Exploring Genetic Associations of Three Types of Risk Factors with Ischemic Stroke: An Integrated Bioinformatics Study

Yi Liu[1#], PhD; Weili Wang[1#], MMedSci; Xin Cui[1*], PhD; Yanming Xie[1*], Prof.

[1]Institute of Basic Research in Clinical Medicine, China Academy of Chinese Medical Sciences Beijing, 100700, China;

[#]Yi Liu and Weili Wang contributed equally to this work.

[*]Yanming Xie and Xin Cui are the joint corresponding authors of this article.

Correspondence:

Yanming Xie No. 16, Nanxiaojie, Dongzhimennei, Dongcheng District, Beijing, China

E-mail: ketizu2018@163.com

Xin Cui No. 16, Nanxiaojie, Dongzhimennei, Dongcheng District, Beijing, China

E-mail: Xinroberttcm@hotmail.com

Short Title: Associations of exposures and ischemic stroke

The total number of words: 8817 words

Abstract：

Background: Ischemic stroke (IS) is a primary cause of disability and mortality globally. More and more reports suggest a strong association between blood pressure, blood glucose, and blood lipids and their metabolic products with IS.

Methods: We extracted the genetic tools of blood pressure, blood glucose, and blood lipids and their metabolites as instrumental variables, which were then paired with GWAS data on IS and a Mendelian randomization (MR) analysis was performed to assess the effect of these exposures on the disease. Following the positive results, colocalization analysis was performed to identify shared genes associated with exposures and IS. We then performed differential expression analysis using the GEO dataset to identify the differentially expressed associated genes (DEAGs) from associated shared genes. Additional analyzes were performed on these DEAGs to obtain their importance scores using four machine learning models. A nomogram was created using genes with high importance scores to predict the level of risk assessment between DEAGs and IS.

Results: There is a positive correlation between blood pressure, blood glucose and the risk of IS onset, while blood lipids and their metabolic products are positively or negatively correlated with the risk. There are 64 shared genes of blood pressure, blood lipids and their metabolic products with IS. Thirteen DEAGs were obtained, and among which FURIN, MAN2A2, HDDC3, ALDH2, and TOMM40 were identified as feature genes for creating the nomogram which can quantitatively predict the risk of IS onset with the expression of these feature genes. By cluster analysis, we found that DEAGs expression underlying immune inflammation, angiogenesis and development, lipid metabolism, etc.

Conclusion: This study suggests a significant association between blood pressure, blood glucose, and blood lipids and their metabolic products with IS, and predicts that these exposures mainly regulate the occurrence, development, and prognosis of IS through mechanisms such as DNA repair, DNA methylation, mitochondrial repair, apoptosis, autophagy, etc.

Key words: Mendelian Randomization; Colocalization analysis; Bioinformatics; Ischemic stroke; Blood pressure; Blood glocose; blood lipids and their metabolic productes;

Abbreviations: Area Under the Curve, AUC; Deoxyribonucleic acid, DNA; Differentially expressed associated genes, DEAGs; differentially expressed genes, DEGs; Extreme Gradient Boosting model, XGB; false discovery rate, FDR; Gene Expression Omnibus , GEO; Gene ontology, GO; gene set variation analysis, GSVA; Generalized Linear, GL; genome-wide association study, GWAS; high-density lipoprotein, HDL; Instrumental variables, IVs; Integrated Epidemiology Unit, IEU; intermediate-density lipoprotein, IDL; inverse variance weighted, IVW; Ischemic stroke, IS; Kyoto Encyclopedia of Genes and Genomes, KEGG; low-density lipoprotein, LDL; Mendelian randomization, MR; Natural killer, NK; one-way analysis of variance, ANOVA; posterior probability, PP; Principal Component Analysis, PCA; Random Forest model, RF; receiver operating characteristic curves, ROC; Single nucleotide polymorphisms, SNPs; Single-sample gene set enrichment analysis, ssGSEA; Support Vector Machine model, SVM; very low-density lipoprotein, VLDL

Ischemic stroke (IS) is a disease caused by the blockage of blood vessels in the brain, leading to local cerebral hypoxia and ischemia, resulting in the death of brain cells. This blockage is usually caused by thrombosis or embolism, causing neurological deficits in the brain area with insufficient blood supply.[1-3] In 2019, the number of patients who died from stroke worldwide reached 6.55 million, making it the main cause of adult disability and death worldwide.[4,5] Among newly diagnosed stroke cases, IS accounted for 62.4%.[6] Therefore, identifying the risk factors for IS is crucial for preventing this disease. Blood pressure, blood glucose, and blood lipids and their metabolic products have received widespread attention as potential factors affecting the risk of IS. A large number of clinical and epidemiological studies have shown that controlling blood pressure, blood glucose, and blood lipids can reduce the risk of IS.[7-9] In recent years, with a deeper understanding of the human genome, we have come to realize that the expression of these biomarkers is largely influenced by genetic information.[10] Therefore, exploring high-risk groups from a genetic perspective and identifying shared genes related to blood pressure, blood glucose, and blood lipids and their metabolic products with IS is of great significance for the precise prevention and treatment of IS.

Mendelian randomization (MR) is a genetic epidemiological research tool that is widely used to assess the potential causal relationship between exposure and disease.[11,12] The working mechanism of MR analysis is similar to that of a natural randomized controlled trial. It uses genetic instruments (single nucleotide polymorphisms, SNPs) that are strongly correlated with exposure factors and are not affected by confounding factors under the principle of MR distribution as instrumental variables (IVs).[13,14] This study used a

two-sample MR analysis to explore the potential causal effects of blood pressure, blood glucose, and blood lipids and their metabolic products on IS. The selected datasets are all from the Integrated Epidemiology Unit (IEU) database (https://gwas.mrcieu.ac.uk/). This database is public and contains nearly 2.45 trillion genetic associations from 42,334 summary datasets of genome-wide association study (GWAS).

Colocalization analysis is a method for assessing whether two correlated traits are driven by the same genetic mechanism. This method often uses the Bayesian model coloc.[15] The coloc model assumes that in each tested region, each trait has at most one association point, and calculates the posterior probability (PP) of all possible association patterns of the two traits through approximate Bayesian factors: 1) H0: no association; 2) H1: only related to trait 1; 3) H2: only related to trait 2; 4) H3: related to both trait 1 and trait 2, but through two independent SNPs; 5) H4: related to both trait 1 and trait 2, through a shared SNP. The posterior probabilities of each association pattern are denoted as PP0 to PP4.[15] A higher PP4 value (e.g., PP4 > 50%) provides colocalization support, indicating the existence of shared genetic variation between the two traits.[16] This study used colocalization analysis to find shared genes associated with blood pressure, blood glucose, blood lipids and their metabolic products, and IS. These shared genes may provide important clues for revealing the pathogenesis of IS.

The Gene Expression Omnibus (GEO) database collects genomic datasets from different species and tissues, which can be used in conjunction with bioinformatics methods to discover evidence of specific gene expression, identify disease prediction factors, and ultimately help us understand the mechanisms by which the genome regulates physiological

and pathological states.[17-19] Therefore, based on the associated shared genes of blood pressure, blood glucose, blood lipids and their metabolic products with IS, this study performed differential analysis on normal samples and IS samples in the GEO dataset, and constructed a machine learning model to screen out feature genes among the associated shared genes. On the one hand, it verifies the MR results through real human samples, and on the other hand, it further screens out biomarkers with higher importance, to further explore the mechanisms of blood pressure, blood glucose, blood lipids and their metabolic products on IS.

