

# Discovery and validation of molecular patterns and immune characteristics in the peripheral blood of ischemic stroke

Lin Cong<sup>1</sup>, Yijie He<sup>1</sup>, Yun Wu<sup>1</sup>, Ze Li<sup>1</sup>, Siwen Ding<sup>1</sup>, Weiwei Liang<sup>1</sup>, Xingjun Xiao<sup>1</sup>, Huixue Zhang<sup>1</sup>, Lihua Wang

Corresp. 1

<sup>1</sup> Department of Neurology, The Second Affiliated Hospital of Harbin Medical University, Harbin Medical University, Harbin, China

Corresponding Author: Lihua Wang

Email address: wang\_lihua0910@163.com

**Background:** Stroke is a disease with high morbidity, disability, and mortality. The immune factor plays a crucial role in the occurrence of ischemic stroke, but the exact mechanism of the immune is not clear. This study aims to identify possible immunological mechanisms by recognizing immune-related biomarkers and evaluating the infiltration pattern of immune cells.

**Methods:** We downloaded datasets of ischemic stroke patients from GEO, applied R language to discover differential expressed genes, and elucidated their biological functions by GO, KEGG analysis, and GSEA analysis. Then the hub genes were obtained using two machine learning algorithms (LASSO and SVM-REF) and the immune cell infiltration pattern was revealed by CIBERSORT. Gene-drug target networks and mRNA-miRNA-lncRNA regulatory networks were constructed using Cytoscape. Finally, we used RT-qPCR to validate the hub genes and applied logistic regression methods to build diagnostic models validated with ROC curves.

**Results:** We screened 188 differentially expressed genes whose functional analysis was enriched to multiple immune-related pathways. Then six hub genes (ANTXR2, BAZ2B, C5AR1, PDK4, PPIH and STK3) were identified using LASSO and SVM-REF. ANTXR2, BAZ2B, C5AR1, PDK4, and STK3 were positively correlated with neutrophils, T cells gamma delta, negatively correlated with T cell follicular helper cells and CD8, while PPIH showed the exact opposite trend. Immune infiltration indicated increased activity of monocytes, macrophages M0, neutrophils and mast cells and decreased infiltration of T cell follicular helper cells and CD8 in the IS group. The ceRNA network consists of 306 miRNA-mRNA interacting pairs and 285 miRNA-lncRNA interacting pairs. RT-qPCR results indicated that the expression levels of BAZ2B, C5AR1, PDK4 and STK3 were significantly increased in patients with ischemic stroke. Finally, we developed a diagnostic model based on these four genes. The AUC value of the model was verified to be 0.999 in the training set and 0.940 in the validation set.

**Conclusion:** Our research explored the immune-related gene expression modules and provided a specific basis for further study of immunomodulatory therapy of IS.

# Discovery and Validation of molecular patterns and Immune characteristics in the peripheral blood of ischemic stroke

Lin Cong<sup>1</sup>, Yijie He<sup>1</sup>, Yun Wu<sup>1</sup>, Ze Li<sup>1</sup>, Siwen Ding<sup>1</sup>, Weiwei Liang<sup>1</sup>, Xingjun Xiao<sup>1</sup>, Huixue Zhang<sup>1</sup>, Lihua Wang<sup>1</sup>

<sup>1</sup> Department of Neurology, The Second Affiliated Hospital of Harbin Medical University, Harbin, China

Corresponding Author:

Lihua Wang <sup>1</sup>

Xuefu Road, Nangang District, Harbin, Heilongjiang, China

Email address: wang\_lihua0910@163.com

## Abstract

**Background:** Stroke is a disease with high morbidity, disability, and mortality. The immune factor plays a crucial role in the occurrence of ischemic stroke, but the exact mechanism of the immune is not clear. This study aims to identify possible immunological mechanisms by recognizing immune-related biomarkers and evaluating the infiltration pattern of immune cells.

**Methods:** We downloaded datasets of ischemic stroke patients from GEO, applied R language to discover differential expressed genes, and elucidated their biological functions by GO, KEGG analysis, and GSEA analysis. Then the hub genes were obtained using two machine learning algorithms (LASSO and SVM-REF) and the immune cell infiltration pattern was revealed by CIBERSORT. Gene-drug target networks and mRNA-miRNA-lncRNA regulatory networks were constructed using Cytoscape. Finally, we used RT-qPCR to validate the hub genes and applied logistic regression methods to build diagnostic models validated with ROC curves.

**Results:** We screened 188 differentially expressed genes whose functional analysis was enriched to multiple immune-related pathways. Then six hub genes (ANTXR2, BAZ2B, C5AR1, PDK4, PPIH and STK3) were identified using LASSO and SVM-REF. ANTXR2, BAZ2B, C5AR1, PDK4, and STK3 were positively correlated with neutrophils, T cells gamma delta, negatively correlated with T cell follicular helper cells and CD8, while PPIH showed the exact opposite trend. Immune infiltration indicated increased activity of monocytes, macrophages M0, neutrophils and mast cells and decreased infiltration of T cell follicular helper cells and CD8 in the IS group. The ceRNA network consists of 306 miRNA-mRNA interacting pairs and 285 miRNA-lncRNA interacting pairs. RT-qPCR results indicated that the expression levels of BAZ2B, C5AR1, PDK4 and STK3 were significantly increased in patients with ischemic stroke.

Finally, we developed a diagnostic model based on these four genes. The AUC value of the model was verified to be 0.999 in the training set and 0.940 in the validation set.

**Conclusion:** Our research explored the immune-related gene expression modules and provided a specific basis for further study of immunomodulatory therapy of IS.

## Introduction

Stroke, of which more than 80% are ischemic (Hasan et al. 2018), is the second leading cause of death and the third leading cause of disability in adults (Campbell & Khatri 2020). It seriously affects patients' quality of life and causes a heavy burden to the family and society. Therefore, our study focuses on discovering novel targets for diagnosing and treating ischemic stroke (IS).

At present, the diagnosis of ischemic stroke (IS) mainly depends on clinical manifestations and neuroimaging, such as computed tomography (CT) (Harston et al. 2015) and magnetic resonance imaging (MRI) (Zameer et al. 2021). The most effective early treatment methods are thrombolysis and intravascular therapy (Cassella & Jagoda 2017; Hasan et al. 2018), both of which have time limitations to administer. Nonetheless, the clinical manifestations of the disease are variable, and the cost of administering CT and MRI scans is expensive and time-consuming. If the symptoms are not typical or the above examinations cannot be performed promptly, diagnosis and treatment may be delayed. At the same time, the above treatment methods carry the risk of bleeding. This has led us to search for new diagnostics and treatments.

Genetic factors are related to the occurrence and prognosis of stroke (Malik et al. 2018; Torres-Aguila et al. 2019). Possible stroke-related gene loci, such as ANK2, CDK6 and KCNK3, were found by whole genome sequencing (Malik et al. 2018). However, the specific mechanism of action is not clear. CYP450, COX-2, PTGIS, TBXAS1, P2RY1, and TGB3, or GPIIa are candidate genes that may be associated with stroke prognosis (Torres-Aguila et al. 2019). Presently, bioinformatics technology has been widely applied to various diseases, so we use bioinformatics technology to screen and identify key genes associated with IS, and build a classification diagnosis model based on these genes, to provide certain auxiliary methods for the diagnosis of ischemic stroke and potential therapeutic targets for the management of diseases.

Immune factors are also closely associated with ischemic stroke (Iadecola et al. 2020; Levard et al. 2021). In the acute phase, the immune response is activated immediately when the brain is in ischemia. Damaged or dead brain cells release damage-associated molecular patterns (DAMPs) (Wu et al. 2022). These substances quickly activate intracranial immune cells (microglia, etc.), release chemokines and cytokines, and trigger a series of inflammatory cascades (Anrather & Iadecola 2016; Levard et al. 2021). More importantly, immune cells in the peripheral blood can penetrate the central nervous system (CNS) through the damaged blood-brain barrier (BBB) (Jayaraj et al. 2019), further exacerbating the damage. At the same time, these DAMPs and

cytokines can enter the bloodstream and activate the systemic immune system and inflammatory response, which can lead to severe immunosuppression and even fatal infections (Iadecola et al. 2020). In the chronic phase, adaptive immune responses to the brain are mobilized due to antigen presentation, which may be the basis of neuropsychiatric sequelae, poststroke dementia/fatigue/depression, and the root cause of post-stroke morbidity (Iadecola et al. 2020). Epidemic-modulated therapy may be a good alternative to IS therapy. Several studies have shown that immune regulation can delay disease progression and improve neurological function and prognosis. These changes highlight the need to maintain a stable immune microenvironment in the central nervous system (Javidi & Magnus 2019; Jayaraj et al. 2019). TGF- $\beta$  1 has a significant neuroprotective effect (Taylor et al. 2017) and can reduce post-stroke infection (Cekanaviciute et al. 2014; Doll et al. 2014). Relevant targeted drugs are now being developed. Specific inhibitors of IL-1  $\beta$  can delay atherosclerosis by interfering with immune pathways associated with atherosclerotic plaque formation (Khambhati et al. 2018). Meanwhile, IL-1 receptor antagonists have been found to effectively reduce peripheral inflammation in patients with acute ischemic stroke and improve clinical outcomes in these patients (Smith et al. 2018). In conclusion, the analysis of immune infiltration analysis and the study of the relationship between target genes and immune cell distribution may help to elucidate the immune-related molecular mechanism of IS and provide the specific basis for immunoregulatory therapy. At the same time, the prediction of drug target gene action network may provide a new target for the treatment of IS.

Competing endogenous RNA (ceRNA) does not represent a specific type of RNA, but a regulatory mechanism. By combining with microRNA response elements (MREs) existing on mRNA, miRNAs inhibit mRNA translation or lead to its degradation, thus achieving the function of regulating gene expression after transcription (Salmena et al. 2011). Different mRNAs can compete to bind to the same miRNA and thus participate in the regulation of gene expression. With the deepening of transcriptome studies, it is found that MREs exist not only on mRNA, but also on lncRNA, circRNA and other types of RNA, which means that the same miRNA can bind to several and multiple types of RNA, and a competitive relationship is formed between RNA molecules that bind to the same miRNA. This forms the ceRNA network. If the regulation of ceRNA is abnormal, it may cause the occurrence of diseases, as shown in the research fields of tumors (Karreth et al. 2015), stroke (Li et al. 2020) and other complex diseases. It has been reported that miRNAs intervene the occurrence and prognosis of stroke by modulating immune and inflammatory factors. Huang L et al. found inhibition of microRNA-210 suppresses pro-inflammatory response and reduces acute brain injury of ischemic stroke in mice (Huang et al. 2018). Mark P. Boldin reported that miR-146a inhibits the production of inflammatory cytokines (Boldin et al. 2011). In addition, the occurrence, progression and prognosis of acute ischemic stroke are also related to abnormal lncRNA expression (Bai et al. 2014; Feng et al. 2019; Fu et al. 2021; Li et al. 2020; Wang et al. 2020). Therefore, they are potential candidates for stroke diagnosis. Meanwhile accumulating evidence demonstrates that lncRNAs regulate target gene

expression by acting as ceRNAs for miRNAs and participate in immunoinflammation in cerebral ischemia-reperfusion injury and various diseases. Knockdown metastasis-associated lung adenocarcinoma transcript 1 (MALAT1), a ceRNA for miR-181c-5p, attenuated inflammatory injury after cerebral ischemia (Cao et al. 2020). lncRNA SNHG14 promotes inflammatory response induced by cerebral ischemia /reperfusion injury by regulating miR-136-5p/ROCK1 (Zhong et al. 2019). The lncRNA small nucleolar RNA host gene 15 (SNHG15) was shown to regulate the expression of programmed death ligand 1 (PD-L1) by inhibiting miR-141, which in turn promoted the resistance of stomach cancer cells to the immune response (Dang et al. 2020). The lncRNA HIX003209 functioned as a ceRNA and enhanced inflammation by sponging miR-6089 via the Toll-like receptor 4 (TLR4)/nuclear factor (NF)- $\kappa$ B pathway in macrophages in rheumatoid arthritis (Yan et al. 2019). The construction of a target gene-miRNA-LncRNA-based CERNA network may provide a molecular basis for understanding stroke and help predict the possible immunological pathways of IS.

In this study, least absolute shrinkage and selection operator (LASSO) regression and support vector machine-recursive feature elimination (SVM-RFE) algorithm were applied to screen essential genes, and their relationship was subsequently analyzed with immune cells. The combined use of two algorithms will make the process more rigorous and standard. Then a ceRNA network of hub genes-miRNA-lncRNA was established. Finally, a logistic regression diagnostic model was built based on hub genes to provide an auxiliary method for the diagnosis of IS. The technical route refers to the following Figure 1.

## Materials & Methods

### Obtaining and processing profile datasets

We obtained and downloaded the ischemic stroke gene expression profile from Gene Expression Omnibus (GEO, <http://www.ncbi.nlm.nih.gov/geo>) (Table 1). The criteria for screening were that the data in the dataset contained both IS and control groups (CTL), and the number of cases in each group was no less than 20. Patients in the IS group were those diagnosed with acute ischemic stroke. In addition, the data were based on whole blood samples with full gene expression profiles. After screening, GSE16561 (Barr et al. 2010) and GSE58294 (Stamova et al. 2014) matched the criteria. GSE16561 was used as a training set. GSE58294 was for the validation of biomarker genes. We applied the Perl language script (<https://www.perl.org/>) to get the gene name corresponding to each probe from the platform annotation file; the names were then converted. In this process, we used the average of all probes' expressions if there were multiple probes corresponding to the same gene name.

### Differential Expression Analysis

We applied R package limma (Ritchie et al. 2015) (version 3.6.0) to perform differential expression analysis to identify DEGs between groups in the dataset GSE16561. FoldChange >

1.5 and adjusted  $P$ -value  $< 0.05$  were taken as significant criteria. DEGs were represented by heatmap, designed by the pheatmap package (Kolde 2012) (version 1.0.12), and volcano maps, created by the ggplot2 package (Wickham 2009) (version 3.3.5).

## Function enrichment analysis

We not only annotate the function of the differentially expressed genes themselves using Gene Ontology (GO) enrichment but also study the pathways through Kyoto Encyclopedia of Genes and Genomes (KEGG) analysis. GO Enrichment and KEGG pathways analysis results were obtained using clusterProfiler (Yu et al. 2012). GSEA software was used to analyze the potential molecular mechanisms and signaling pathways of Hub genes in IS patients. C2 (c2.cp. Kegg. v7.4.symbols. gmt) was selected as the reference gene set for analysis. The R package clusterProfiler also conducted the GSEA analysis of the IS and CTL groups, with an adjusted  $P$ -value  $< 0.05$ .

## Screening for hub genes

Two machine learning methods (LASSO and SVM-REF) were used to identify the hub genes of IS. These two methods were first applied to analyze breast tumors (Wang & Liu 2015), and later also to neurological diseases such as AD (Liu et al. 2021). LASSO regression is characterized by variable selection and complexity adjustment while fitting a generalized linear model. Thus, whether the target-dependent/response variable is continuous, binary or multivariate discrete, it can be modelled and predicted (Friedman et al. 2010). The LASSO analysis was undertaken using the glmnet package (Engelbrechtsen & Bohlin 2019) with the following parameters: response type set to binomial, alpha set to 1, and 10-fold cross-validation to adjust the optimal value of the parameter  $\lambda$ . SVM has shown many unique advantages in solving small samples, and nonlinear and high-dimensional pattern recognition, and can be extended to other machine learning problems such as function fitting. SVM-RFE, on the other hand, is used as a practical feature selection technique to find the best variables by removing the feature vectors generated by SVM (Wang & Liu 2015). The algorithm was constructed by the e1071 package, and  $\text{half.v} = 100$  and  $k = 5$  are used as parameter criteria. The hub genes were the intersecting genes obtained by the two machine learning algorithms.

