to the biological pathways that may affect AMD, and the use of these differential genes is a good theoretical basis for further research.

Second, we used two machine learning models to screen for disease signature genes. Sixteen genes were obtained using the LASS0 algorithm and 40 genes were obtained using the SVM algorithm, and the genes obtained by these two algorithms were intersected to obtain 15 genes as disease signature genes. The ROC curves showed that the evaluation of these genes and the predictive performance of the model were good. To further verify whether our screened disease signature genes are representative of the disease, we performed GSVA enrichment analysis and obtained c12orf76 (Fig. 4A), LUC7L3 (Fig. 4B), LOC202025 (Fig. 4C), NPIPB6 (Fig. 4D), ZDHHC11 (Figure These five genes are involved in glucose metabolism, and CHRAC1 (Fig. 4F) and EIF4A2 (Fig. 4G) are involved in fatty acid metabolism. It is well known that diet and nutrition have a strong epidemiological link to the onset and progression of AMD. Rowan, Sheldon scholars have argued that A lower glycemic diet is associated with protection against AMD in humans, and switching from a higher to a lower glycemic diet prevents the AMD phenotype in mice [13]. Pennington, Katie L also showed that lipid metabolism is involved in the development of AMD [14]. The two studies mentioned above confirm the accuracy of our model and that we can regulate the progression of AMD and even cure or avoid it completely by regulating the expression of these genes. In immune cell infiltration analysis, we obtained that several genes are closely related to the level of immune cell infiltration, such as HNRNPH1, which is negatively correlated with M2 macrophages, and related studies have shown that the number of M2 macrophages is increased