1. Study design

In this study, aligned with our research objective, we selected GWAS summary datasets for blood pressure, blood glucose, and blood lipids and their metabolic products with IS. We conducted a two-sample MR analysis and chose positive results for further gene colocalization analysis to identify genes associated with the exposure factors and IS. Subsequently, we retrieved two datasets of IS (training and validation sets), encompassing normal and IS groups, from the GEO database (https://www.ncbi.nlm.nih.gov/geo/). Differential analysis, correlation analysis, and immune infiltration analysis were carried out on the normal and IS groups of the training set. With the help of machine learning, we derived feature genes related to blood pressure, blood glucose, blood lipids and their metabolic products, and IS. Following this, the results of the machine learning model were validated using the validation set, and two rounds of clustering analysis on the IS group samples were performed to further investigate the functional mechanisms of the exposure-related genes (Figure 1).

1

2 1.1 Data sources

3 In accordance with the goals of this study, we utilized the IEU database, conducting
4 searches with the keywords of "VLDL (very low-density lipoprotein)", "LDL (low-density
5 lipoprotein)", "IDL (intermediate-density lipoprotein)", "HDL (high-density lipoprotein)",
6 "apoliprotein", "triglyceride", "fatty acid class", "systolic pressure", "diastolic pressure",
7 "blood glucose", "glycosylated hemoglobin", and "IS". This can identify genetic instruments
8 for blood pressure, blood glucose, and blood lipids and their metabolic products as exposure
9 variables. Concurrently, IS-related genetic instruments were selected as outcome variables for
10 an MR analysis, aiming to explore potential causal relationships between exposures and
11 outcomes. By using "ischemic stroke" as the keyword, setting the data type as array-based
12 expression profiles, and specifying the species as homo sapiens, we retrieved samples from
13 the GEO database (https://www.ncbi.nlm.nih.gov/geo/) to obtain gene expression and clinical
14 data of IS patients and healthy individuals. Perl code was employed for gene symbol
15 annotation and data adjustment, thus deriving expression levels of genes related to blood
16 pressure, blood glucose, blood lipids and their metabolites, and IS, in both the normal and IS
17 groups.

18

19 1.2 Mendelian Randomization analysis

20 We utilized the "TwoSampleMR" package to conduct a two-sample MR analysis on the
21 relationship between blood pressure, blood glucose, and blood lipids and their metabolic
22 products with IS.[20] In the analysis, we set the parameters as follows: 1) The selected genetic

instruments need to have a strong correlation with the exposure factors, and the corresponding P-value should be less than 5×10^{-8}; 2) In conducting Linkage Disequilibrium (LD) clumping, we set the r^2 threshold at 0.01; 3) When performing clumping analysis, we set the window size as 10,000 kilobase pairs.[21] These stringent parameter settings help us to accurately reveal the causal relationship between exposures and outcomes.

1.3 Colocalization analysis

After completing the two-sample MR analysis, we selected positive results that demonstrated significant associations between the exposure factors and outcome in the MR analysis. We used the "ieugwasr" and "coloc" packages to conduct colocalization analysis on these positive results to determine if the exposure factors and outcome might be associated within the same gene region. We set a posterior probability (PP4) greater than 50% as the standard, indicating that the genes in that region are associated with both exposure and outcomes[16]. Finally, we employed the "gassocplot2" package to visualize the colocalization results, clearly demonstrating the associated shared genes of blood pressure, blood glucose, and blood lipids and their metabolic products with IS.

1.4 Identification of differently expressed associated genes (DEAGs) and analyses of DEAGs

Expression levels of genes associated with blood pressure, blood glucose, blood lipids and their metabolic products, and IS were extracted from both the normal and IS groups. Differential expression analysis was performed using the "limma", "pheatmap", "ggpubr" and other R packages, and results were presented as box plots and heatmaps. Genes with a p-value less than 0.05 were defined as differentially expressed associated genes (DEAGs).

Using Perl coding, DEAGs were located on chromosomes and their positions were displayed on a circos plot created with the "Rcircos" package. In addition, the "cor" command was used to calculate the correlation coefficient between each two DEAGs and the results were visualized to illustrate the correlation among DEAGs.

1.5 Analysis of immune cells in IS samples

We performed 1000 simulations using the CIBERSORT command in R to obtain a total relative amount of immune cells equal to 1, then visualized the content of immune cells in each sample using a bar plot. Single-sample gene set enrichment analysis (ssGSEA) was performed using the "GSVA" and "GSABase" packages to compare the differences in immune cell contents between the normal group and the IS group, and the results of ssGSEA are presented as box plot. We matched the differentially expressed associated genes (DEAGs) with the ssGSEA scores, performed a correlation test to obtain the correlation coefficient, and then visualized the results as a heat map.

1.6 Identification of feature genes and construction and validation of nomogram based on ML

Expression data from DEAGs was utilized to construct four predictive models: Random Forest model (RF), Support Vector Machine model (SVM), Generalized Linear Model (GL), and Extreme Gradient Boosting model (XGB). The results from these four models were calculated based on the prediction function. Feature genes within DEAGs were then selected through comprehensive analysis of reverse cumulative distribution plots, residual box plots, and receiver operating characteristic (ROC) curves. After selecting the optimal model, line plots were constructed using the expression levels of the feature genes in both the normal and

IS groups. Lastly, decision and calibration curves were drawn to assess the accuracy of the line plots. Another dataset that includes both a normal and IS group was obtained from the GEO database, and a machine learning model was constructed using the same R language approach as before. The ROC was plotted to validate the machine learning model constructed in the test dataset.

1.7 Clustering of DEAGs and analysis between DEAG clusters.

The "ConsensusClusterPlus" package in R was used for clustering, employing Euclidean distance type and allowing up to nine clusters. Expression levels between clusters were compared using heatmaps and box plots, and principal component analysis (PCA) was applied to assess differences between clusters. Following this, a ssGSEA was performed on the DEAG clusters, generating bar plots of the amount of individual immune cells in each sample within the different clusters and comparing the differences in immune cell content among different clusters. Gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analyses were performed using gmt files downloaded from the GSEA platform (http://www.gsea-msigdb.org/), and gene set variation analysis (GSVA) was conducted in R language to analyze the expression of enrichment items between clusters. Finally, under the filtering conditions of $|logFC| > 1$ and an adjusted P-value < 0.05, differential expression analysis was performed on the gene expression of DEAG clusters. Intersection of DEAG clusters through a Venn diagram yielded differentially expressed genes (DEGs).

1.8 Clustering of DEGs and analysis between DEG clusters

The same clustering method as in section 1.7 was applied to cluster DEGs, and the DEG cluster with the highest precision was selected. Based on the DEGs clustering results, we compared the expression levels of DEGs, the differences in DEAGs expression, and the immune cell content in different clusters. These results were visualized using heatmaps and box plots.