## Evaluation of immune cell infiltration

Analysis of immune cell infiltration relied on CIBERSORT, which is a widespread tool to research the proportion of 22 immune cells in specimens. Furthermore, we analysed the Spearman correlation between immune cells and key genes. The corrplot software package (Friendly 2002) calculated the Pearson correlation coefficient between hub genes and each immune cell.

## Construction of the ceRNA Network

Firstly, the miRanda, MicroRNADB, and TargetScan databases were used to simultaneously predict the target miRNAs of hub genes and get the intersection. Obtained miRNA-mRNA interaction pair file. Secondly, the spongeScan database (<http://spongescan.rc.ufl.edu>), was used to predict the lncRNA based on the result file from the previous step. Finally, the ceRNA network was formed, and the feature files were jointly built based on the lncRNA-miRNA-mRNA files. Visualise ceRNA regulatory networks using Cytoscape (Shannon et al. 2003).

## Gene- Drug Target Network Analysis

The Drug-Gene Interaction Database (DGIDB, <http://www.dgidb.org/>), an online database that collects data on drug and gene interactions, is the source of material for building the network. The hub gene-drug interaction pair was compared with the DGIDB database to identify immunomodulatory drugs that may be related to ischemic stroke. Gene-Drug Target Network was constructed based on the above information.

## Samples collection

Our samples were all obtained from patients attending the Department of Neurology of the Second Affiliated Hospital of Harbin Medical University, approved by the Medical Ethics Committee of the Second Affiliated Hospital of Harbin Medical University (NO.KY2022-285) (Document S1), and all enrolled individuals signed an informed consent form (Document S2). 5 cases in the disease group and 5 cases in the healthy control group. Disease group (IS group) inclusion criteria: Patients were suffering from the 1st acute ischemic stroke confirmed by MRI, meeting the diagnostic criteria in the International Classification of Diseases (9th Revision). Exclusion criteria: Patients with a history of hematologic, type 1 diabetes, autoimmune, thyroid, neoplastic, renal or liver diseases were excluded. Fasting venous blood samples of 5ml were collected from both groups. Peripheral blood for the disease group was required to be collected within 24 hours of onset.

## Real-Time Quantitative PCR (RT-qPCR)

RT-qPCR was used to detect the expression of six hub genes in the whole peripheral blood of clinical samples. Total RNA was isolated from each sample using a Trizol reagent according to the manufacturer's instructions. Reverse transcription from RNA to cDNA was performed using HiScript Q RT SuperMix for qPCR (+gDNA wiper) (Vazyme Biotech, Nanjing). RT-qPCR was performed on cDNA samples using ChamQ SYBR Color qPCR Master Mix (2X) (Vazyme Biotech, Nanjing). Results were analyzed by the  $2^{-\Delta\Delta C_t}$  method and expressed as fold changes, with GAPDH selected as the internal reference gene. The PCR primers were designed by Majorbio (Majorbio, Shanghai) (Table S1). A *P*-value less than 0.05 was considered statistically significant.

## Statistical analysis

The Shapiro-Wilk test was used to assess whether the continuous data were normally distributed in the control and IS groups. Normally distributed data were counted using the independent samples t-test. Non-normally distributed data were evaluated using the Wilcoxon-Mann Whitney test. A  $P$ -value  $< 0.05$  was considered statistically significant.

## Development and Validation of a Diagnostic Model

On the one hand, boxplots were used to show the expression of hub genes in GSE58294 as a validation dataset. We considered the difference statistically significant at a  $P$ -value  $< 0.05$ . On the other hand, a logistic regression algorithm was used to build a diagnostic model for IS classification based on crucial genes. The Proc R package was applied to generate the ROC curve and calculate the area under the ROC curve (AUC) (Stamova et al. 2014) to verify the accuracy of the genes and the model in dataset GSE16561. In addition, we calculated the AUC values of the key genes and logistic regression models in dataset GSE58294 to validate the diagnostic models' classification performance. To avoid the occurrence of overfitting, we chose to reduce the complexity of the model and reduce the parameters to make the model simpler. When training the model, the dataset was partitioned in a 7:3 ratio for training the model and predicting the accuracy of the created model, respectively, to balance the accuracy. Finally, the visual analysis of the above process was completed using ggplot2.

## Results

### Differential Expression Analysis

The volcano plot (Fig. 2A) showed that by using R package limma, we obtained 188 differential expression genes (Table S2), of which 85 genes were down-regulated in expression and 103 genes were up-regulated in expression. The top 50 differentially expressed genes are shown in a heatmap (Fig. 2B), created by the pheatmap package.

### Function enrichment analysis

The GO analysis results (Table S3) were presented in a barplot (Figure 3A), a bubble plot (Figure 3B), and a circle plot (Figure 3C). GO-Biological Process (BP) shows that DEGs were primarily involved in immune-related biological processes, including immune response—regulating signaling pathway, positive regulation of cytokine production, and immune response—regulating cell surface receptor signaling pathway; GO- Molecular Function (MF) enrichment results showed DEGs were primarily associated with integrin binding oxidoreductase activity, acting on NAD(P)H quinone or similar compound as acceptor and immunoglobulin binding and GO-Cellular Component (CC) analysis was significantly enriched in platelet alpha granule, secretory granule lumen, etc. KEGG pathways of each module were listed in Figure 4 and Table S4, mainly enriched in MAPK signaling pathway, Cell adhesion molecules, Complement and coagulation, The Cascades, and so on. The GO and KEGG pathway analysis results suggested that the immune system played a crucial role in IS.



We further conducted a GSEA analysis to avoid missing the functions of some insignificant but biologically significant genes, biological characteristics, and regulatory networks. GSEA results (Figure 5; Table S5) showed that the pathway involved axon guidance, complement and coagulation cascades, Fc gamma r mediated phagocytosis, focal adhesion, and regulation of actin cytoskeleton in the IS group (Figure 5A). Complement and coagulation cascades, the identified pathway of IS, were consistent with the result of KEGG, and the other pathways are related to immunity or cellular metabolism. Genes in the treatment group (IS group) were enriched at the top, which means that this gene set was upregulated. Conversely, genes in the control group (Figure 5B) were enriched at the bottom with a down-regulation trend.

### Identification of the hub gene

To explore the biomarkers of IS, we performed feature screening through LASSO regression and the SVM-RFE algorithm. 12 genes were identified as the most associated with IS by LASSO regression (Fig. 6A). The SVM-RFE algorithm evaluated 34 characteristic genes in IS (Fig. 6B). Both machine algorithms were subjected to 10-fold cross-validation to ensure the accuracy of the results. Then, 6 differential expression genes (ANTXR2, BAZ2B, C5AR1, PDK4, PPIH, and STK3) (Fig. 6C) were identified as the hub genes by these two joint algorithms for subsequent research.

### Results of Immune Cell Infiltration Analysis

The immune cell infiltration analysis results (Table S6) were presented by barplot (Figure 7A) and heatmap (Figure 7B). Barplot showed a difference in the percentage of immune cells between the IS group and CTL. Compared with the CTL, Monocytes, macrophage M0, Neutrophils, and Mast cells activated infiltration increased, but T cells follicular helper, CD8 T cells infiltration decreased in the IS group. The relationship between immune cells and key genes is presented in Figure 8. ANTXR2, BAZ2B, C5AR1, PDK4, and STK3 were positively correlated with neutrophils and negatively correlated with T cells follicular helper and CD8 T cells; but the PPIH gene was positively associated with T cells follicular helper and CD8 T cells and negatively correlated with Mast cells activated and neutrophils; Only PPIH was a down-regulated expression gene, while the other five genes (ANTXR2, BAZ2B, C5AR1, PDK4, and STK3) were all up-regulated expression. Meanwhile, the correlation between individual hub genes and immune cells is shown by a Scatter plot (Figure S1-8). R-value indicates positive and negative correlation.  $P$ -value<0.05 is statistically significant. These evidence suggest that changes in the immune microenvironment of IS patients may be related to these 6 Hub genes.

### Construction of ceRNA Networks for Hub Genes

Many studies have confirmed that ceRNA regulatory networks play a role in the biology and pathophysiology of various diseases. We also constructed networks (Figure 9; Table S7) to determine whether 6 hub genes have similar regulatory relationships. 6 hub genes were separately entered into the database, yielding a total of 306 miRNA-mRNA interaction pairs, of

which there were 117 ANT XR2-miRNA interaction pairs 87 BAZ2B-miRNA interaction pairs, and 285 miRNA-lncRNA interaction pairs. We hypothesised that ANT XR2 and BAZ2B might be involved in the regulation of ceRNA. According to the value of a degree (Table S8), hsa-miR-766-3p, hsa-miR-149-3p, hsa-miR-1972, hsa-miR-186-5p, hsa-miR-1207-5p, MUC19 and LINC01002 may play a vital role in the ceRNA network.

### Establishment of Drug-Hub Genes Regulatory Networks

Drugs related to the 6 essential genes were screened out, and 3 genes, including BAZ2B, C5AR1, and STK3, had target drugs. A gene-drug network is shown in Figure 10 and Table S9. We retrieved a total of 58 drugs acting on the Hub gene. BAZ2B is predicted to have the most targets of drug action, a total of 30.25 drugs targeted STK3. Only one drug targeted C5AR1.

### Validation of Hub Genes and Diagnosis Model

The results of RT-qPCR showed that the expression of BAZ2B, C5AR1, PDK4, and STK3 was significantly higher in the ischemic stroke group (Figure 11 and Table S10). A diagnosis model for IS classification was established using a logistic regression algorithm based on the 4 key genes. The results show that the diagnostic model could distinguish patients from normal samples. The ROC curves for the HUB genes and models are shown in Fig 12A and Fig12B. The AUC values are as follows: BAZ2B (AUC=0.892), C5AR1 (AUC=0.966), PDK4 (AUC=0.891), STK3 (AUC=0.966), and Model (AUC=0.999). Finally, we verified the accuracy of hub genes (Figure 12C) and the model (Figure 12D) using another dataset, GSE58294. AUC values for genes and models were as follows: BAZ2B (AUC=0.936), C5AR1 (AUC=0.734), PDK4 (AUC=0.697), STK3 (AUC=0.751), and Model (AUC=0.940). The boxplot (Figure 13) shows the differential expression of hub genes in the CTL and IS groups in the validation dataset. Our established model performs well in distinguishing between IS and normal samples.

### Discussion

Ischemic stroke is a disease that threatens people's health worldwide, with high incidence, high disability rate, and high mortality. In addition to the current diagnosis and therapy, it is urgent to develop more convenient diagnosis and treatment programs to help solve this problem. Primary ischemic brain injury is combined by multiple mechanisms, including excitotoxicity, oxidative stress, apoptosis, and inflammation (Moskowitz et al. 2010). After that, a series of immune cascade reactions can further aggravate the damage to the brain tissue, such as the production of proinflammatory cytokines and the activation of destructive serine proteases (Anrather & Iadecola 2016; Xu et al. 2011). In our research, we explored possible biomarkers of IS, mechanisms of action, and potential drug targets in terms of inflammation and immunity.

In this study, 188 DEGs were observed through R package limma. By applying the joint LASSO and SVM algorithm, we finally obtained 6 genes (ANT XR2, BAZ2B, C5AR1, PDK4, PPIH, and STK3) most associated with IS. However only BAZ2B, C5AR1, PDK4, and STK3 were

upregulated expression in ischemic stroke patients as verified by RT-PCR. C5AR1 has been proven to play a crucial role in regulating inflammatory and neurocognitive functions in ischemic stroke, Alzheimer's disease, malaria, and neuropathic pain (Brandolini et al. 2019; McDonald et al. 2015; Moriconi et al. 2014). A marked increase in C5AR1 expression was observed in the MCAO- and OGD-induced models. Meanwhile, C5AR1 inhibitors have significant neuroprotective effects and remarkably inhibit neuro-inflammation and apoptosis in primary cortical neurons and MCAO-induced stroke models (Brandolini et al. 2019; Shi et al. 2017). TLR4 and C5AR1 promote apoptosis and inflammation by activating the cAMP/PKA/I- $\kappa$ B/NF- $\kappa$ B signaling pathway during brain ischemia-reperfusion (Kim & Jang 2017; Shi et al. 2017; Zaal et al. 2017b). Some study shows that C5AR1 can inhibit human monocyte-derived dendritic cells, potentially exacerbating inflammatory responses (Zaal et al. 2017a; Zaal et al. 2017b).

However, there are no studies on the correlation between BAZ2B, PDK4, and STK3 genes and ischemic stroke. The biological function of BAZ2B, remains unclear, besides the involvement in nucleosome remodelling (Oppikofer et al. 2017). However it includes at least four functional domains that could encode or bind multiple proteins or DNA to perform various roles. Due to this characteristic, we speculate that BAZ2B may be associated with many diseases, including ischemic stroke. But current studies are limited to suggesting that it is associated with neurodevelopment and that its functional loss and haploinsufficiency are one of the causes of autism and may act through transcriptional regulation (Krupp et al. 2017; Scott et al. 2020). Another study found that BAZ2B activates M2 macrophages to participate in the inflammatory response (Xia et al. 2021), indicating its research value in immunity. PDK4 is a mitochondrial matrix enzyme essential in cellular energy regulation, which regulates the pyruvate dehydrogenase complex in the central nervous system and has important implications for neuron-glia metabolic interactions (Jha et al. 2012). Atherosclerosis is the result of cholesterol and lipid deposition in the arterial wall, usually associated with calcification, and is the most common cause of ischemic stroke. Ma et al. found that PDK4 could promote vascular calcification by interfering with autophagic activity and metabolic reprogramming (Ma et al. 2020), contributing to the development of atherosclerosis. STK3 (MST2) is a component of the MAPK module and Hippo signaling pathway. In studies of cardiovascular disease, STK3 was found to mediate mir155 to initiate inflammation and redox stress, leading to vascular smooth muscle cell proliferation and remodeling. It regulates the ERK1/2 signaling pathway by competing with MAP2K of the MAPK pathway, which in turn triggers inflammation and oxidative stress after vascular injury (Thiriet 2018). And in ischemic stroke, the MAPK pathway also acts by regulating cytokines, inflammation, apoptosis and death (Qin et al. 2022; Sun & Nan 2016). STK3 is highly expressed in most cell types in the brain and may play a role in stroke by regulating the MAPK pathway. It has also been described as a substrate for CASP6 to intervene in neurodegeneration and apoptosis (Riechers et al. 2016; van Raam et al. 2013). Cho et al. found that STK3 increases phagocytosis of adipocytes, leading to obesity due to reduced catabolic function of adipocytes. In obese human patients, STK3 expression levels were elevated. STK3

inhibitors improved metabolic patterns in obese mouse models, suggesting that there may be a viable pathway to investigate and develop drugs targeting STK3 to treat obesity-related diseases including stroke (Cho et al. 2021). All of the above suggest that STK3 may be relevant to the onset of ischemic stroke, but the exact mechanism needs to be further investigated.