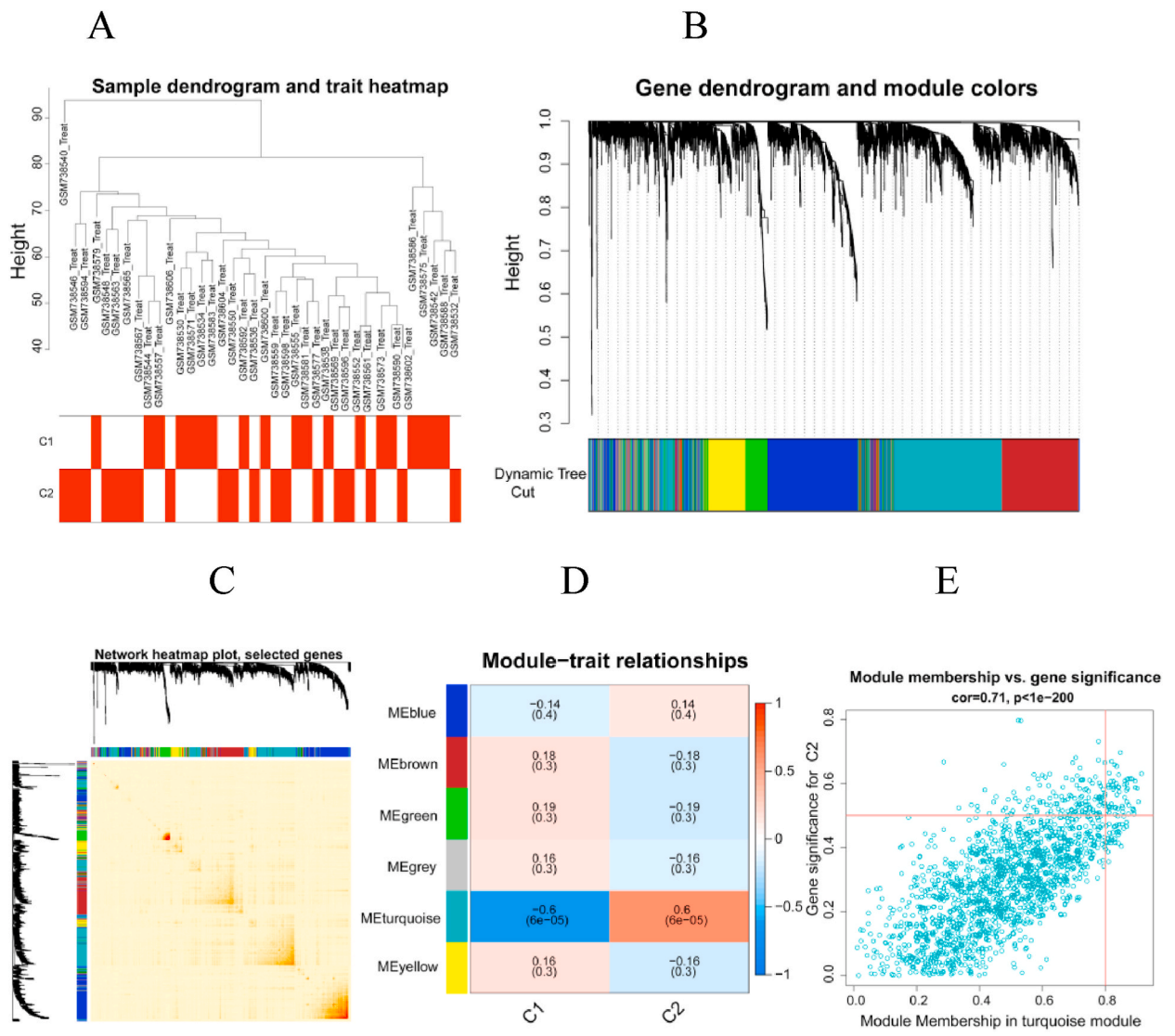

Fig. 8. Weighted gene co-expression network analysis. (A) Sample clustering dendrogram detecting outliers in C1 and C2 clusters of WGCNA. (B) Clustered dendrogram of genes with topology-based overlap dissimilarity and the specified module colors. (C) Gene co-expression network visualized as a heat map. Lighter colors indicate low co-expression, and darkening red indicates high co-expression. Darker colors on the diagonal lines are modules. (D) Module-trait associations. Each row corresponds to a module, and each column corresponds to a trait. Each cell contains the corresponding correlation and P value. The table is color-coded by correlation according to the color legend. (E) Scatter plot between each gene in the module and the C2 cluster.

in AMD [15]. And it has been shown that melatonin attenuates choroidal neovascularization by regulating macrophage/microglia polarization through inhibition of the RhoA/ROCK signaling pathway [16]. This suggests that macrophages play an important role in the development of AMD, and the 15 disease signature genes we obtained may be targets for future immunotherapy of AMD. In conclusion, Ferroptosis, abnormal glucose metabolism, fatty acid metabolism, and the degree of immune cell infiltration may play a vital role in the development of AMD and may be a direction for future treatment.

Third, to further investigate the transcriptome information characteristics of AMD samples, we performed cluster clustering analysis based on these 50 differential gene expressions. AMD samples were classified into two clusters, C1 and C2. From the immune cell infiltration, we obtained that C1 was related to oxid ative phosphorylation, while the C2 cluster was related to glycine serine and threonine metabolism and glycosphingolipid biosynthesis ganglio series. The results again validate that the protein synthesis and glucose metabolism we obtained above are involved in the development of AMD.

Fourth, based on the transcriptome information of 84 samples obtained from the GEO database, we performed WGCNA to screen the optimal module MEturquoise and the 52 genes it contained, and MEturquoise was positively correlated with C2 clinical features (Fig. 8D). The five highest scoring predictive genes in the SVM model were HGD, KDF1, GJB1, CNIH3 and CHADL, and we used these