1.9 Construction of DEAG scores

The PCA method was employed to calculate the expression levels of DEAGs for each sample, yielding DEAG scores (The formula is shown below).[22] Using R packages such as "limma" and "ggpubr", we performed differential analysis on the scores of DEAGs in both significantly differentially expressed core gene clusters and DEG clusters. We created box plots to illustrate the scores of significantly differentially expressed core genes in samples within the significantly differentially expressed core gene and DEG clusters. In addition, an alluvial diagram was drawn using the package "ggalluvial" to visualize the relationships and overall processes among DEAG clusters, DEG clusters, samples with higher scores, and samples with lower scores of DEAGs.

$$\text{DEAG Score} = \sum (\text{PC1}_i + \text{PC2}_i)$$

1.10 Statistical analysis

This study carried out MR analysis and colocalization analysis using R V4.2.1. In the MR analysis, if the number of SNPs for the exposure factor is less than 5 after filtering, we will use the Wald ratio method for analysis; otherwise, we will opt for the inverse variance

weighted (IVW) method for analysis. In colocalization analysis, to maximize the acquisition of shared genes between "exposure" and "outcome", we set the chromosomal locus parameter range from 75 kilobase pairs to 500 kilobase pairs, aiming to include a wider area around the SNP to enhance the likelihood of discovering shared genes related to both exposure and outcome. In the bioinformatics part, we used Strawberry Perl 5.32.1.1 for GEO dataset extraction and data annotation, and all other statistical analyses were performed with R V4.2.1. For two independent samples, we applied a t-test, while for two paired samples we used the Wilcoxon paired-rank sum test, and for three or more groups of data, we employed one-way analysis of variance (ANOVA) and the Kruskal-Wallis rank sum test. The Spearman rank correlation test was used for correlation analysis. We set a P value < 0.05 or a false discovery rate (FDR) < 0.05 after correction by the Benjamini-Hochberg method as being statistically significant.

2 Results

2.1 Results of collection of GWAS and GEO datasets

We retrieved 516 relevant exposure datasets from the ieu database based on the set keywords, and the IS dataset came from 440,328 participants of European ancestry, including 34,217 cases and 406,111 controls. These datasets were used in two-sample MR analysis for IS to explore their potential causal relationships. By searching the GEO database for IS datasets, we filtered out two datasets that met the requirements, namely GSE16561 and GSE22255, using the former as the test set and the latter as the validation set. The former is

total RNA data from human peripheral blood samples, including 39 IS samples (diagnosed by MRI) and 24 normal samples; the latter is RNA expression data from human peripheral blood monocytes, including 20 IS samples and 20 normal samples.

2.2 Results of Mendelian randomization analysis

Among the 516 exposure datasets, 114 showed a positive association with IS. These positive results further suggested that increased levels of Apolipoprotein A-I and HDL may reduce the risk of IS, while elevated levels of VLDL, LDL, IDL, Apolipoprotein B, triglycerides, fatty acids, systolic blood pressure, diastolic blood pressure, fasting blood sugar, and glycated hemoglobin might increase the risk of IS. (Please see Attachment 1 for detailed analysis results)

2.3 Results of colocalization analysis

Upon completion of the two-sample MR analysis, we further subjected the 114 positive results to colocalization analysis to study the potential shared genes between the exposure factors and IS. The results showed that the PP4 was greater than 50% for 70 of the positive results, including exposure factors such as LDL, IDL, VLDL, Apolipoprotein B, fatty acids, systolic and diastolic blood pressure. This suggests that these exposure factors may have genes shared with IS. The colocalization results were visualized using the "gassocplot2" package, and we finally identified 64 shared genes between these exposure factors and IS. These genes may play a key role in the association between the selected exposure factors and

IS. (Please see Attachment 1 for detailed analysis results and Attachment 2 for specific shared genes situations)

2.4 Results of DEAGs identification and analysis of DEAGs.

Through colocalization analysis, we identified 64 associatied shared genes between blood pressure, blood sugar,and blood lipids and their metabolites with IS. Differential analysis between the IS group and the normal group in the test dataset revealed that 13 associatied shared genes exhibited statistically significant differences. Among them, ERCC2, FEN1, HDDC3, TOMM40, TRAPPC6A were highly expressed in the normal group, while ALDH2, FADS2, FES, FURIN, MAN2A2, PVRL2, SYT7, and FAM109A were highly expressed in the IS group, as shown in Figures A and B. The specific positions of these significantly differentially expressed association genes on the chromosome are shown in Figure C. The correlation analysis between each pair of DEAGs in the IS samples showed that there is a certain correlation between DEAGs, which are mainly positive correlations, as shown in Figures D and E.

2.5 Results of Immune cell infiltration analysis, immune cell difference analysis and correlation analysis of IS samples

Through immune cell infiltration analysis, we obtained the types and content of immune cells expressed in each sample, as shown in Figure A. ssGSEA analysis revealed (Figure B) that the expression differences of 11 types of immune cells in the control group and IS group were statistically significant. Among them, B cells naive, T cells CD8, T cells CD4 memory

activated, T cells follicular helper, Natural killer (NK) cells activated, and Dendritic cells activated were highly expressed in the control group, while T cells gamma delta, Macrophages M0, Macrophages M2, Mast cells activated, and Neutrophils were highly expressed in the IS group. The correlation analysis between DEAGs and immune cells showed that there is a certain correlation between DEAGs and immune cells, and the levels of positive and negative correlations are comparable, as shown in Figure C.

2.6 Results of selection of machine learning models, construction of nomogram, and verification

The results of building SVM, RF, XGB, and GL machine learning prediction models using DEAGs data were obtained. From the residual boxplots, inverse cumulative distribution plots, and ROC curves (Figure A, C, and D), it can be observed that the GL method had the highest area under the ROC curve, the lowest residual values, and the lowest inverse cumulative values. Therefore, the GL method was considered the most accurate and chosen as the best model for further analysis. The GL model provided importance scores for the selected feature genes, as shown in Figure B, revealing nine feature genes. Among them, the top five genes with the highest importance scores were used to construct the column line graph (including FURIN, MAN2A2, HDDC3, ALDH2, and TOMM40). Then, separate scoring scales were obtained for these five feature genes. The risk rate of the co-located feature genes associated with blood pressure, blood glucose, and blood lipids and their metabolic products in the occurrence of IS was assessed by calculating the sum of the feature gene expression scores (Figure E). The predictive accuracy of the model was evaluated by the

distance between the solid line and the dotted line in the calibration curve (Figure F) and the distance between the red and gray lines in the decision curve (Figure G), indicating high accuracy. The GL model was validated using the validation dataset (GSE22255) and the top five feature genes with the highest importance scores were included in the model validation and ROC curve analysis. The results showed an Area Under the Curve (AUC) value of 0.528, with a 95% confidence interval of 0.167-0.889. Therefore, based on the validation of the model, it can be concluded that the model based on the GEO dataset has good accuracy.

2.7 Results of Clustering of DEAGs, Analysis of DEAGs Expression between Clusters, Immune Cell Analysis, and GO and KEGG Enrichment Analysis

Clustering analysis based on the expression of DEAGs revealed two distinct clusters with the highest accuracy, as shown in Figure A. Consequently, the IS samples were divided into two groups: C1 and C2, as depicted in Figures B and C. Subsequently, DEAGs expression analysis was performed between the two DEAG clusters (Figure D and E), indicating that ALDH2, FEN1, FES, and FURIN were significantly upregulated in C1. PCA analysis (Figure F) demonstrated that DEAGs can differentiate between C1 and C2. Moreover, ssGSEA analysis (Figure G) identified three statistically significant immune cell types between C1 and C2. Specifically, NK cells resting and Macrophages M2 were upregulated in C1, while Mast cells resting were upregulated in C2. The respective content of different immune cells in each C1 and C2 sample is illustrated in Figure H.