The critical pathways enriched by GO and KEGG analysis were related to immunity, including the immune response—regulating signaling pathway, immune response—regulating cell surface receptor signaling pathway, and MAPK signaling pathway. Notably, C5AR1 and STK3 were enriched in the positive regulatory pathway of the MAPK cascade. Previous studies have found that MAPK signaling pathway in inflammation and BBB dysfunction MAPK comprises three main effectors: ERK1/2, JNK, and p38 (Qin et al. 2022; Sun & Nan 2016). C5AR1 affects inflammation in bone by activating the MAPK pathway and regulating gene expression in pathways associated with insulin and transforming growth factor- $\beta$  (Modinger et al. 2018). STK3, as described above, mediates cardiovascular injury through the MAPK pathway and also promotes apoptosis by inducing the activation of JNK (Chen et al. 2018). These results are consistent with our bioinformatics analysis and suggest that the MAPK signaling pathway plays an important role in these key gene-mediated biological processes. In addition, the results of the GSEA enrichment analysis showed that C5AR1 was involved in complement, and coagulation cascades pathway which had been proven to correlate with IS (Berkowitz et al. 2021). The functions of PDK4 are all focused on the involvement of lipid metabolism in the process of atherosclerosis. These results are also in accordance with the role of C5AR1 and PDK4 found in some studies (Berkowitz et al. 2021; Ma et al. 2020).

Many stroke models have confirmed that neutrophils in the acute phase of ischemic stroke are among the earliest immune cells recruited into the brain tissue (Gokhan et al. 2013; Jickling et al. 2015; Kaito et al. 2013). Activated neutrophils may contribute to the development of ischemic stroke by stimulating the systemic inflammatory response and destroying the BBB. Van Duijn et al. found that CD8 T cells can exacerbate the inflammatory response by secreting various inflammatory cytokines, resulting in increased instability of atherosclerotic plaques. But CD8 T cells also have an anti-atherosclerotic effect, which is mediated by stimulation of inhibitory receptor production and cytolytic killing of antigen-presenting cells (van Duijn et al. 2018). The results of immunological infiltration analysis by CIBERSORT revealed that Monocytes, macrophage M0, Neutrophils, and Mast cells activated were increased, while T cells follicular helper and CD8 decreased infiltration in the IS group, which was consistent with the previous studies (Zheng et al. 2022). Our research also showed that neutrophils were significantly and positively correlated with gene ANTXR2, BAZ2B, C5AR1, PDK4, and STK3, negatively correlated with PPIH; CD8 was significantly and negatively correlated with gene ANTXR2, BAZ2B, C5AR1, PDK4, and STK3, positively correlated with PPIH. This evidence further suggests that neutrophils play a significant role, and hub genes are involved in regulating the immune microenvironment in IS patients.

The ceRNA network we conducted included nodes (6 mRNAs, 249 miRNAs, 218 lncRNAs) with 591 edges. In these nodes, the miRNAs with high degrees, hsa-miR-766-3p, hsa-miR-149-3p, and hsa-miR-186-5p, have been reported as essential molecules in cerebral ischemic/reperfusion injury (Cai et al. 2019; Hu et al. 2020; You et al. 2022). Of these miRNAs, hsa-miR-766-3p is known to be involved in immune responses or inflammation in various diseases. It contributes to anti-inflammatory responses through the indirect inhibition of NF- $\kappa$ B signaling in rheumatoid arthritis (RA) (Hayakawa et al. 2019). The miR-766-3p/NR3C2 axis is participating in the protection against cerebral ischemia and reperfusion (Cai et al. 2019). LINC00689, one of the highest degrees lncRNAs in the ceRNA network, is a key regulator of the toll-like receptor (TLR) signaling pathway controlling innate immunity (Liu et al. 2020). However, the mode of regulation between them and hub genes still needs further validation. Meanwhile, we constructed a Hub gene-based Gene-Drug regulatory network and predicted possible targeted drugs. The result reflected that BAZ2B had the most targets of drug action and the likely reason was that it had multiple functional areas. The network provides a theoretical basis for developing targeted immunotherapies for IS.

Finally, based on the four key genes mentioned above, we constructed an IS diagnostic model using logistic regression methods. We also verified the validity of the model in a publicly available dataset. Bioinformatics analysis combined with machine learning methods achieved good results.

However, there are some limitations to our study. First, although we analyzed clinical data and performed RT-qPCR validation, the sample size was relatively limited. Second, more in vivo and in vitro studies are needed to identify and validate the underlying mechanisms between these genes and stroke. These will be the most critical aspects of our future research. We hope to complete it in further studies to gain insight into the immunological pathogenesis of ischemic stroke and to provide additional options for clinical treatment.

## Conclusions

We explored 6 hub genes (ANTXR2, BAZ2B, C5AR1, PDK4, PPIH, and STK3) for ischemic stroke and revealed the immune cell infiltration pattern via bioinformatics analysis and machine learning algorithm. Secondly, we successfully constructed the ceRNA and Drug-Gene networks, which provide new ideas to explore the possible molecular mechanisms and develop targeted drugs of IS. Finally, BAZ2B, C5AR1, PDK4, and STK3 were verified to be the biological markers of ischemic stroke. Based on the 4 genes, we developed a diagnostic model and validated its effectiveness, providing additional tools for screening and diagnosing ischemic stroke.

## Acknowledgements

442 We want to thank the GEO database for the data support

# 443 References

- 444 Anrather J, and Iadecola C. 2016. Inflammation and Stroke: An Overview. *Neurotherapeutics*  
445 13:661-670. 10.1007/s13311-016-0483-x
- 446 Bai Y, Nie S, Jiang G, Zhou Y, Zhou M, Zhao Y, Li S, Wang F, Lv Q, Huang Y, Yang Q, Li Q, Li  
447 Y, Xia Y, Liu Y, Liu J, Qian J, Li B, Wu G, Wu Y, Wang B, Cheng X, Yang Y, Ke T, Li H,  
448 Ren X, Ma X, Liao Y, Xu C, Tu X, and Wang QK. 2014. Regulation of CARD8  
449 expression by ANRIL and association of CARD8 single nucleotide polymorphism  
450 rs2043211 (p.C10X) with ischemic stroke. *Stroke* 45:383-388.  
451 10.1161/STROKEAHA.113.003393
- 452 Barr TL, Conley Y, Ding J, Dillman A, Warach S, Singleton A, and Matarin M. 2010. Genomic  
453 biomarkers and cellular pathways of ischemic stroke by RNA gene expression profiling.  
454 *Neurology* 75:1009-1014. 10.1212/WNL.0b013e3181f2b37f
- 455 Berkowitz S, Chapman J, Dori A, Gofrit SG, Maggio N, and Shavit-Stein E. 2021. Complement  
456 and Coagulation System Crosstalk in Synaptic and Neural Conduction in the Central and  
457 Peripheral Nervous Systems. *Biomedicines* 9. 10.3390/biomedicines9121950
- 458 Boldin MP, Taganov KD, Rao DS, Yang L, Zhao JL, Kalwani M, Garcia-Flores Y, Luong M,  
459 Devrekanli A, Xu J, Sun G, Tay J, Linsley PS, and Baltimore D. 2011. miR-146a is a  
460 significant brake on autoimmunity, myeloproliferation, and cancer in mice. *J Exp Med*  
461 208:1189-1201. 10.1084/jem.20101823
- 462 Brandolini L, Grannonico M, Bianchini G, Colanardi A, Sebastiani P, Paladini A, Piroli A,  
463 Allegretti M, Varrassi G, and Di Loreto S. 2019. The Novel C5aR Antagonist DF3016A  
464 Protects Neurons Against Ischemic Neuroinflammatory Injury. *Neurotox Res* 36:163-  
465 174. 10.1007/s12640-019-00026-w
- 466 Cai J, Shangguan S, Li G, Cai Y, Chen Y, Ma G, Miao Z, Liu L, and Deng Y. 2019. Knockdown  
467 of lncRNA Gm11974 protect against cerebral ischemic reperfusion through miR-766-  
468 3p/NR3C2 axis. *Artif Cells Nanomed Biotechnol* 47:3847-3853.  
469 10.1080/21691401.2019.1666859
- 470 Campbell BCV, and Khatrri P. 2020. Stroke. *Lancet* 396:129-142. 10.1016/S0140-  
471 6736(20)31179-X
- 472 Cao DW, Liu MM, Duan R, Tao YF, Zhou JS, Fang WR, Zhu JR, Niu L, and Sun JG. 2020. The  
473 lncRNA Malat1 functions as a ceRNA to contribute to berberine-mediated inhibition of  
474 HMGB1 by sponging miR-181c-5p in poststroke inflammation. *Acta Pharmacol Sin*  
475 41:22-33. 10.1038/s41401-019-0284-y
- 476 Cassella CR, and Jagoda A. 2017. Ischemic Stroke: Advances in Diagnosis and Management.  
477 *Emerg Med Clin North Am* 35:911-930. 10.1016/j.emc.2017.07.007
- 478 Cekanaviciute E, Fathali N, Doyle KP, Williams AM, Han J, and Buckwalter MS. 2014.  
479 Astrocytic transforming growth factor-beta signaling reduces subacute  
480 neuroinflammation after stroke in mice. *Glia* 62:1227-1240. 10.1002/glia.22675
- 481 Chen S, Fang Y, Xu S, Reis C, and Zhang J. 2018. Mammalian Sterile20-like Kinases:  
482 Signalings and Roles in Central Nervous System. *Aging Dis* 9:537-552.  
483 10.14336/AD.2017.0702
- 484 Cho YK, Son Y, Saha A, Kim D, Choi C, Kim M, Park JH, Im H, Han J, Kim K, Jung YS, Yun J,  
485 Bae EJ, Seong JK, Lee MO, Lee S, Granneman JG, and Lee YH. 2021. STK3/STK4  
486 signalling in adipocytes regulates mitophagy and energy expenditure. *Nat Metab* 3:428-  
487 441. 10.1038/s42255-021-00362-2
- 488 Dang S, Malik A, Chen J, Qu J, Yin K, Cui L, and Gu M. 2020. LncRNA SNHG15 Contributes to  
489 Immuno-Escape of Gastric Cancer Through Targeting miR141/PD-L1. *Onco Targets*  
490 *Ther* 13:8547-8556. 10.2147/OTT.S251625

- Doll DN, Barr TL, and Simpkins JW. 2014. Cytokines: their role in stroke and potential use as biomarkers and therapeutic targets. *Aging Dis* 5:294-306. 10.14336/AD.2014.0500294
- Engebretsen S, and Bohlin J. 2019. Statistical predictions with glmnet. *Clin Epigenetics* 11:123. 10.1186/s13148-019-0730-1
- Feng L, Guo J, and Ai F. 2019. Circulating long noncoding RNA ANRIL downregulation correlates with increased risk, higher disease severity and elevated pro-inflammatory cytokines in patients with acute ischemic stroke. *J Clin Lab Anal* 33:e22629. 10.1002/jcla.22629
- Friedman J, Hastie T, and Tibshirani R. 2010. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw* 33:1-22.
- Friendly M. 2002. Corrgams. *The American Statistician* 56:316-324. 10.1198/000313002533
- Fu J, Yu Q, Xiao J, and Li S. 2021. Long noncoding RNA as a biomarker for the prognosis of ischemic stroke: A protocol for meta-analysis and bioinformatics analysis. *Medicine (Baltimore)* 100:e25596. 10.1097/MD.00000000000025596
- Gokhan S, Ozhasenekler A, Mansur Durgun H, Akil E, Ustundag M, and Orak M. 2013. Neutrophil lymphocyte ratios in stroke subtypes and transient ischemic attack. *Eur Rev Med Pharmacol Sci* 17:653-657.
- Harston GW, Rane N, Shaya G, Thandeswaran S, Cellerini M, Sheerin F, and Kennedy J. 2015. Imaging biomarkers in acute ischemic stroke trials: a systematic review. *AJNR Am J Neuroradiol* 36:839-843. 10.3174/ajnr.A4208
- Hasan TF, Rabinstein AA, Middlebrooks EH, Haranhalli N, Silliman SL, Meschia JF, and Tawk RG. 2018. Diagnosis and Management of Acute Ischemic Stroke. *Mayo Clin Proc* 93:523-538. 10.1016/j.mayocp.2018.02.013
- Hayakawa K, Kawasaki M, Hirai T, Yoshida Y, Tsushima H, Fujishiro M, Ikeda K, Morimoto S, Takamori K, and Sekigawa I. 2019. MicroRNA-766-3p Contributes to Anti-Inflammatory Responses through the Indirect Inhibition of NF-kappaB Signaling. *Int J Mol Sci* 20. 10.3390/ijms20040809
- Hu X, Xiang Z, Zhang W, Yu Z, Xin X, Zhang R, Deng Y, and Yuan Q. 2020. Protective effect of DLX6-AS1 silencing against cerebral ischemia/reperfusion induced impairments. *Aging (Albany NY)* 12:23096-23113. 10.18632/aging.104070
- Huang L, Ma Q, Li Y, Li B, and Zhang L. 2018. Inhibition of microRNA-210 suppresses pro-inflammatory response and reduces acute brain injury of ischemic stroke in mice. *Exp Neurol* 300:41-50. 10.1016/j.expneurol.2017.10.024
- Iadecola C, Buckwalter MS, and Anrather J. 2020. Immune responses to stroke: mechanisms, modulation, and therapeutic potential. *J Clin Invest* 130:2777-2788. 10.1172/JCI135530
- Javidi E, and Magnus T. 2019. Autoimmunity After Ischemic Stroke and Brain Injury. *Front Immunol* 10:686. 10.3389/fimmu.2019.00686
- Jayaraj RL, Azimullah S, Beiram R, Jalal FY, and Rosenberg GA. 2019. Neuroinflammation: friend and foe for ischemic stroke. *J Neuroinflammation* 16:142. 10.1186/s12974-019-1516-2
- Jha MK, Jeon S, and Suk K. 2012. Pyruvate Dehydrogenase Kinases in the Nervous System: Their Principal Functions in Neuronal-glial Metabolic Interaction and Neuro-metabolic Disorders. *Curr Neuroparmacol* 10:393-403. 10.2174/157015912804143586
- Jickling GC, Liu D, Ander BP, Stamova B, Zhan X, and Sharp FR. 2015. Targeting neutrophils in ischemic stroke: translational insights from experimental studies. *J Cereb Blood Flow Metab* 35:888-901. 10.1038/jcbfm.2015.45
- Kaito M, Araya S, Gondo Y, Fujita M, Minato N, Nakanishi M, and Matsui M. 2013. Relevance of distinct monocyte subsets to clinical course of ischemic stroke patients. *PLoS One* 8:e69409. 10.1371/journal.pone.0069409
- Karreth FA, Reschke M, Ruocco A, Ng C, Chapuy B, Leopold V, Sjoberg M, Keane TM, Verma A, Ala U, Tay Y, Wu D, Seitzer N, Velasco-Herrera Mdel C, Bothmer A, Fung J,