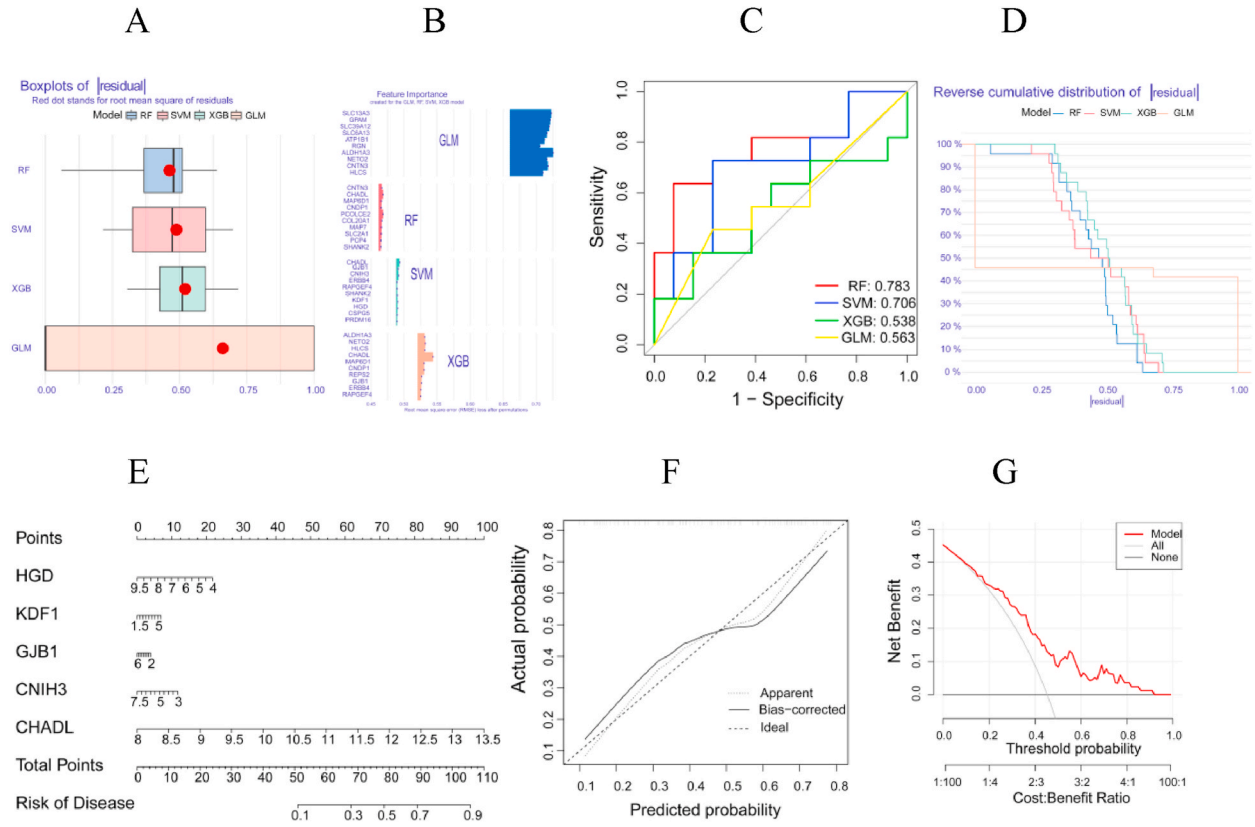

Fig. 9. (A) Boxplots of |residual|, the vertical axis indicates the model type, and the horizontal axis indicates the residual size; The smaller the residual, the higher the model accuracy. (B) Feature Importance. The Vertical axis indicates the model type; the horizontal axis indicates Root mean square error (RMSE) loss after permutations, the smaller the Root mean square error (RMSE) loss after permutations, the higher the model accuracy. The smaller the Root mean square error (RMSE) loss after permutations, the higher the model accuracy. (C) ROC curve. The larger the area under the curve, the higher the model's accuracy. (D) Reverse cumulative distribution of |residual|. The smaller the |residual|, the higher the model's accuracy. (E) Nomo clinical prediction model. (F) Calibration curves. (G) DCA decision curve.

five predictive genes to construct an AMD clinical prediction model. The ROC and DCA curves showed that this model has accurate predictive ability. The expression levels of these five genes can be measured before the onset of AMD to predict the risk of AMD in individual patients, providing a theoretical basis for treating the disease before it occurs.

5. Conclusion

In conclusion, we combined machine learning model, WGCNA and cluster clustering analysis, not only predicted the possible pathogenesis of AMD, but also constructed a clinical prediction model for AMD, which provides a theoretical basis for the treatment of AMD and screening of high-risk groups. This study has some limitations due to the limitations of disease type and database; experimental validation due to the influence of COVID-9. If conditions permit, we will conduct clinical and animal trials to further explore the cause of AMD. We believe that the treatment of AMD will gain further progress caused by our study.