In GSVA analysis (Figures I and J), it can be observed that compared to C1, C2 showed upregulation of several GO biological processes, including Interleukin 6 production, positive

regulation of interleukin 6 production, cellular response to osmotic stress, neutrophil degeneration, response to lipoteichoic acid, positive regulation of reactive oxygen species metabolic process, and macropinocytosis. On the other hand, C2 exhibited downregulation of branching involved in blood vessel morphogenesis, cellular response to potassium ion, myotube cell development, regulation of heart morphogenesis, canonical Wnt signaling pathway involved in osteoblast differentiation, gonadotropin secretion, vascular process in the circulatory system, and blood vessel maturation in GO biological processes. In GO molecular functions, C2 showed upregulation of leukotriene C4 synthase activity and downregulation of sphingosine 1-phosphate phosphatase activity. In GO cellular components, C2 exhibited upregulation of early phagosome, specific granule lumen, and proton-transporting two-sector ATPase complex catalytic domain, while downregulation of photoreceptor disc membrane. In KEGG pathways, C2 demonstrated upregulation of other glycan degradation, toll-like receptor signaling pathway, sphingolipid metabolism, pentose phosphate pathway, pathogenic Escherichia coli infection, glutathione metabolism, Fc gamma R-mediated phagocytosis, neurotrophin signaling pathway, cell signaling in Helicobacter pylori infection, progesterone-mediated oocyte maturation, and endocytosis. Conversely, C2 showed downregulation of alanine, aspartate, and glutamate metabolism, cytokine-cytokine receptor interaction, taste transduction, ECM receptor interaction, hedgehog signaling pathway, arachidonic acid metabolism, neuroactive ligand-receptor interaction, ascorbate and aldarate metabolism, pentose and glucuronate interconversions, and nitrogen metabolism in KEGG pathways.

2.8 Results of DEGs screening, DEGs clustering, and analyzes of DEG clusters

After filtering for significantly differentially expressed genes (DEGs) between C1 and C2 samples, a total of 4 DEGs were identified (Figure A). These 4 DEGs were further subjected to clustering analysis within the DEAGs cluster, resulting in two distinct clusters, namely CI and CII (Figure B). Analysis of DEAGs expression in these two clusters in IS samples revealed that RNASE2, RNASE3, MMP9, and CAMP exhibited statistically significant differential expression, with higher expression in CI and lower expression in CII (Figure C).

The differential expression analysis of DEAGs based on DEGs clustering was performed on the samples, as shown in Figure D. The results revealed that these genes were mainly upregulated in the CI group and downregulated in the CII group. Among them, the significantly differentially expressed genes included ALDH2, FES, MAN2A2 (upregulated in CI), and TRAPPCA (upregulated in CII). Additionally, using ssGSEA based on DEGs clustering (using the same gene set file as above), statistically significant differences in immune cells were observed, including activated B cells (downregulated in CI), activated dendritic cells, immature dendritic cells, myeloid-derived suppressor cells (MDSCs), macrophages, mast cells, monocytes, neutrophils, and plasmacytoid dendritic cells (upregulated in CI), as shown in Figure E.

2.9 Results of DEAGs scoring, differential analysis of DEAGs score between clusters, and construction of the alluvial plot

The differential analysis of DEGs clustering based on the principal component analysis

(PCA) scores showed statistically significant differences between the clusters (Figure A). Specifically, C1 had lower scores, while C2 had higher scores. However, there was no statistically significant difference observed in the DEGs clustering (Figure B). The scatter plot in Figure C displayed that the C1 and C2 clusters of DEAGs corresponded mainly to the CI and CII clusters of DEGs, respectively. However, there was no clear correspondence observed between the high and low scores of DEAGs and the DEGs clustering.

3 Discussion

3.1. The relationships of blood pressure, blood glucose, and blood lipids and their metabolic products with IS

Clinical studies have demonstrated that effective management of various physiological indicators can significantly impact the risk of IS. For example, blood pressure control has been shown to significantly reduce the risk of IS recurrence.[23] Elevated fasting blood glucose levels are significantly associated with an increased risk of IS.[24] Furthermore, increased levels of glycated hemoglobin, regardless of diabetes status, are associated with an increased risk of IS.[25] Lipid management also plays a significant role in IS risk. There are significant differences in VLDL levels between the control and IS groups.[26] However, the role and mechanisms of VLDL level changes in IS pathogenesis require further investigation. Medications that lower LDL levels can significantly reduce the risk of IS.[27] Increased levels of IDL and apolipoprotein B are associated with an increased risk of IS.[28,29] Controlling triglyceride levels can lower the risk of IS.[30] Both polyunsaturated fatty acids and saturated

fatty acids are also associated with IS.[31,32] However, elevated levels of HDL and apolipoprotein A-I are associated with a lower risk of IS.[33,34] Taking all the above clinical studies into consideration, it can be concluded that blood pressure, blood glucose, and blood lipids and their metabolic products play critical roles in influencing the risk of IS.

The results of MR analysis indicate a potential causal relationship between blood pressure (diastolic and systolic), blood glucose (fasting blood glucose and glycated hemoglobin), blood lipids and their various metabolic products (VLDL, LDL, IDL, HDL, apolipoprotein A-I, apolipoprotein B, triglycerides, Eicosapentaenoate, Docosapentaenoate, Stearidonate, Docosahexaenoic acid) with IS. Increasing levels of high-density lipoprotein (HDL) and apolipoprotein A-I are associated with a decreased risk of IS. However, elevated blood pressure (including systolic and diastolic), blood glucose (fasting blood glucose and glycated hemoglobin), various low-density lipoproteins such as VLDL, LDL, IDL, as well as apolipoprotein B, triglycerides, Eicosapentaenoate, Docosapentaenoate, Stearidonate, Docosahexaenoic acid are associated with an increased risk of IS.

3.2 Discussion of mechanisms associated shared genes

Through MR analysis, we found that blood pressure, blood glucose, blood lipids and their metabolic products have a significant impact on the risk of IS. To further explore the specific mechanisms underlying these effects, we conducted a colocalization analysis. The results showed that there are 64 associated shared genes between LDL, IDL, VLDL, apolipoprotein B, fatty acids, systolic blood pressure, and diastolic blood pressure with IS. Although a positive association between blood glucose and IS risk has been established, we

did not identify any associated shared genes due to the PP4 being less than 50%.

The differential analysis of the 64 associated shared genes between normal and IS samples, using the GEO dataset, revealed 13 genes with significant differential expression. Among them, ERCC2, a transcription factor (TF) involved in nucleotide excision repair of damaged DNA,[35] has been clinically associated with increased stroke risk,[36,37] and animal studies have shown its neuroprotective role in preventing ischemia-reperfusion injury.[38] FEN1, another TF encoding a protein involved in rDNA and mitochondrial DNA repair,[39] has been implicated in the pathogenesis of IS and is considered a therapeutic target for IS treatment.[40,41] HDDC3, a protein encoded by HDDC3 gene, participates in the starvation response, and starvation-induced autophagy is known to protect neurons and regulate IS.[42,43] TOMM40, encoding the translocase of the outer mitochondrial membrane 40 homolog, has been identified as a potential susceptibility locus for IS based on a study conducted in Japan.[44] ALDH2, encoding aldehyde dehydrogenase, has shown neuroprotective effects by clearing 4-hydroxy-2-nonenal and reducing mitochondrial-associated cell apoptosis through JNK-mediated cystathionine-β-synthase-3 activation, making it a potential target for IS intervention. Case-control studies have also suggested an association between ALDH2 polymorphisms and IS risk in the Han Chinese population.[45-47] FADS2, encoding fatty acid desaturase 2, has been associated with IS risk and lipid levels in case-control studies, although the specific mechanisms are still under investigation.[48] The product of the FES gene exhibits tyrosine-specific protein kinase activity. Inhibitors of this class of enzymes have shown potential therapeutic value for various diseases, including IS, but the involvement of FES in IS is yet to be confirmed.[49] FURIN, encoding proprotein convertase subtilisin/kexin

type 1, has been found to be upregulated within 24 hours after cerebral ischemia in animal experiments,[50] and case-control studies have revealed lower DNA methylation levels in the IS region, suggesting its association with IS.[51] SYT7, encoding synaptotagmin 7, has been identified to have significant DNA methylation correlation with IS in a large-scale sequencing study.[52] The roles of FAM109A, MAN2A2, PVRL2, and others in IS require further investigation.