- Langellotto F, Rodig SJ, Elemento O, Shipp MA, Adams DJ, Chiarle R, and Pandolfi PP. 2015. The BRAF pseudogene functions as a competitive endogenous RNA and induces lymphoma in vivo. *Cell* 161:319-332. 10.1016/j.cell.2015.02.043
- Khambhati J, Engels M, Allard-Ratick M, Sandesara PB, Quyyumi AA, and Sperling L. 2018. Immunotherapy for the prevention of atherosclerotic cardiovascular disease: Promise and possibilities. *Atherosclerosis* 276:1-9. 10.1016/j.atherosclerosis.2018.07.007
- Kim SH, and Jang YS. 2017. Yersinia enterocolitica Exploits Signal Crosstalk between Complement 5a Receptor and Toll-like Receptor 1/2 and 4 to Avoid the Bacterial Clearance in M cells. *Immune Netw* 17:228-236. 10.4110/in.2017.17.4.228
- Kolde R. 2012. Pheatmap: pretty heatmaps. *R package version* 1:726.
- Krupp DR, Barnard RA, Duffourd Y, Evans SA, Mulqueen RM, Bernier R, Riviere JB, Fombonne E, and O'Roak BJ. 2017. Exonic Mosaic Mutations Contribute Risk for Autism Spectrum Disorder. *Am J Hum Genet* 101:369-390. 10.1016/j.ajhg.2017.07.016
- Levard D, Buendia I, Lanquetin A, Glavan M, Vivien D, and Rubio M. 2021. Filling the gaps on stroke research: Focus on inflammation and immunity. *Brain Behav Immun* 91:649-667. 10.1016/j.bbi.2020.09.025
- Li P, Duan S, and Fu A. 2020. Long noncoding RNA NEAT1 correlates with higher disease risk, worse disease condition, decreased miR-124 and miR-125a and predicts poor recurrence-free survival of acute ischemic stroke. *J Clin Lab Anal* 34:e23056. 10.1002/jcla.23056
- Liu J, Wang Y, Chu Y, Xu R, Zhang D, and Wang X. 2020. Identification of a TLR-Induced Four-lncRNA Signature as a Novel Prognostic Biomarker in Esophageal Carcinoma. *Front Cell Dev Biol* 8:649. 10.3389/fcell.2020.00649
- Liu Z, Li H, and Pan S. 2021. Discovery and Validation of Key Biomarkers Based on Immune Infiltrates in Alzheimer's Disease. *Front Genet* 12:658323. 10.3389/fgene.2021.658323
- Ma WQ, Sun XJ, Zhu Y, and Liu NF. 2020. PDK4 promotes vascular calcification by interfering with autophagic activity and metabolic reprogramming. *Cell Death Dis* 11:991. 10.1038/s41419-020-03162-w
- Malik R, Chauhan G, Traylor M, Sargurupremraj M, Okada Y, Mishra A, Rutten-Jacobs L, Giese AK, van der Laan SW, Gretarsdottir S, Anderson CD, Chong M, Adams HHH, Ago T, Almgren P, Amouyel P, Ay H, Bartz TM, Benavente OR, Bevan S, Boncoraglio GB, Brown RD, Jr., Butterworth AS, Carrera C, Carty CL, Chasman DI, Chen WM, Cole JW, Correa A, Cotlarciuc I, Cruchaga C, Danesh J, de Bakker PIW, DeStefano AL, den Hoed M, Duan Q, Engelter ST, Falcone GJ, Gottesman RF, Grewal RP, Gudnason V, Gustafsson S, Haessler J, Harris TB, Hassan A, Havulinna AS, Heckbert SR, Holliday EG, Howard G, Hsu FC, Hyacinth HI, Ikram MA, Ingelsson E, Irvin MR, Jian X, Jimenez-Conde J, Johnson JA, Jukema JW, Kanai M, Keene KL, Kissela BM, Kleindorfer DO, Kooperberg C, Kubo M, Lange LA, Langefeld CD, Langenberg C, Launer LJ, Lee JM, Lemmens R, Leys D, Lewis CM, Lin WY, Lindgren AG, Lorentzen E, Magnusson PK, Maguire J, Manichaikul A, McArdle PF, Meschia JF, Mitchell BD, Mosley TH, Nalls MA, Ninomiya T, O'Donnell MJ, Psaty BM, Pulit SL, Rannikmae K, Reiner AP, Rexrode KM, Rice K, Rich SS, Ridker PM, Rost NS, Rothwell PM, Rotter JI, Rundek T, Sacco RL, Sakaue S, Sale MM, Salomaa V, Sapkota BR, Schmidt R, Schmidt CO, Schminke U, Sharma P, Slowik A, Sudlow CLM, Tanislav C, Tatlisumak T, Taylor KD, Thijs VNS, Thorleifsson G, Thorsteinsdottir U, Tiedt S, Trompet S, Tzourio C, van Duijn CM, Walters M, Wareham NJ, Wassertheil-Smoller S, Wilson JG, Wiggins KL, Yang Q, Yusuf S, Consortium AF, Cohorts for H, Aging Research in Genomic Epidemiology C, International Genomics of Blood Pressure C, Consortium I, Starnet, Bis JC, Pastinen T, Ruusalepp A, Schadt EE, Koplev S, Bjorkegren JLM, Codoni V, Civelek M, Smith NL, Tregouet DA, Christophersen IE, Roselli C, Lubitz SA, Ellinor PT, Tai ES, Kooner JS, Kato N, He J, van der Harst P, Elliott P, Chambers JC, Takeuchi F, Johnson AD,



BioBank Japan Cooperative Hospital G, Consortium C, Consortium E-C, Consortium EP-I, International Stroke Genetics C, Consortium M, Neurology Working Group of the CC, Network NSG, Study UKYLD, Consortium M, Sanghera DK, Melander O, Jern C, Strbian D, Fernandez-Cadenas I, Longstreth WT, Jr., Rolfs A, Hata J, Woo D, Rosand J, Pare G, Hopewell JC, Saleheen D, Stefansson K, Worrall BB, Kittner SJ, Seshadri S, Fornage M, Markus HS, Howson JMM, Kamatani Y, Debette S, and Dichgans M. 2018. Multiancestry genome-wide association study of 520,000 subjects identifies 32 loci associated with stroke and stroke subtypes. *Nat Genet* 50:524-537. 10.1038/s41588-018-0058-3

McDonald CR, Cahill LS, Ho KT, Yang J, Kim H, Silver KL, Ward PA, Mount HT, Liles WC, Sled JG, and Kain KC. 2015. Experimental Malaria in Pregnancy Induces Neurocognitive Injury in Uninfected Offspring via a C5a-C5a Receptor Dependent Pathway. *PLoS Pathog* 11:e1005140. 10.1371/journal.ppat.1005140

Modinger Y, Rapp A, Pazmandi J, Vikman A, Holzmann K, Haffner-Luntzer M, Huber-Lang M, and Ignatius A. 2018. C5aR1 interacts with TLR2 in osteoblasts and stimulates the osteoclast-inducing chemokine CXCL10. *J Cell Mol Med* 22:6002-6014. 10.1111/jcmm.13873

Moriconi A, Cunha TM, Souza GR, Lopes AH, Cunha FQ, Carneiro VL, Pinto LG, Brandolini L, Aramini A, Bizzarri C, Bianchini G, Beccari AR, Fanton M, Bruno A, Costantino G, Bertini R, Galliera E, Locati M, Ferreira SH, Teixeira MM, and Allegretti M. 2014. Targeting the minor pocket of C5aR for the rational design of an oral allosteric inhibitor for inflammatory and neuropathic pain relief. *Proc Natl Acad Sci U S A* 111:16937-16942. 10.1073/pnas.1417365111

Moskowitz MA, Lo EH, and Iadecola C. 2010. The science of stroke: mechanisms in search of treatments. *Neuron* 67:181-198. 10.1016/j.neuron.2010.07.002

Oppikofer M, Bai T, Gan Y, Haley B, Liu P, Sandoval W, Ciferri C, and Cochran AG. 2017. Expansion of the ISWI chromatin remodeler family with new active complexes. *EMBO Rep* 18:1697-1706. 10.15252/embr.201744011

Qin C, Yang S, Chu YH, Zhang H, Pang XW, Chen L, Zhou LQ, Chen M, Tian DS, and Wang W. 2022. Signaling pathways involved in ischemic stroke: molecular mechanisms and therapeutic interventions. *Signal Transduct Target Ther* 7:215. 10.1038/s41392-022-01064-1

Riechers SP, Butland S, Deng Y, Skotte N, Ehrnhoefer DE, Russ J, Laine J, Laroche M, Pouladi MA, Wanker EE, Hayden MR, and Graham RK. 2016. Interactome network analysis identifies multiple caspase-6 interactors involved in the pathogenesis of HD. *Hum Mol Genet* 25:1600-1618. 10.1093/hmg/ddw036

Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, and Smyth GK. 2015. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 43:e47. 10.1093/nar/gkv007

Salmena L, Poliseno L, Tay Y, Kats L, and Pandolfi PP. 2011. A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? *Cell* 146:353-358. 10.1016/j.cell.2011.07.014

Scott TM, Guo H, Eichler EE, Rosenfeld JA, Pang K, Liu Z, Lalani S, Bi W, Yang Y, Bacino CA, Streff H, Lewis AM, Koenig MK, Thiffault I, Bellomo A, Everman DB, Jones JR, Stevenson RE, Bernier R, Gilissen C, Pfundt R, Hiatt SM, Cooper GM, Holder JL, and Scott DA. 2020. BAZ2B haploinsufficiency as a cause of developmental delay, intellectual disability, and autism spectrum disorder. *Hum Mutat* 41:921-925. 10.1002/humu.23992

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, and Ideker T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 13:2498-2504. 10.1101/gr.1239303

- Shi YW, Zhang XC, Chen C, Tang M, Wang ZW, Liang XM, Ding F, and Wang CP. 2017. Schisantherin A attenuates ischemia/reperfusion-induced neuronal injury in rats via regulation of TLR4 and C5aR1 signaling pathways. *Brain Behav Immun* 66:244-256. 10.1016/j.bbi.2017.07.004
- Smith CJ, Hulme S, Vail A, Heal C, Parry-Jones AR, Scarth S, Hopkins K, Hoadley M, Allan SM, Rothwell NJ, Hopkins SJ, and Tyrrell PJ. 2018. SCIL-STROKE (Subcutaneous Interleukin-1 Receptor Antagonist in Ischemic Stroke): A Randomized Controlled Phase 2 Trial. *Stroke* 49:1210-1216. 10.1161/STROKEAHA.118.020750
- Stamova B, Jickling GC, Ander BP, Zhan X, Liu D, Turner R, Ho C, Khoury JC, Bushnell C, Pancioli A, Jauch EC, Broderick JP, and Sharp FR. 2014. Gene expression in peripheral immune cells following cardioembolic stroke is sexually dimorphic. *PLoS One* 9:e102550. 10.1371/journal.pone.0102550
- Sun J, and Nan G. 2016. The Mitogen-Activated Protein Kinase (MAPK) Signaling Pathway as a Discovery Target in Stroke. *J Mol Neurosci* 59:90-98. 10.1007/s12031-016-0717-8
- Taylor RA, Chang CF, Goods BA, Hammond MD, Mac Grory B, Ai Y, Steinschneider AF, Renfro SC, Askenase MH, McCullough LD, Kasner SE, Mullen MT, Hafler DA, Love JC, and Sansing LH. 2017. TGF-beta1 modulates microglial phenotype and promotes recovery after intracerebral hemorrhage. *J Clin Invest* 127:280-292. 10.1172/JCI88647
- Thiriet M. 2018. Cardiovascular Risk Factors and Markers. *Vasculopathies*, 91-198.
- Torres-Aguila NP, Carrera C, Muino E, Cullell N, Carcel-Marquez J, Gallego-Fabrega C, Gonzalez-Sanchez J, Bustamante A, Delgado P, Ibanez L, Heitsch L, Krupinski J, Montaner J, Marti-Fabregas J, Cruchaga C, Lee JM, Fernandez-Cadenas I, and Acute Endophenotypes Group of the International Stroke Genetics C. 2019. Clinical Variables and Genetic Risk Factors Associated with the Acute Outcome of Ischemic Stroke: A Systematic Review. *J Stroke* 21:276-289. 10.5853/jos.2019.01522
- van Duijn J, Kuiper J, and Slutter B. 2018. The many faces of CD8+ T cells in atherosclerosis. *Curr Opin Lipidol* 29:411-416. 10.1097/MOL.0000000000000541
- van Raam BJ, Ehrnhoefer DE, Hayden MR, and Salvesen GS. 2013. Intrinsic cleavage of receptor-interacting protein kinase-1 by caspase-6. *Cell Death Differ* 20:86-96. 10.1038/cdd.2012.98
- Wang M, Chen W, Geng Y, Xu C, Tao X, and Zhang Y. 2020. Long Non-Coding RNA MEG3 Promotes Apoptosis of Vascular Cells and is Associated with Poor Prognosis in Ischemic Stroke. *J Atheroscler Thromb* 27:718-726. 10.5551/jat.50674
- Wang Q, and Liu X. 2015. Screening of feature genes in distinguishing different types of breast cancer using support vector machine. *Onco Targets Ther* 8:2311-2317. 10.2147/OTT.S85271
- Wickham H. 2009. ggplot2: Elegant graphics for data analysis (Use R) Springer-Verlag. New York.
- Wu F, Liu Z, Zhou L, Ye D, Zhu Y, Huang K, Weng Y, Xiong X, Zhan R, and Shen J. 2022. Systemic immune responses after ischemic stroke: From the center to the periphery. *Front Immunol* 13:911661. 10.3389/fimmu.2022.911661
- Xia L, Wang X, Liu L, Fu J, Xiao W, Liang Q, Han X, Huang S, Sun L, Gao Y, Zhang C, Yang L, Wang L, Qian L, and Zhou Y. 2021. Inc-BAZ2B promotes M2 macrophage activation and inflammation in children with asthma through stabilizing BAZ2B pre-mRNA. *J Allergy Clin Immunol* 147:921-932 e929. 10.1016/j.jaci.2020.06.034
- Xu J, Zhang X, Monestier M, Esmon NL, and Esmon CT. 2011. Extracellular histones are mediators of death through TLR2 and TLR4 in mouse fatal liver injury. *J Immunol* 187:2626-2631. 10.4049/jimmunol.1003930
- Yan S, Wang P, Wang J, Yang J, Lu H, Jin C, Cheng M, and Xu D. 2019. Long Non-coding RNA HIX003209 Promotes Inflammation by Sponging miR-6089 via TLR4/NF-kappaB

Signaling Pathway in Rheumatoid Arthritis. *Front Immunol* 10:2218. 10.3389/fimmu.2019.02218

You J, Qian F, Huang Y, Guo Y, Lv Y, Yang Y, Lu X, Guo T, Wang J, and Gu B. 2022. lncRNA WT1-AS attenuates hypoxia/ischemia-induced neuronal injury during cerebral ischemic stroke via miR-186-5p/XIAP axis. *Open Med (Wars)* 17:1338-1349. 10.1515/med-2022-0528

Yu G, Wang LG, Han Y, and He QY. 2012. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* 16:284-287. 10.1089/omi.2011.0118

Zaal A, Dieker M, Oudenampsen M, Turksma AW, Lissenberg-Thunnissen SN, Wouters D, van Ham SM, and Ten Brinke A. 2017a. Anaphylatoxin C5a Regulates 6-Sulfo-LacNAc Dendritic Cell Function in Human through Crosstalk with Toll-Like Receptor-Induced CREB Signaling. *Front Immunol* 8:818. 10.3389/fimmu.2017.00818