Author contribution statement

Daoxin Han, Xiaoli He: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Data availability statement

Data will be made available on request.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to

influence the work reported in this paper.

Acknowledgements

The authors thank the GEO StarBase and GSVA, DAVID Online Tools for allowing them to upload the useful datasets.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2023.e16981.

References

[1] R. Klein, B.E.K. Klein, K.L.P. Linton, Prevalence of age-related maculopathy, Ophthalmology 99 (6) (1992) 933–943, https://doi.org/10.1016/S0161-6420(92)31871-8.
[2] W.L. Wong, et al., Global prevalence of age-related macular degeneration and disease burden projection for 2020 and 2040: a systematic review and meta-analysis, Lancet Global Health 2 (2) (2014) e106–e116, https://doi.org/10.1016/S2214-109X(13)70145-1.
[3] Updates on the Epidemiology of Age-Related Macular Degeneration, Asia-Pacific Journal of Ophthalmology [Preprint] (2017), https://doi.org/10.22608/APO.2017251.
[4] S.L. Tyson, Utility values associated with blindness in an adult population, Evid. Base Eye Care 3 (1) (2002) 54–55, https://doi.org/10.1097/00132578-200201000-00026.
[5] A. Hernandez Trillo, C.M. Dickinson, The impact of visual and nonvisual factors on quality of life and adaptation in adults with visual impairment, Investigative Opthalmology & Visual Science 53 (7) (2012) 4234, https://doi.org/10.1167/iovs.12-9580.
[6] R.A. Williams, The psychosocial impact of macular degeneration, Arch. Ophthalmol. 116 (4) (1998) 514, https://doi.org/10.1001/archopht.116.4.514.
[7] K. Spooner, The burden of neovascular age-related macular degeneration: a patient's perspective, Clin. Ophthalmol. 12 (2018) 2483–2491, https://doi.org/10.2147/OPTH.S185052.
[8] M.M. Brown, et al., Age-related macular degeneration: economic burden and value-based medicine analysis, Can. J. Ophthalmol. 40 (3) (2005) 277–287, https://doi.org/10.1016/S0008-4182(05)80070-5.
[9] A. Lançon, R. Frazzi, N. Latruffe, Anti-oxidant, anti-inflammatory and anti-angiogenic properties of resveratrol in ocular diseases, Molecules 21 (3) (2016) 304, https://doi.org/10.3390/molecules21030304.
[10] J.K. Luttrull, B.W.L. Margolis, Functionally guided retinal protective therapy for dry age-related macular and inherited retinal degenerations: a pilot study, Investigative Opthalmology & Visual Science 57 (1) (2016) 265, https://doi.org/10.1167/iovs.15-18163.
[11] M. van Lookeren Campagne, E.C. Strauss, B.L. Yaspan, Age-related macular degeneration: complement in action, Immunobiology 221 (6) (2016) 733–739, https://doi.org/10.1016/j.imbio.2015.11.007.
[12] P. Langfelder, S. Horvath, WGCNA: an R package for weighted correlation network analysis, BMC Bioinf. 9 (1) (2008) 559, https://doi.org/10.1186/1471-2105-9-559.
[13] T. Wei, et al., Interferon-γ induces retinal pigment epithelial cell Ferroptosis by a JAK1-2/STAT1/SLC7A11 signaling pathway in Age-related Macular Degeneration, FEBS J. 289 (7) (2022) 1968–1983, https://doi.org/10.1111/febs.16272.
[14] Y. Sun, et al., Glutathione depletion induces ferroptosis, autophagy, and premature cell senescence in retinal pigment epithelial cells, Cell Death Dis. 9 (7) (2018) 753, https://doi.org/10.1038/s41419-018-0794-4.
[15] S. Rowan, A. Taylor, Gut microbiota modify risk for dietary glycemia-induced age-related macular degeneration, Gut Microb. (2018) 1–6, https://doi.org/10.1080/19490976.2018.1435247.
[16] K.L. Pennington, M.M. DeAngelis, Epidemiology of age-related macular degeneration (AMD): associations with cardiovascular disease phenotypes and lipid factors, Eye and Vision 3 (1) (2016) 34, https://doi.org/10.1186/s40662-016-0063-5.