These genes primarily function through DNA repair, DNA methylation, mitochondrial repair, cell apoptosis, and autophagy to regulate lipid levels, blood pressure, and neuronal protection, thereby affecting the occurrence, development, and prognosis of IS. The associated shared genes are predominantly upregulated in the IS group, and they exhibit mostly positive regulatory relationships, indicating potential synergistic effects among these genes in IS. Differential expression analysis of immune cells between the IS and normal groups reveals significant expression differences in 50% of the immune cells, suggesting the involvement of immune cells in the regulation of IS by these associated shared genes. Moreover, the comparable levels of positive and negative regulation of immune cells by DEAGs suggest that DEAGs may have bidirectional regulatory effects on the immune system's role in IS.

The feature genes obtained from the GL model constructed through machine learning (FURIN, MAN2A2, HDDC3, ALDH2, TOMM40) have significant importance in regulating the associated shared genes and the risk of developing IS. The column plot provides a quantitative prediction of the aforementioned importance and risk of disease. Furthermore,

the construction of the GL model using the validation dataset indicates that the column plot has high accuracy.

The IS samples were clustered into C1 and C2 groups based on the expression of DEAGs, with C1 showing predominantly high expression of DEAGs and C2 showing low expression. The two clusters were significantly distinguishable according to the PCA results. Furthermore, immune cell infiltration analysis of DEAGs clustering revealed significant differences in immune cell expression between C1 and C2 groups of IS patients. The GSVA results indicated significant expression differences between C1 and C2 groups in various biological processes and pathways, including immune inflammation, vascular development and formation, reactive oxygen species metabolism, carbohydrate metabolism, lipid metabolism, fatty acid metabolism, amino acid metabolism, and neural regulation.

Differential analysis between the C1 and C2 groups identified 4 DEGs, which have been extensively studied and shown to be associated with IS incidence, IS prognosis, and IS complications.[53-55] Subsequently, IS samples were further clustered into CI and CII groups based on the expression of DEGs. Differential analysis revealed that DEGs were highly expressed in the CI group and lowly expressed in the CII group, while DEAGs were mainly upregulated in the CI group and downregulated in the CII group. Immune cell infiltration analysis of DEGs clustering demonstrated significant differences in immune cell expression between the CI and CII groups of IS patients, with upregulation observed in the CI group and downregulation observed in the CII group. PCA-based DEAG scoring and subsequent differential analysis revealed significant differences in DEAG scoring between the C1 and C2 groups, while no significant differences were observed between the CI and CII groups.

Furthermore, the construction of Sankey diagrams provided a visual representation of the correspondence between the two clustering results and the high/low DEAGs\ scoring groups, with C1 and C2 corresponding to CI, CII, and the high/low DEAG groups, while the correspondence between CI and CII was less clear.

3.3 Limitations of this study

Although MR analysis and colocalization analysis provide valuable insights into the risk factors and underlying mechanisms of IS from the perspective of exposure factors and genes, these methods also have their limitations. The pathogenesis of IS is extremely complex and involves intricate interactions among multiple environmental factors, physiological indicators, and genetic factors. Therefore, in order to comprehensively understand and elucidate the risk factors and mechanisms of IS, a holistic and multifactorial research approach is needed. Additionally, the GWAS data included in our study primarily originate from sample populations of European ancestry. Hence, extrapolating these results directly to non-European populations may have limitations. Similarly, bioinformatics methods also have certain limitations. Although they have provided multi-faceted and multi-dimensional analyses of human blood samples, elucidating specific mechanisms of interaction between blood pressure, blood glucose, blood lipids and their metabolic products in relation to IS, further research and clinical applications require careful consideration. Moreover, the blood samples used in this study were only collected from the United States and Portugal, and whether the results can be extrapolated to other countries, ethnicities, and populations remains to be investigated and confirmed. Given the relatively small sample size, there is a greater risk of bias in the conclusions. Therefore, large-scale studies or animal experiments can be conducted to further

validate and deepen the conclusions.

4 Conclusion and prospects

This study first integrated MR analysis, colocalization analysis, and bioinformatics analysis to conduct an in-depth study of the association of blood pressure, blood glucose, and blood lipids and their metabolic products with IS. Firstly, the results of MR analysis showed that there is a positive correlation between blood pressure, blood glucose and the risk of IS onset, and blood lipids and their metabolic products have either a positive or negative correlation with the risk of IS onset.Furthermore, through colocalization analysis, we identified 64 associated shared genes with blood pressure, blood lipids and their metabolic products in relation to IS. Additionally, using the GEO dataset, we identified 13 DEAGs. These genes primarily function through deoxyribonucleic acid (DNA) repair, DNA methylation, mitochondrial repair, cell apoptosis, and autophagy, regulating lipid metabolism, blood pressure, and neuroprotection, thereby influencing the occurrence, progression, and prognosis of IS. Subsequently, through the construction of machine learning models, we selected feature genes. Cluster analysis revealed that the expression of DEAGs in IS may also involve mechanisms such as immune inflammation, vascular development, lipid metabolism, and fatty acid metabolism.

The findings of this study will contribute to a deeper understanding of the pathophysiological mechanisms underlying IS, optimize treatment strategies, and drive the development of new drugs. The methodological approach used in this study, which combines genetic epidemiology with bioinformatics, offers a unique perspective for elucidating causal

relationships between exposure factors and diseases, as well as unraveling genetic mechanisms. This research strategy, along with the identification of the associated shared genes, provides a new theoretical framework and practical approach for the prevention, diagnosis, and treatment of IS.

Article Information

Acknowledgments

The authors would like to thanks the institutions for providing GWAS summary statistics and human gene expression datasets.

Source of Founding

National Administration of Traditional Chinese Medicine 2021 Qihuang Scholar Support Project (National TCM Education Letter 2022 No. 6) ; Yanming Xie's National Renowned Traditional Chinese Medicine Expert Inheritance Studio Construction Project (National TCM Education Letter 2022 No. 75).

Disclosures

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References:

1. Hankey GJ. Stroke. Lancet. 2017;389:641-654. doi: 10.1016/S0140-6736(16)30962-X.

2. Lee RHC, Lee MHH, Wu CYC, Couto E Silva A, Possoit HE, Hsieh TH, Minagar A, Lin HW. Cerebral ischemia and neuroregeneration. Neural Regen Res. 2018;13:373-385. doi: 10.4103/1673-5374.228711.