Zaal A, Nota B, Moore KS, Dieker M, van Ham SM, and Ten Brinke A. 2017b. TLR4 and C5aR crosstalk in dendritic cells induces a core regulatory network of RSK2, PI3Kbeta, SGK1, and FOXO transcription factors. *J Leukoc Biol* 102:1035-1054. 10.1189/jlb.2MA0217-058R

Zameer S, Siddiqui AS, and Riaz R. 2021. Multimodality Imaging in Acute Ischemic Stroke. *Curr Med Imaging* 17:567-577. 10.2174/1573405616666201130094948

Zheng PF, Chen LZ, Liu P, Pan HW, Fan WJ, and Liu ZY. 2022. Identification of immune-related key genes in the peripheral blood of ischaemic stroke patients using a weighted gene coexpression network analysis and machine learning. *J Transl Med* 20:361. 10.1186/s12967-022-03562-w

Zhong Y, Yu C, and Qin W. 2019. lncRNA SNHG14 promotes inflammatory response induced by cerebral ischemia/reperfusion injury through regulating miR-136-5p /ROCK1. *Cancer Gene Ther* 26:234-247. 10.1038/s41417-018-0067-5

**Table 1** (on next page)

*Basic information on gene expression profiling*

**Table 1 Basic information on gene expression profiling**

1

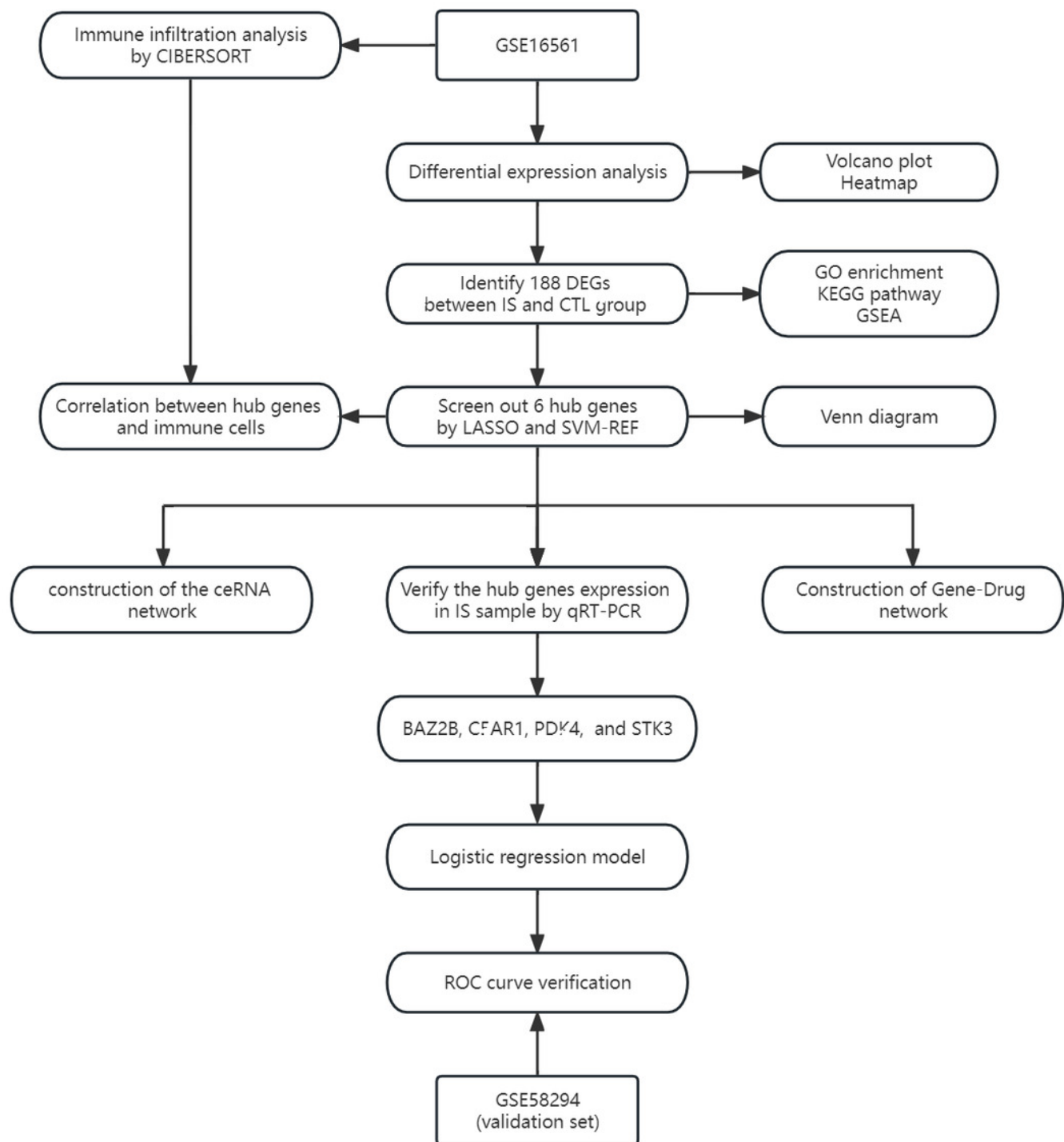
Location	Dataset ID	Platform	Type	Number
Whole blood samples	GSE16561	GPL6883	Microarray	39IS VS 24 CTL
Whole blood samples	GSE58294	GPL570	Microarray	69IS VS 23CTL

2

# Figure 1

*The technical route*

*Abbreviations: IS, ischemic stroke; CTL, Healthy Control*



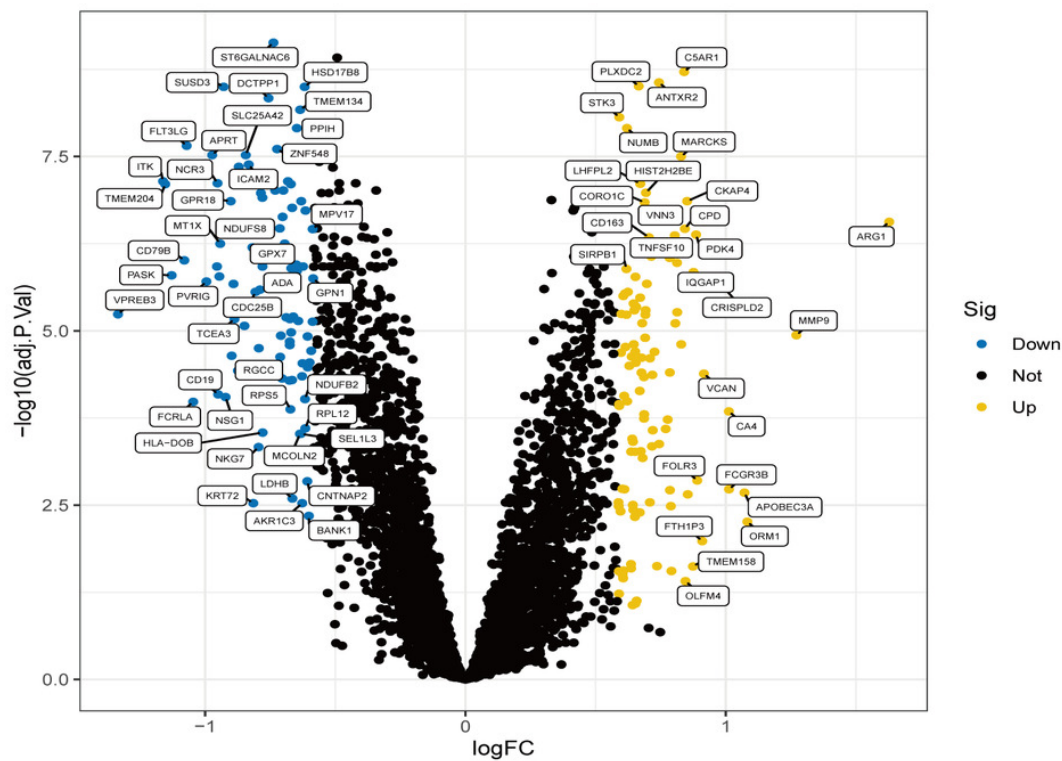
# Figure 2

Differential Expression Analysis

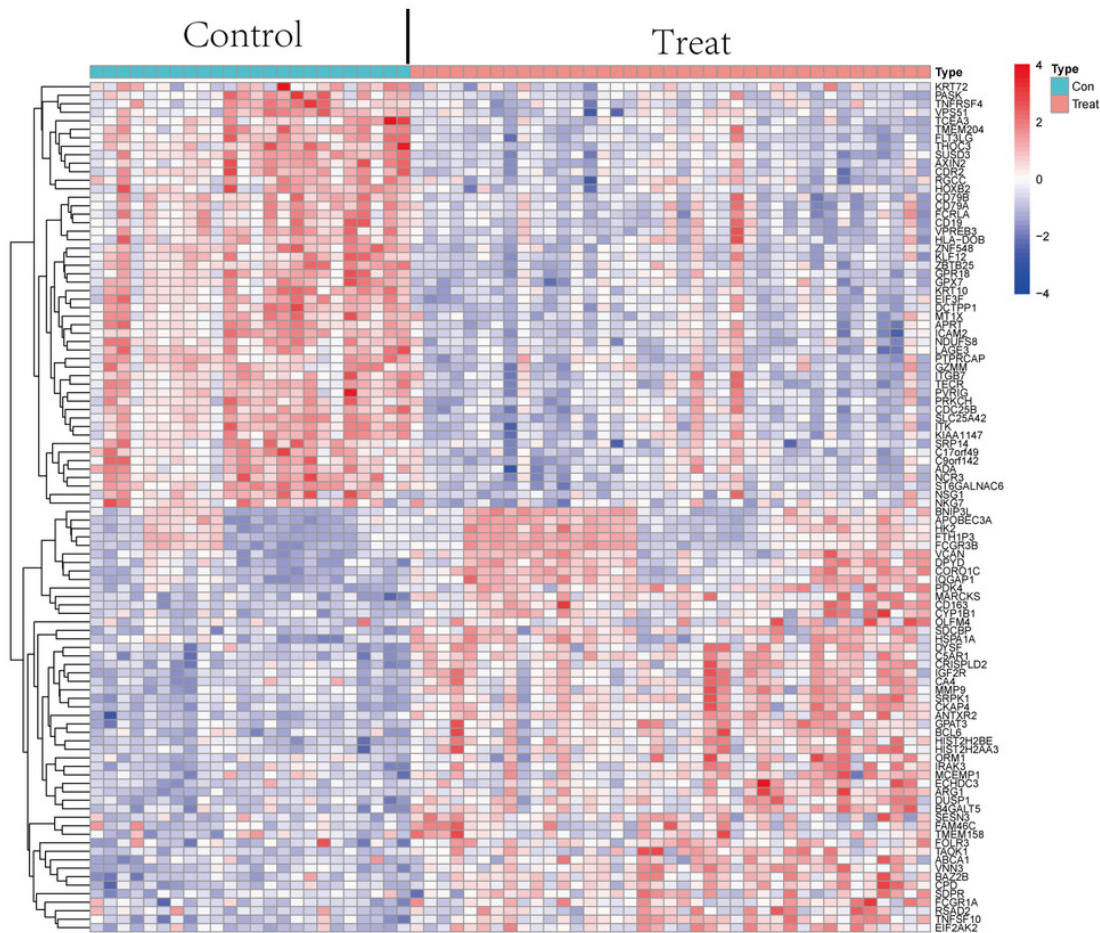
(A) The differential expression analysis results are shown in the volcano plot. The x-axis represents  $\log_2$  (fold change), and the y-axis represents  $-\log_{10}$  (adjusted p. value). Green dots represent downregulated genes, red dots represent upregulated genes, and black dots represent genes with no evident differential expression. (B) Heatmap of DEGs. Each column in the graph represents a sample, each row represents a gene, and the expression status of the gene is indicated from high to low in orange to blue, respectively, and at the top of the heatmap, red/blue represents the IS group/control group, respectively. Abbreviations: IS (ischemic stroke), Control (Healthy Control), DEGs (differential expression genes)



A



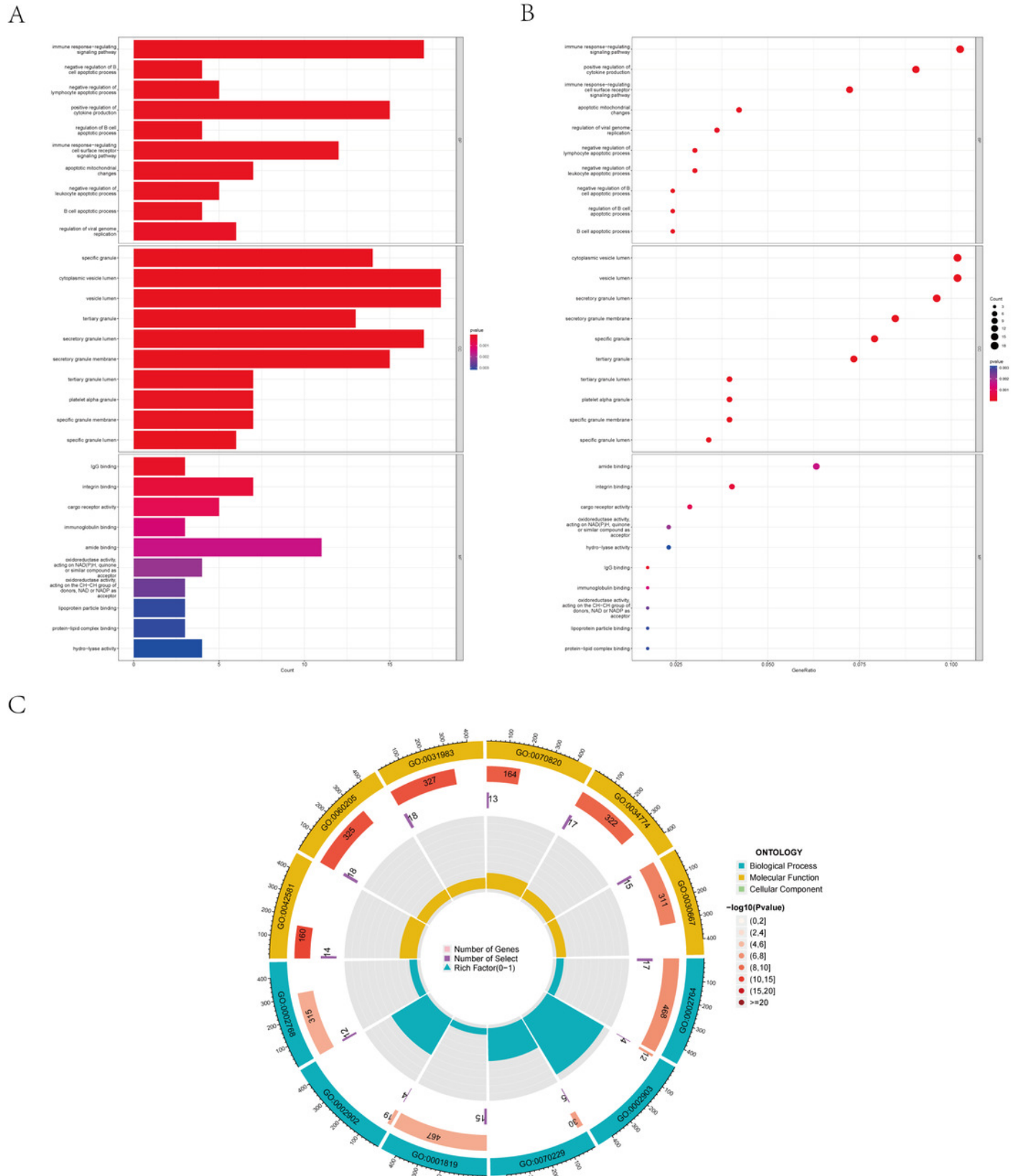
B



# Figure 3

*Show the top 10 significantly enriched BP (biological process), CC (cellular component) considerably, and MF (molecular function) in the barplot and the bubble plot.*

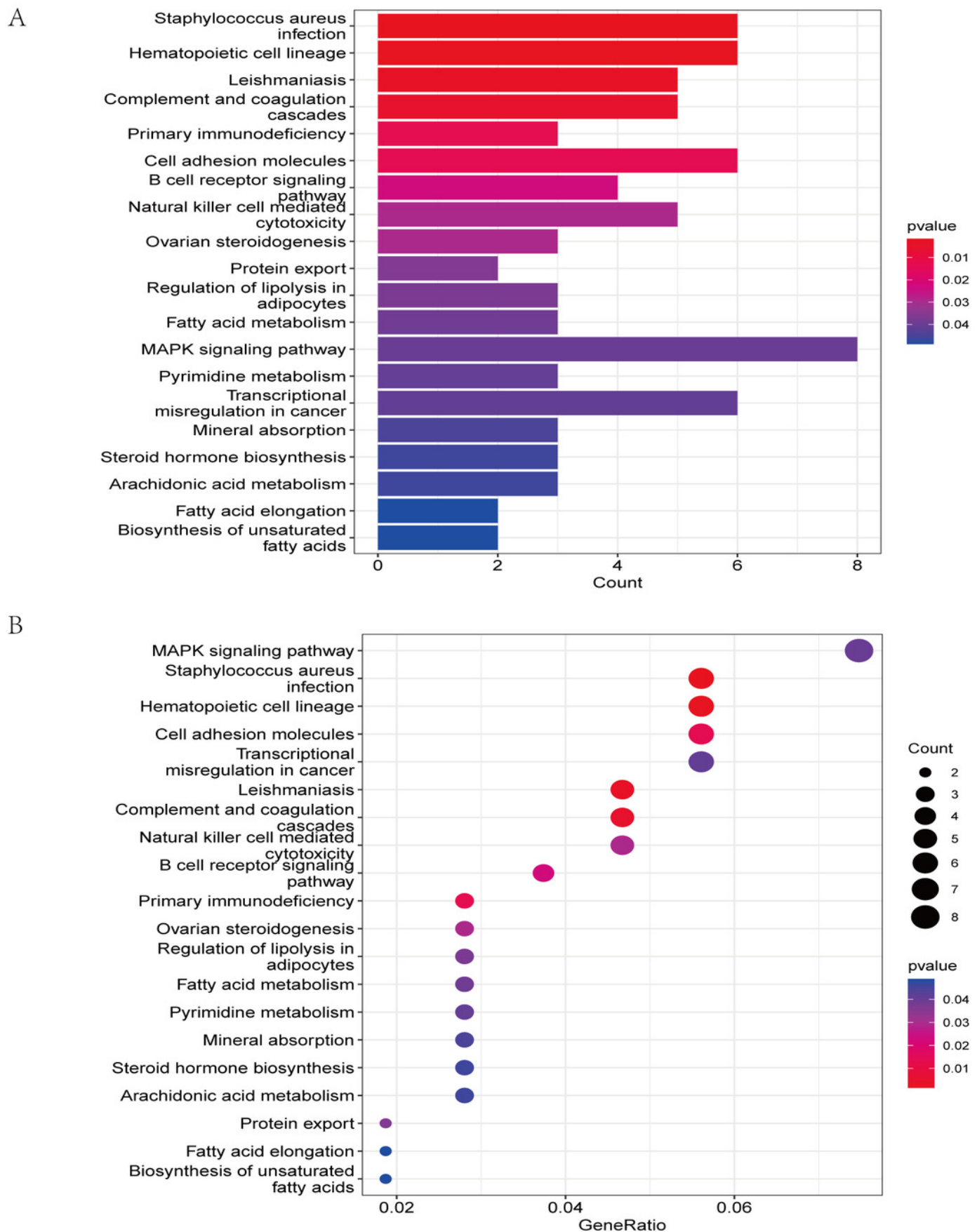
*(A) In the barplot, the bar size represents the number of genes, and the colour shades represent the p-value. (B) In the bubble plot, the x-axis represents the gene ratio, the y-axis represents the  $-\log_{10}$  (FDR) value, the bubble size represents the number of genes, and the colour shades represent the size of the p-value. (A) and (B) The  $p\text{-value} \leq 0.001$  is shown in red;  $0.001 < p\text{-value} \leq 0.002$  is shown in pink;  $0.002 < p\text{-value} \leq 0.003$  is shown in purple;  $p\text{-value} > 0.003$  is shown in blue. Immunoglobulin binding and amide binding are shown in pink. Oxidoreductase activity, acting on NAD(P)H, quinone or similar compound as acceptor and oxidoreductase activity, acting on the CH-CH group of donors, NAD or NADP as acceptor are shown in purple. Lipoprotein particle binding, protein-lipid complex binding and hydro-lyase activity are shown in blue. Others are shown in red. (C) The Gene Ontology (GO) enrichment analysis results of DEGs are shown as circle plots. First circle: enriched classification, outside the process, is the sitting scale for the number of genes. Different colours represent different classifications; Second circle: the number of genes in the category in the background and the Q or P value. The more genes the longer the bar and the smaller the value the redder the colour; Third circle: the total number of foreground genes; Fourth circle: RichFactor value for each classification (the number of foreground genes divided by the number of background genes in that classification), with each cell of the auxiliary background line indicating 0.1.*



# Figure 4

*The KEGG pathway analysis results of DEGs are shown as the barplot and the bubble plot.*

*(A) In the barplot, the bar size represents the number of genes, and the colour shades represent the p-value. (B) In the bubble plot, the x-axis represents the gene ratio, the y-axis represents the  $-\log_{10}$  (FDR) value, the bubble size represents the number of genes, and the colour shades represent the size of the p-value. The  $p\text{-value} \leq 0.01$  is shown in red;  $0.01 < p\text{-value} \leq 0.02$  is shown in light red;  $0.02 < p\text{-value} \leq 0.03$  is shown in light purple;  $0.03 < p\text{-value} \leq 0.04$  is shown in purple;  $p\text{-value} > 0.04$  is shown in blue. Staphylococcus aureus infection, Hematopoietic cell lineage, Leishmaniasis and Complement and coagulation cascades are shown in red. Primary immunodeficiency and Cell adhesion molecules are shown in light red.*



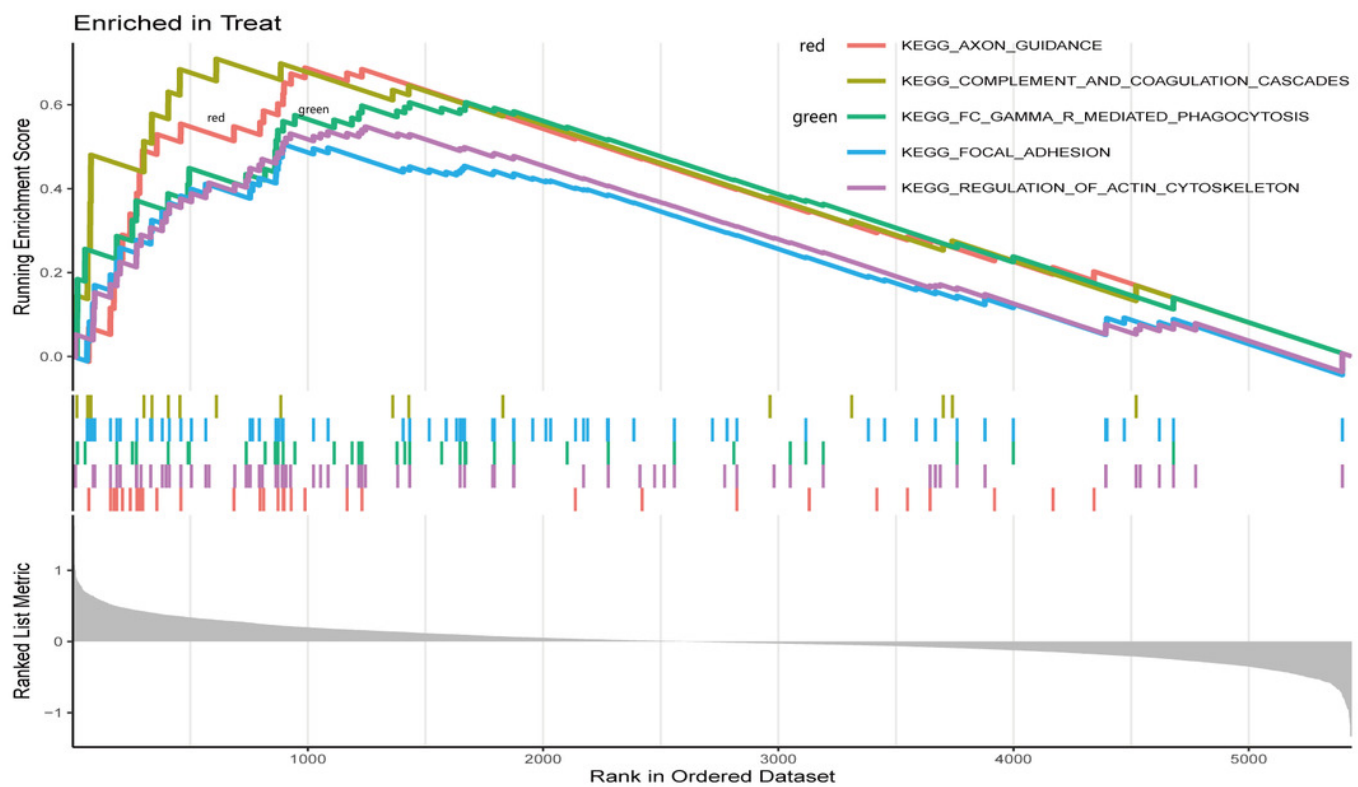
# Figure 5

*Gene Set Enrichment Analysis of IS and CTL group in the GSE dataset.*

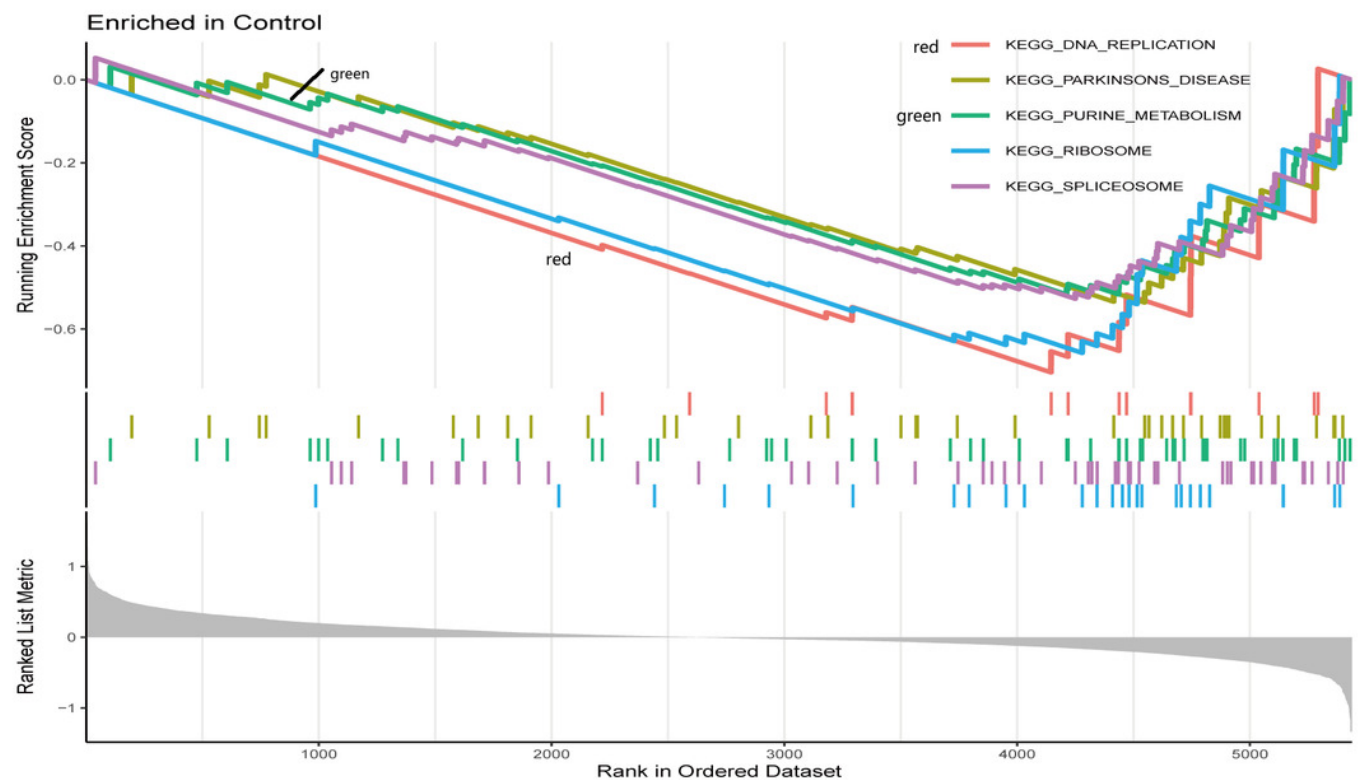
*The different colours represent the different pathways obtained by enrichment. The figure has three parts. Part I: The top part is a line graph of the Gene Enrichment Score. The vertical axis is the corresponding Running ES. There is a peak at the top or bottom of the line. The core genes are those before the peak for the treatment group with a positive Enrichment score, while for the CTL group with a negative Enrichment score, the core genes are those after the peak. The horizontal axis represents each gene under this gene set and corresponds to the barcode-like vertical line in the second part. The second part: Hits, each vertical line corresponds to a gene under this gene set. The third part: is the distribution of rank values for all genes, and the vertical coordinate is the ranked list metric, which is the value of the rated amount of that gene. (A) the GSEA results of IS group. (B) the GSEA results of the CTL group.*