3. Powers WJ, Rabinstein AA, Ackerson T, Adeoye OM, Bambakidis NC, Becker K, Biller J, Brown M, Demaerschalk BM, Hoh B, et al. Guidelines for the Early Management of Patients With Acute Ischemic Stroke: 2019 Update to the 2018 Guidelines for the Early Management of Acute Ischemic Stroke: A Guideline for Healthcare Professionals From the American Heart Association/American Stroke Association. Stroke. 2019;50:e344-e418. doi: 10.1161/STR.0000000000000211.

4. Romano JG, Rundek T. Expanding Treatment for Acute Ischemic Stroke beyond Revascularization. N Engl J Med. 2023;388:2095-2096. doi: 10.1056/NEJMe2303184.

5. Roth GA, Mensah GA, Johnson CO, Addolorato G, Ammirati E, Baddour LM, Barengo NC, Beaton AZ, Benjamin EJ, Benziger CP, et al. Global Burden of Cardiovascular Diseases and Risk Factors, 1990-2019: Update From the GBD 2019 Study. J Am Coll Cardiol. 2020;76:2982-3021. doi: 10.1016/j.jacc.2020.11.010.

6. GBD 2019 Stroke Collaborators. Global, regional, and national burden of stroke and its risk factors, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet Neurol. 2021;20:795-820. doi: 10.1016/S1474-4422(21)00252-0.

7. Hansson L, Zanchetti A, Carruthers SG, Dahlöf B, Elmfeldt D, Julius S, Ménard J, Rahn KH, Wedel H, Westerling S. Effects of intensive blood-pressure lowering and low-dose aspirin in patients with hypertension: principal results of the Hypertension Optimal

Treatment (HOT) randomised trial. HOT Study Group. Lancet. 1998;351:1755-1762. doi: 10.1016/s0140-6736(98)04311-6.

8. Duckworth W, Abraira C, Moritz T, Reda D, Emanuele N, Reaven PD, Zieve FJ, Marks J, Davis SN, Hayward R,et al. Glucose control and vascular complications in veterans with type 2 diabetes. N Engl J Med. 2009;360:129-139. doi: 10.1056/NEJMoa0808431.

9. Amarenco P, Labreuche J. Lipid management in the prevention of stroke: review and updated meta-analysis of statins for stroke prevention. Lancet Neurol. 2009;8:453-463. doi: 10.1016/S1474-4422(09)70058-4.

10. Kim YS, Leventhal BL. Genetic epidemiology and insights into interactive genetic and environmental effects in autism spectrum disorders. Biol Psychiatry. 2015;77:66-74. doi: 10.1016/j.biopsych.2014.11.001.

11. Smith GD, Ebrahim S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? Int J Epidemiol. 2003;32:1-22. doi: 10.1093/ije/dyg070.

12. Grover S, Del Greco M F, Stein CM, Ziegler A. Mendelian Randomization. Methods Mol Biol. 2017;1666:581-628. doi: 10.1007/978-1-4939-7274-6_29.

13. Verduijn M, Siegerink B, Jager KJ, Zoccali C, Dekker FW. Mendelian randomization: use of genetics to enable causal inference in observational studies. Nephrol Dial Transplant. 2010;25:1394-1398. doi: 10.1093/ndt/gfq098.

14. Davies NM, Holmes MV, Davey Smith G. Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians. BMJ. 2018;362:k601. doi: 10.1136/bmj.k601.

15. Giambartolomei C, Vukcevic D, Schadt EE, Franke L, Hingorani AD, Wallace C, Plagnol

V. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS Genet. 2014;10:e1004383. doi: 10.1371/journal.pgen.1004383.

16. Yarmolinsky J, Bouras E, Constantinescu A, Burrows K, Bull CJ, Vincent EE, Martin RM, Dimopoulou O, Lewis SJ, Moreno V, et al. Genetically proxied glucose-lowering drug target perturbation and risk of cancer: a Mendelian randomisation analysis. Diabetologia. 2023. doi: 10.1007/s00125-023-05925-4.

17. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Marshall KA,et al. NCBI GEO: archive for high-throughput functional genomic data. Nucleic Acids Res. 2009;37(Database issue):D885-890. doi: 10.1093/nar/gkn764.

18. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, et al. NCBI GEO: archive for functional genomics data sets--update. Nucleic Acids Res. 2013;41(Database issue):D991-995. doi: 10.1093/nar/gks1193.

19. Xu B, Wang L, Zhan H, Zhao L, Wang Y, Shen M, Xu K, Li L, Luo X, Zhou S, et al. Investigation of the Mechanism of Complement System in Diabetic Nephropathy via Bioinformatics Analysis. J Diabetes Res. 2021;2021:5546199. doi: 10.1155/2021/5546199.

20. Hemani G, Zheng J, Elsworth B, Wade KH, Haberland V, Baird D, Laurin C, Burgess S, Bowden J, Langdon R,et al. The MR-Base platform supports systematic causal inference across the human phenome. Elife. 2018;7:e34408. doi: 10.7554/eLife.34408.

21. Xiang M, Wang Y, Gao Z, Wang J, Chen Q, Sun Z, Liang J, Xu J. Exploring causal correlations between inflammatory cytokines and systemic lupus erythematosus: A

Mendelian randomization. Front Immunol. 202;13:985729. doi: 10.3389/fimmu.2022.985729.

22. Zhang B, Wu Q, Li B, Wang D, Wang L, Zhou YL. m6A regulator-mediated methylation modification patterns and tumor microenvironment infiltration characterization in gastric cancer. Mol Cancer. 2020;19:53. doi: 10.1186/s12943-020-01170-0.

23.Biffi A, Anderson CD, Battey TW, Ayres AM, Greenberg SM, Viswanathan A, Rosand J. Association Between Blood Pressure Control and Risk of Recurrent Intracerebral Hemorrhage. JAMA. 2015;314:904-912. doi: 10.1001/jama.2015.10082.

24. Boden-Albala B, Cammack S, Chong J, Wang C, Wright C, Rundek T, Elkind MS, Paik MC, Sacco RL. Diabetes, fasting glucose levels, and risk of ischemic stroke and vascular events: findings from the Northern Manhattan Study (NOMAS). Diabetes Care. 2008;31:1132-1137. doi: 10.2337/dc07-0797.

25. Nomani AZ, Nabi S, Ahmed S, Iqbal M, Rajput HM, Rao S. High HbA1c is associated with higher risk of ischaemic stroke in Pakistani population without diabetes. Stroke Vasc Neurol. 2016;1:133-139. doi: 10.1136/svn-2016-000018.

26. Bansal BC, Sood AK, Bansal CB. Familial hyperlipidemia in stroke in the young. Stroke. 1986;17:1142-1145. doi: 10.1161/01.str.17.6.1142.

27. Schwartz GG, Steg PG, Szarek M, Bhatt DL, Bittner VA, Diaz R, Edelberg JM, Goodman SG, Hanotin C, Harrington RA, et al. Alirocumab and Cardiovascular Outcomes after Acute Coronary Syndrome. N Engl J Med. 2018;379:2097-2107. doi: 10.1056/NEJMoa1801174.

28. Berger JS, McGinn AP, Howard BV, Kuller L, Manson JE, Otvos J, Curb JD, Eaton CB, Kaplan RC, Lynch JK, et al. Lipid and lipoprotein biomarkers and the risk of ischemic stroke in postmenopausal women. Stroke. 2012;43:958-66. doi: 10.1161/STROKEAHA.111.641324.

29. Johannesen CDL, Mortensen MB, Langsted A, Nordestgaard BG. ApoB and Non-HDL Cholesterol Versus LDL Cholesterol for Ischemic Stroke Risk. Ann Neurol. 2022;92:379-389. doi: 10.1002/ana.26425.

30. Das Pradhan A, Glynn RJ, Fruchart JC, MacFadyen JG, Zaharris ES, Everett BM, Campbell SE, Oshima R, Amarenco P, Blom DJ, et al. Triglyceride Lowering with Pemafibrate to Reduce Cardiovascular Risk. N Engl J Med. 2022 ;387:1923-1934. doi: 10.1056/NEJMoa2210645.

31. Saber H, Yakoob MY, Shi P, Longstreth WT Jr, Lemaitre RN, Siscovick D, Rexrode KM, Willett WC, Mozaffarian D. Omega-3 Fatty Acids and Incident Ischemic Stroke and Its Atherothrombotic and Cardioembolic Subtypes in 3 US Cohorts. Stroke. 2017;48:2678-2685. doi: 10.1161/STROKEAHA.117.018235.

32. Venø SK, Schmidt EB, Jakobsen MU, Lundbye-Christensen S, Bach FW, Overvad K. Substitution of Linoleic Acid for Other Macronutrients and the Risk of Ischemic Stroke. Stroke. 2017;48:3190-3195. doi: 10.1161/STROKEAHA.117.017935.

33. Curb JD, Abbott RD, Rodriguez BL, Masaki KH, Chen R, Popper JS, Petrovitch H, Ross GW, Schatz IJ, Belleau GC, et al. High density lipoprotein cholesterol and the risk of stroke in elderly men: the Honolulu heart program. Am J Epidemiol. 2004;160:150-157. doi: 10.1093/aje/kwh177.

34. O'Donnell MJ, McQueen M, Sniderman A, Pare G, Wang X, Hankey GJ, Rangarajan S, Chin SL, Rao-Melacini P, Ferguson J, et al. Association of Lipids, Lipoproteins, and Apolipoproteins with Stroke Subtypes in an International Case Control Study (INTERSTROKE). J Stroke. 2022;24:224-235. doi: 10.5853/jos.2021.02152.

35. National Center for Biotechnology Information (2023). PubChem Gene Summary for Gene 2068, ERCC2 - ERCC excision repair 2, TFIIH core complex helicase subunit (human). Retrieved June 14, 2023 from https://pubchem.ncbi.nlm.nih.gov/gene/ERCC2/human.

36. Shyu HY, Shieh JC, Ji-Ho L, Wang HW, Cheng CW. Polymorphisms of DNA repair pathway genes and cigarette smoking in relation to susceptibility to large artery atherosclerotic stroke among ethnic Chinese in Taiwan. J Atheroscler Thromb. 2012;19:316-325. doi: 10.5551/jat.10967.

37. LD, Dawsey SM, Dong ZW, Taylor PR, Mark SD. A prospective study of polymorphisms of DNA repair genes XRCC1, XPD23 and APE/ref-1 and risk of stroke in Linxian, China. J Epidemiol Community Health. 2007;61:737-741. doi: 10.1136/jech.2006.048934.

38. Zhang J, Guo F, Zhou R, Xiang C, Zhang Y, Gao J, Cao G, Yang H. Proteomics and transcriptome reveal the key transcription factors mediating the protection of Panax notoginseng saponins (PNS) against cerebral ischemia/reperfusion injury. Phytomedicine. 2021;92:153613. doi: 10.1016/j.phymed.2021.153613.

39. National Center for Biotechnology Information (2023). PubChem Gene Summary for Gene 2237, FEN1 - flap structure-specific endonuclease 1 (human). Retrieved June 15, 2023 from https://pubchem.ncbi.nlm.nih.gov/gene/FEN1/human.

40. He Z, Ning N, Zhou Q, Khoshnam SE, Farzaneh M. Mitochondria as a therapeutic target for ischemic stroke. Free Radic Biol Med. 2020;146:45-58. doi: 10.1016/j.freeradbiomed.2019.11.005.

41. Andrabi SS, Parvez S, Tabassum H. Ischemic stroke and mitochondria: mechanisms and targets. Protoplasma. 2020;257:335-343. doi: 10.1007/s00709-019-01439-2.

42. Sun D, Lee G, Lee JH, Kim HY, Rhee HW, Park SY, Kim KJ, Kim Y, Kim BY, Hong JI, et al. A metazoan ortholog of SpoT hydrolyzes ppGpp and functions in starvation responses. Nat Struct Mol Biol. 2010;17:1188-94. doi: 10.1038/nsmb.1906.

43. He S, Wang C, Dong H, Xia F, Zhou H, Jiang X, Pei C, Ren H, Li H, Li R, et al. Immune-related GTPase M (IRGM1) regulates neuronal autophagy in a mouse model of stroke. Autophagy. 2012;8:1621-1627. doi: 10.4161/auto.21561.

44. Yamase Y, Horibe H, Ueyama C, Fujimaki T, Oguri M, Kato K, Arai M, Watanabe S, Yamada Y. Association of *TOMM40* and *SLC22A4* polymorphisms with ischemic stroke. Biomed Rep. 2015;3:491-498. doi: 10.3892/br.2015.457.

45. Guo JM, Liu AJ, Zang P, Dong WZ, Ying L, Wang W, Xu P, Song XR, Cai J, Zhang SQ, et al. ALDH2 protects against stroke by clearing 4-HNE. Cell Res. 2013;23:915-930. doi: 10.1038/cr.2013.69.

46. Xia P, Zhang F, Yuan Y, Chen C, Huang Y, Li L, Wang E, Guo Q, Ye Z. ALDH 2 conferred neuroprotection on cerebral ischemic injury by alleviating mitochondria-related apoptosis through JNK/caspase-3 signing pathway. Int J Biol Sci. 2020;16:1303-1323. doi: 10.7150/ijbs.38962.

47. Sun S, He J, Zhang Y, Xiao R, Yan M, Ren Y, Zhu Y, Jin T, Xia Y. Genetic

polymorphisms in the ALDH2 gene and the risk of ischemic stroke in a Chinese han population. Oncotarget. 2017;8:101936-101943. doi: 10.18632/oncotarget.21803.

48. Yang Q, Yin RX, Cao XL, Wu DF, Chen WX, Zhou YJ. Association of two polymorphisms in the FADS1/FADS2 gene cluster and the risk of coronary artery disease and ischemic stroke. Int J Clin Exp Pathol. 2015;8:7318-7331.

49. Gągało I, Rusiecka I, Kocić I. Tyrosine Kinase Inhibitor as a new Therapy for Ischemic Stroke and other Neurologic Diseases: is there any Hope for a Better Outcome? Curr Neuropharmacol. 2015;13:836-844. doi: 10.2174/1570159x13666150518235504.

50. Yokota N, Uchijima M, Nishizawa S, Namba H, Koide Y. Identification of differentially expressed genes in rat hippocampus after transient global cerebral ischemia using subtractive cDNA cloning based on polymerase chain reaction. Stroke. 2001;32:168-174. doi: 10.1161/01.str.32.1.168.