A



B

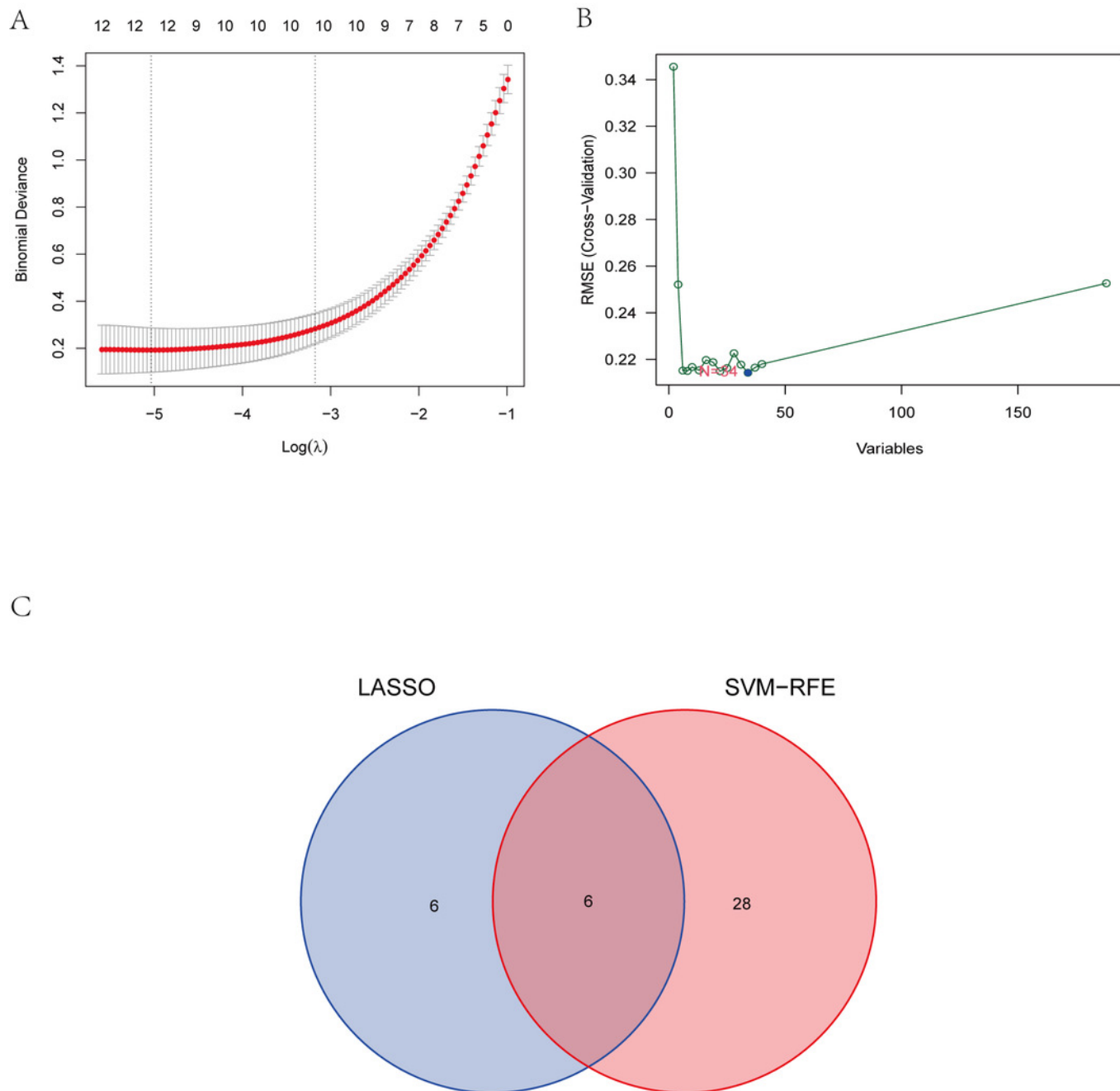


# Figure 6

*The identification of 6 Hub genes.*

*(A) The LASSO logistic regression algorithm identified 12 IS-related features with a 10-fold cross-validation set for selecting the penalty parameter to determine the optimal lambda value. (B) A total of 34 feature genes were filtered out using the SVM-RFE algorithm. (C) Venn diagram of genes extracted from LASSO and SVM-RFE methods. Abbreviations: LASSO (least absolute shrinkage and selection operator), SVM (support vector machine), RFE (recursive feature elimination).*

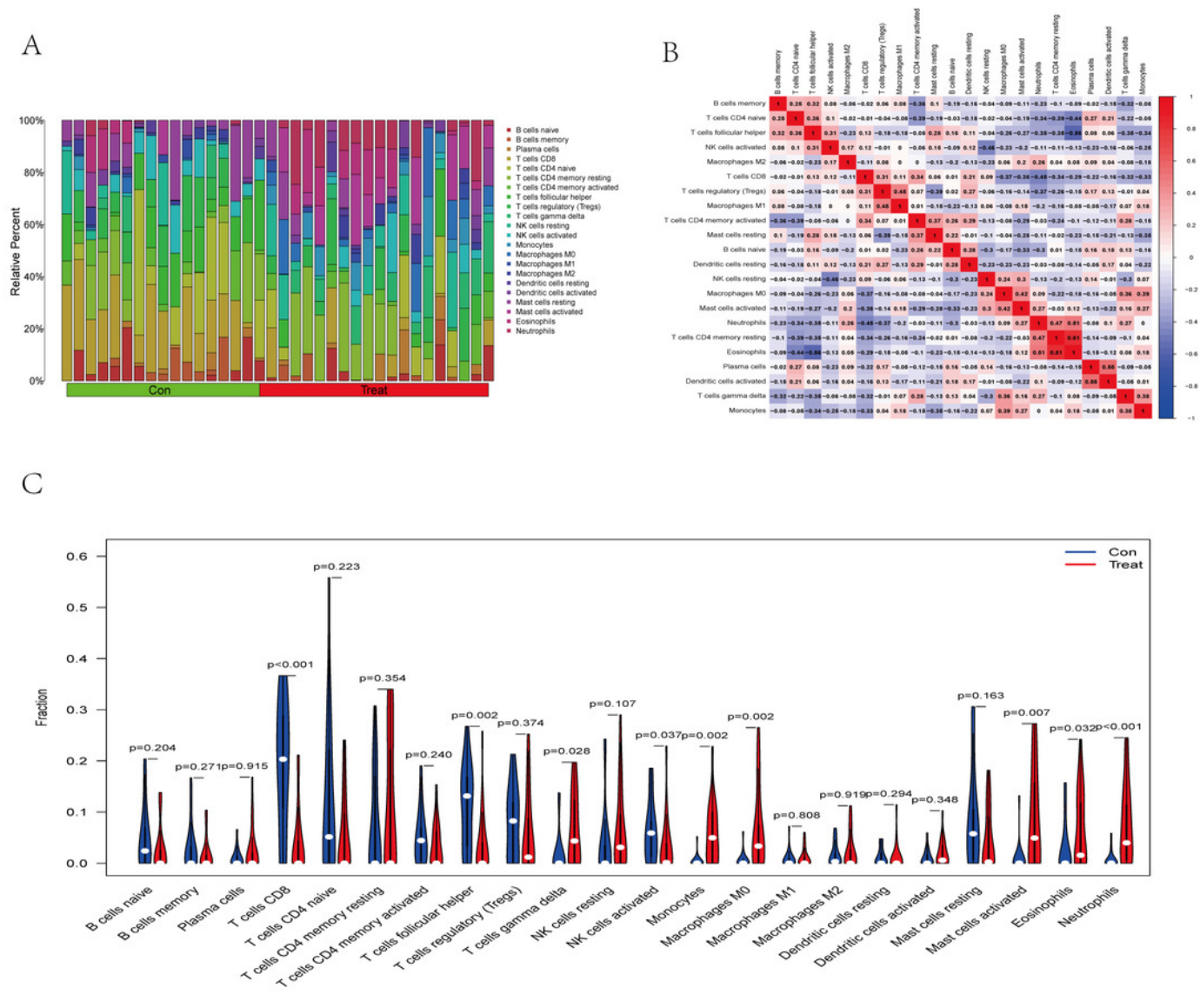




# Figure 7

Immune infiltration landscape between IS and Control obtained by CIBERSORT analysis.

(A) The bar-plot diagram indicates the relative percentage of different types of immune cells between IS and CTL. (B) The heatmap shows the correlation in the infiltration of innate immune cells. (C) Violin plot showing the difference in immune cell infiltration between IS (red) and Control (purple),  $P < 0.05$ , was considered statistically significant.

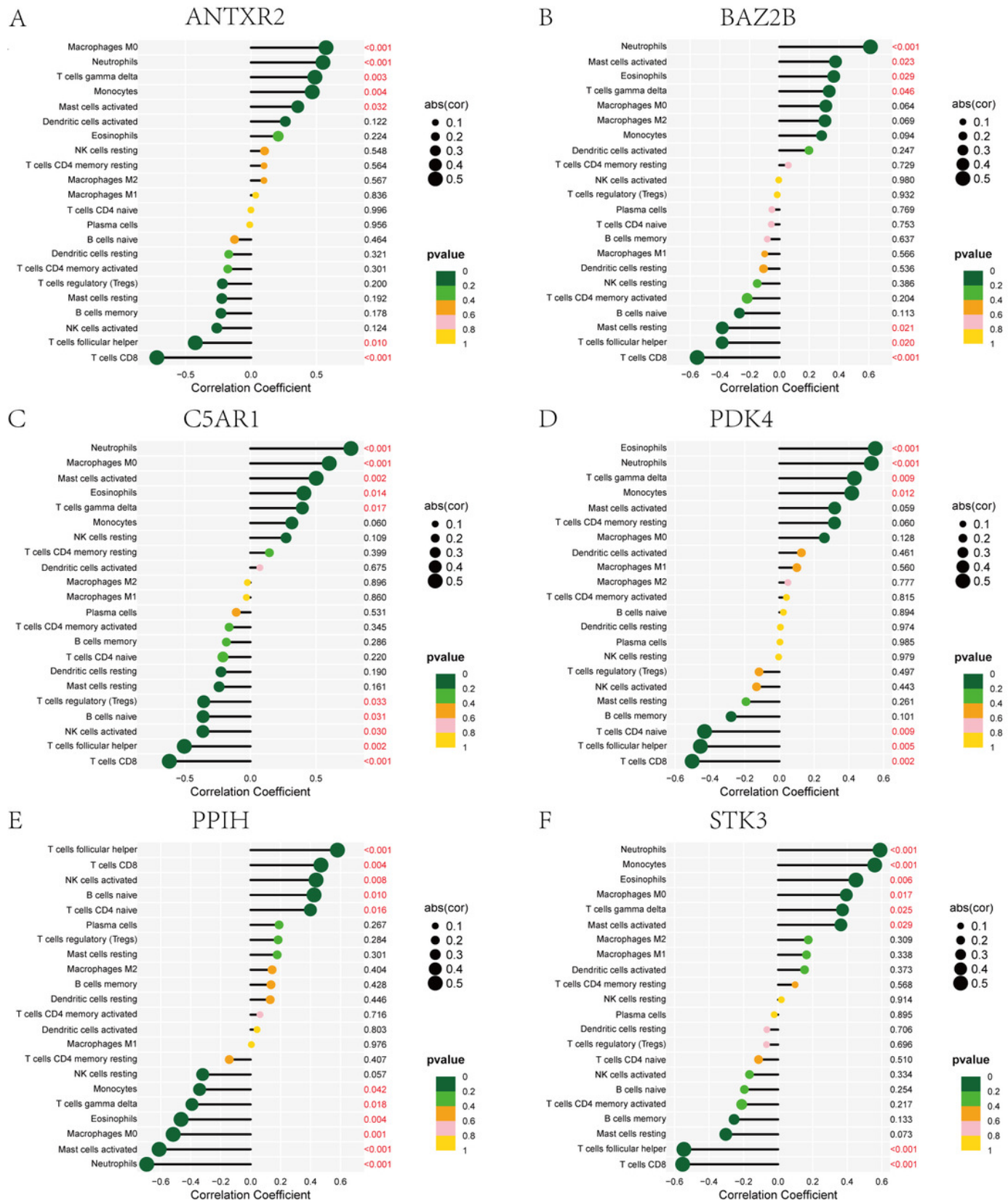


# Figure 8

*Lollipop charts show the correlations between hub genes and the infiltration level.*

*The p-value  $\leq 0.2$  is shown in green;  $0.2 < p\text{-value} \leq 0.4$  is shown in lime green;  $0.4 < p\text{-value} \leq 0.6$  is shown in orange;  $0.6 < p\text{-value} \leq 0.8$  is shown in pink;  $0.8 < p\text{-value} \leq 1$  is shown in yellow.*

(A) ANTXR2 (B) BAZ2B (C)C5AR1 (D)PDK4 (E) PPIH (F) STK3

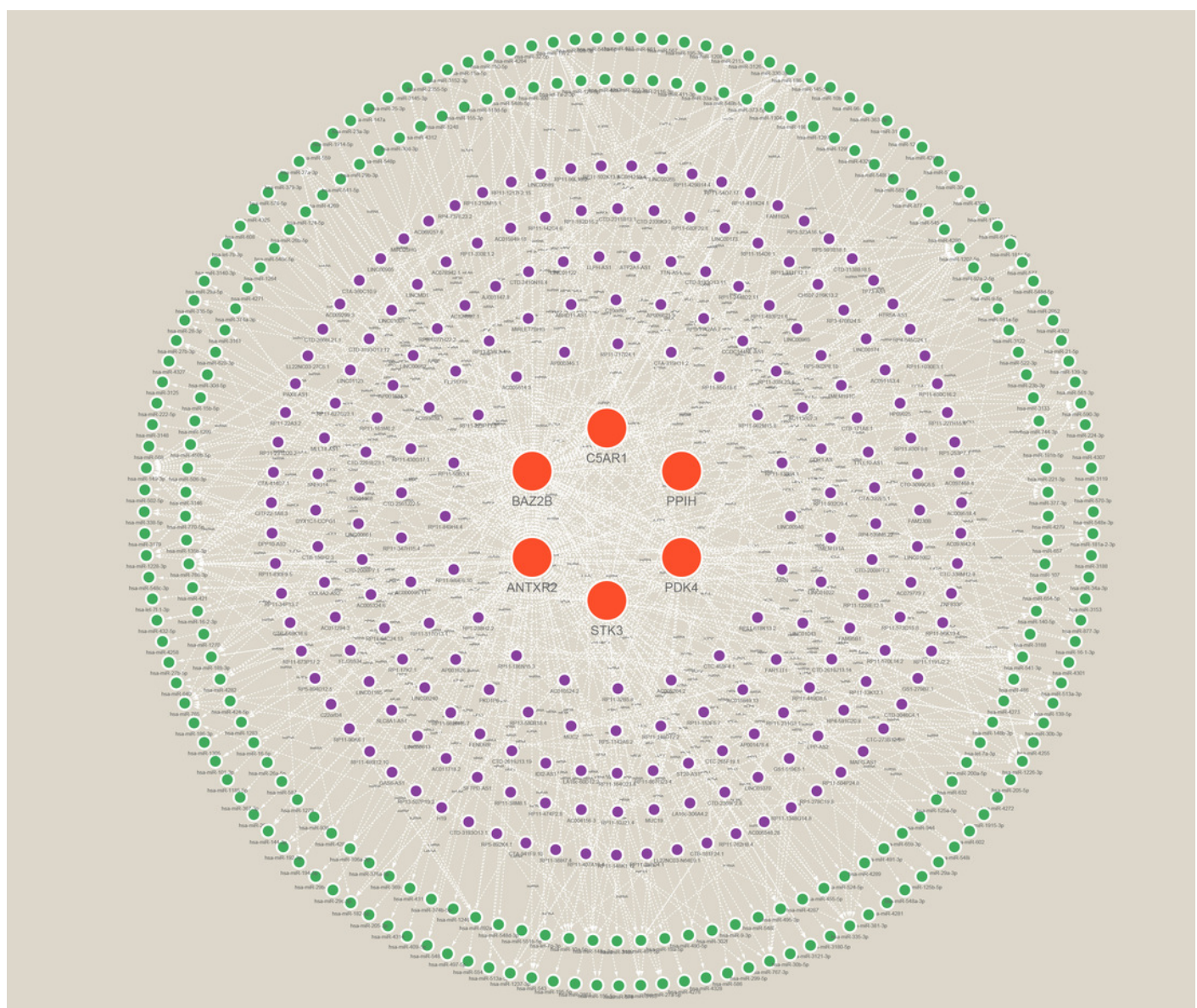




# Figure 9

*A ceRNA network based on Hub genes.*

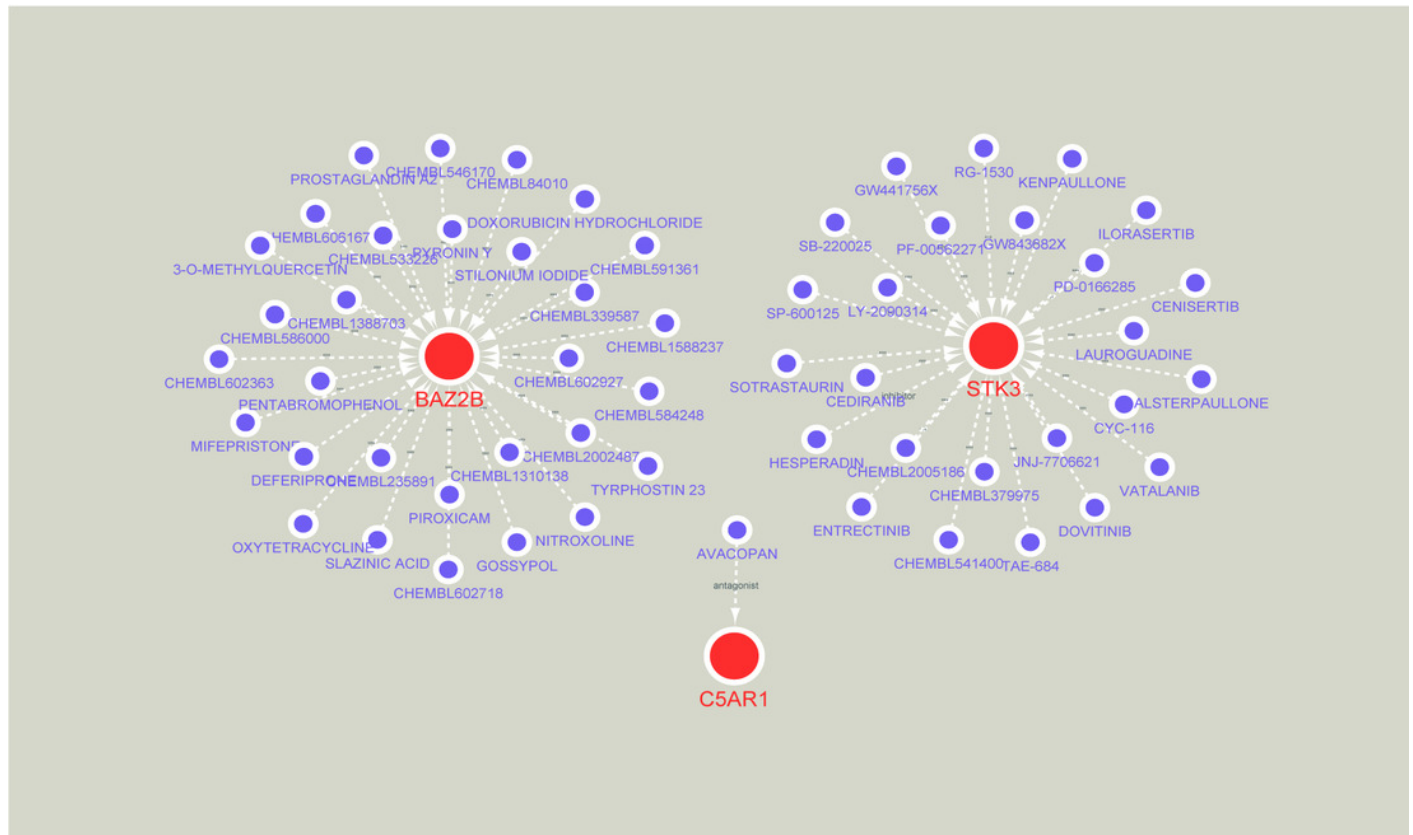
*The network includes nodes (6 mRNAs, 249 miRNAs, 218 lncRNAs) with 591 edges. The most central part uses a red orbs representing the hub genes. The outermost two laps are shown with green orbs representing the miRNAs. And purple orbs represent lncRNAs.*



# Figure 10

*Prediction of marker gene-targeted drugs.*

*Red orbs represent up-regulated mRNA and purple orbs represent drugs.*



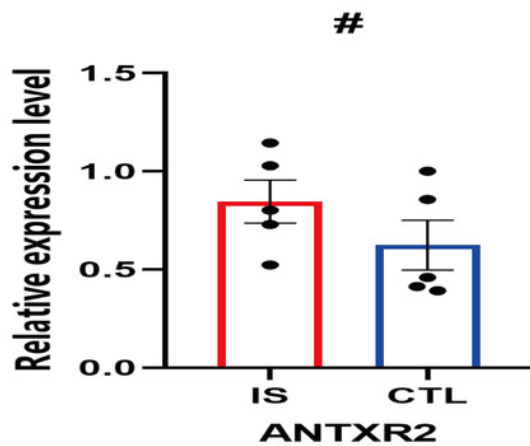
# Figure 11

*Prediction of marker gene-targeted drugs.*

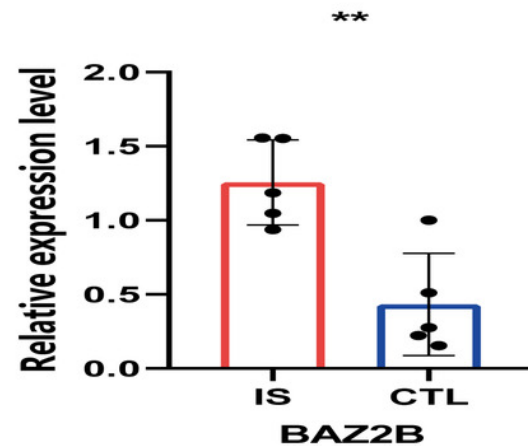
*The relative expression levels of (A) ANTXR2,  $P=0.2233$  (B) BAZ2B,  $P=0.0034$  (C) C5AR1,  $P=0.0018$  (D) PDK4,  $P=0.0026$  (E) PPIH,  $P=0.9792$  (F) STK3,  $P=0.0226$  in IS and CTL. # $P>0.05$ , \*  $0.01 \leq p\text{-value} \leq 0.05$  \*\* $P<0.01$*

*Results for the IS group are shown in red. Results for the CTL group are shown in blue.*

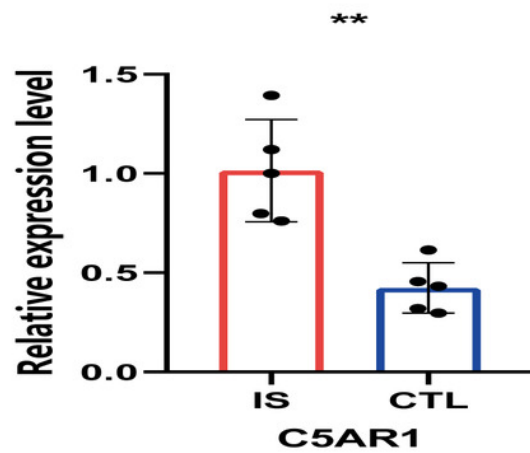
A



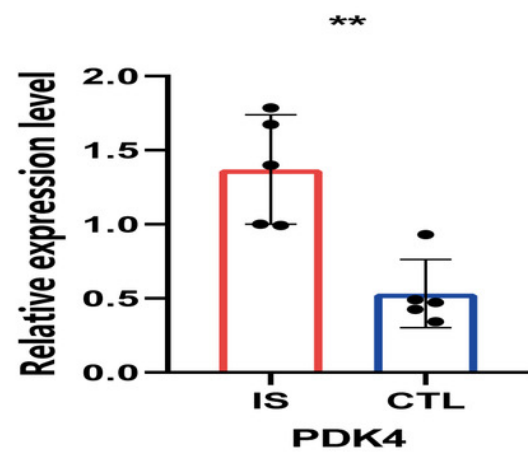
B



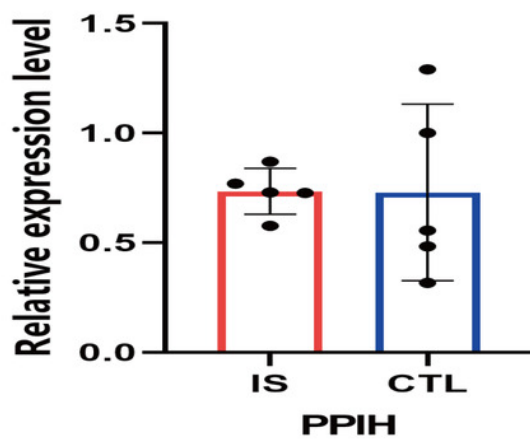
C



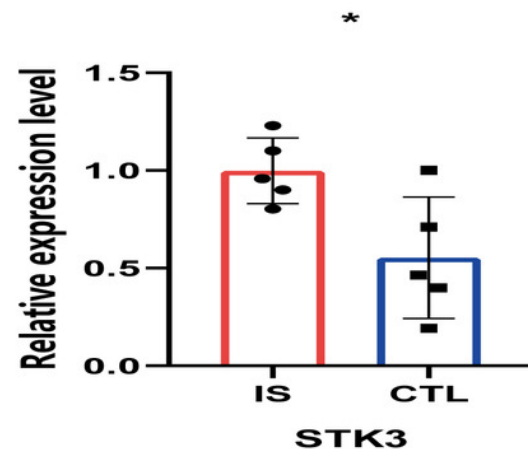
D



E



F



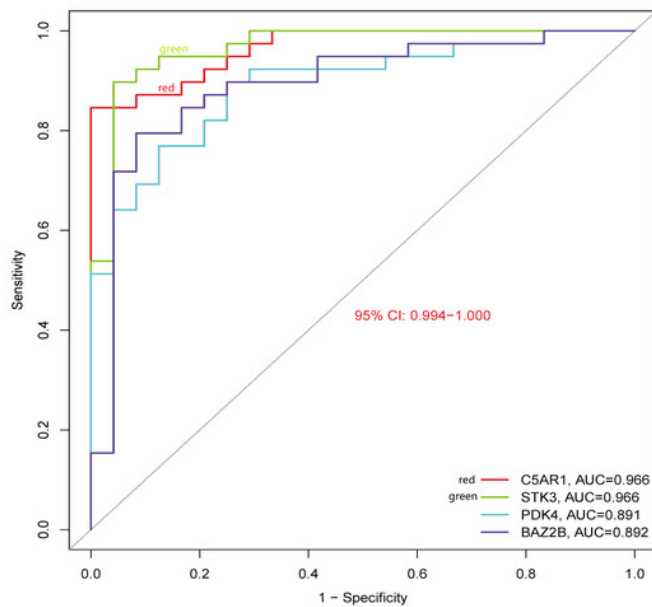


# Figure 12

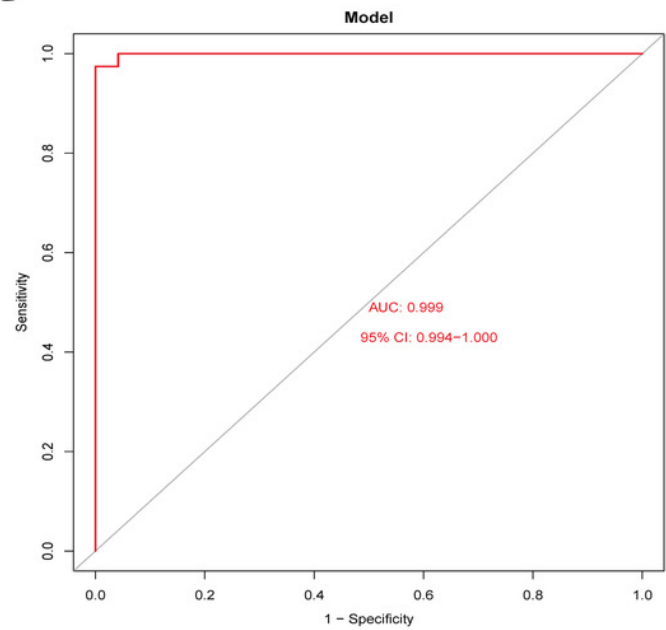
*Diagnostic model building and ROC curve validation.*

*(A) ROC curves for evaluating the accuracy of hub genes of the training set. (B) ROC curves for evaluating the accuracy of the model of the training set. (C) ROC curves for assessing the accuracy of hub genes using the validation set. (D) ROC curves for assessing the accuracy of the model using the validation set.*

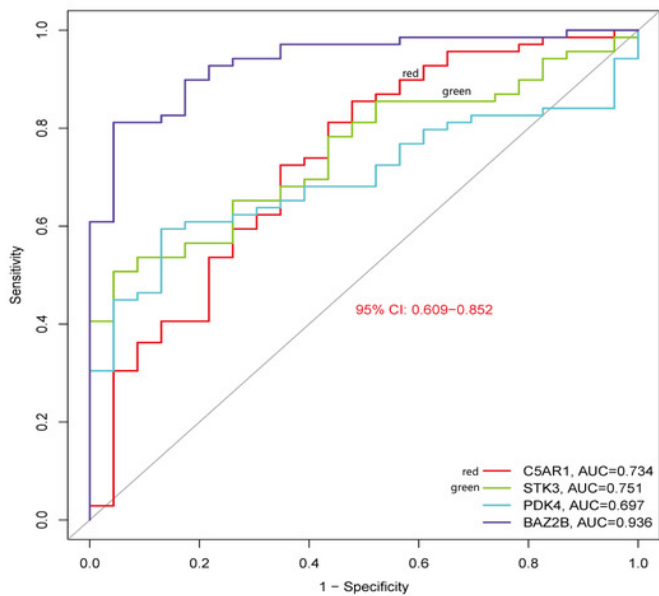
A



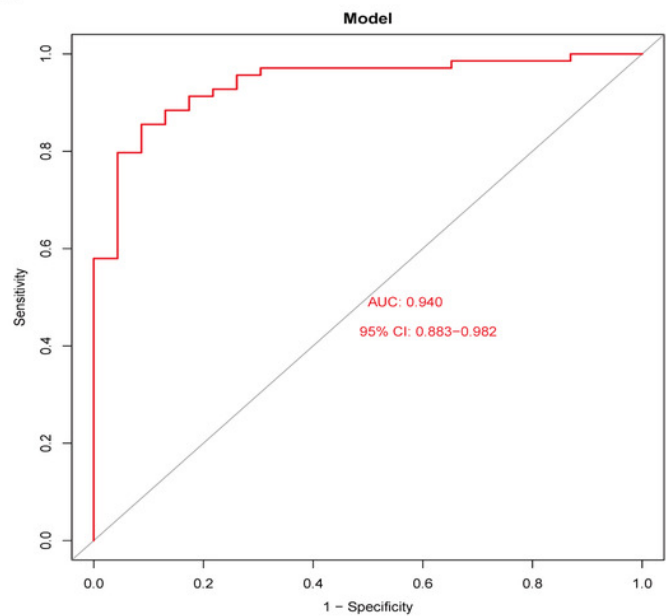
B



C



D



# Figure 13

*External validation of the key genes.*

*The expression levels of hub genes (A) BAZ2B (B) C5AR1 (C) PDK4 (D) STK3 in the GSE58294 testing set. The result for the Treat group is shown in red. The result for the Con group is shown in blue.*