51. Peng H, Fan Y, Li J, Zheng X, Zhong C, Zhu Z, He Y, Zhang M, Zhang Y. DNA Methylation of the Natriuretic Peptide System Genes and Ischemic Stroke: Gene-Based and Gene Set Analyses. Neurol Genet. 2022;8:e679. doi: 10.1212/NXG.0000000000000679.

52. Zhang H, Mo X, Wang A, Peng H, Guo D, Zhong C, Zhu Z, Xu T, Zhang Y. Association of DNA Methylation in Blood Pressure-Related Genes With Ischemic Stroke Risk and Prognosis. Front Cardiovasc Med. 2022;9:796245. doi: 10.3389/fcvm.2022.796245.

53. Sundström J, Söderholm M, Borné Y, Nilsson J, Persson M, Östling G, Melander O, Orho-Melander M, Engström G. Eosinophil Cationic Protein, Carotid Plaque, and Incidence of Stroke. Stroke. 2017;48:2686-2692. doi: 10.1161/STROKEAHA.117.018450.

54. Zhong C, Yang J, Xu T, Xu T, Peng Y, Wang A, Wang J, Peng H, Li Q, Ju Z, et al. Serum matrix metalloproteinase-9 levels and prognosis of acute ischemic stroke. Neurology. 2017;89:805-812. doi: 10.1212/WNL.0000000000004257.

55. Ponsaerts L, Alders L, Schepers M, de Oliveira RMW, Prickaerts J, Vanmierlo T, Bronckaers A. Neuroinflammation in Ischemic Stroke: Inhibition of cAMP-Specific Phosphodiesterases (PDEs) to the Rescue. Biomedicines. 2021;9:703. doi: 10.3390/biomedicines9070703.

TABLE 1| Gene name conversion table

Gene description	Gene symbol
ERCC excision repair 2, TFIIH core complex helicase subunit	ERCC2
flap structure-specific endonuclease	FEN1
HD domain containing 3	HDDC3
translocase of outer mitochondrial membrane 40	TOMM40
trafficking protein particle complex subunit 6A	TRAPPC6A
aldehyde dehydrogenase 2 family member	ALDH2
fatty acid desaturase 2	FADS2
FES proto-oncogene, tyrosine kinase	FES
furin, paired basic amino acid cleaving enzyme	FURIN
mannosidase alpha class 2A member 2	MAN2A2
nectin cell adhesion molecule 2	PVRL2
synaptotagmin 7	SYT7
PH domain containing endocytic trafficking adaptor 1	FAM109A

FIGURE 1 | Flow diagram of this study design

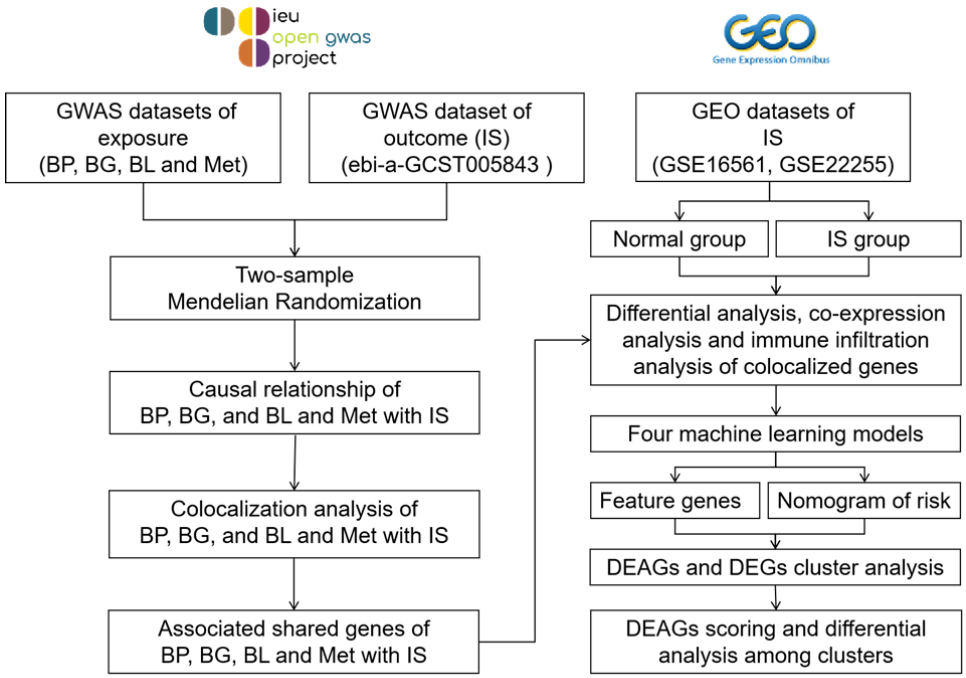


FIGURE 2 | (A) Box plot of expression difference analysis of associated shared genes between normal samples and IS samples; (B) Heat map of DEAGs expression in normal and IS samples; (C) Circle plot of chromosome location of DEAGs; (D) DEAGs correlation network; (E) Correlation analysis between the two DEAGs

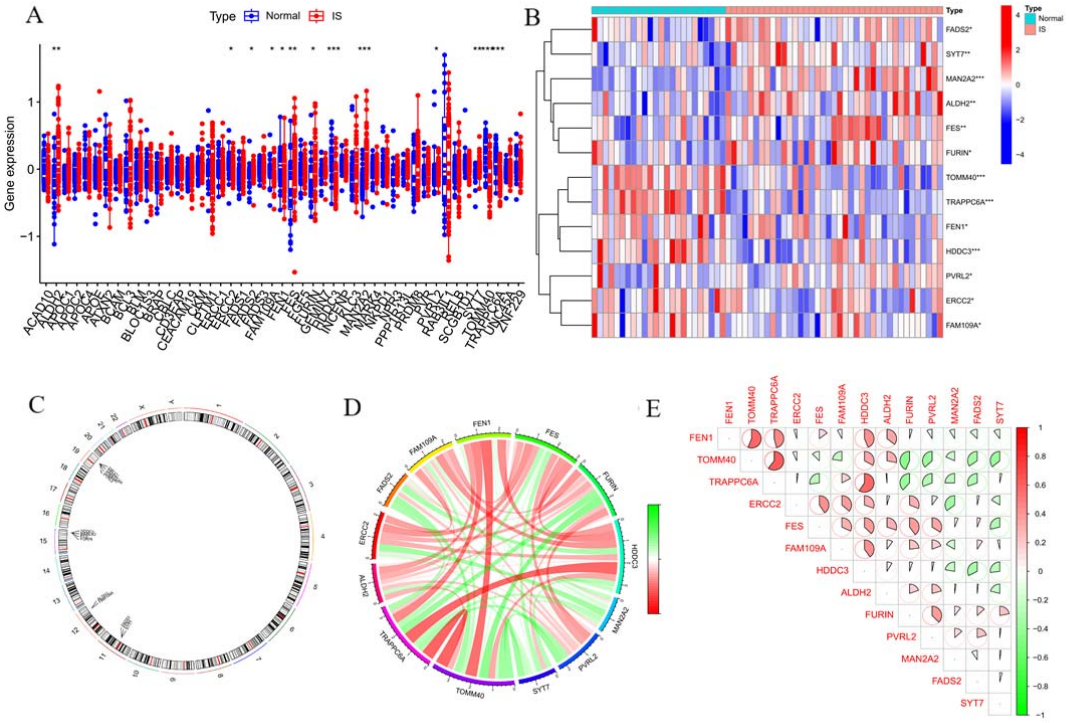


FIGURE 3 | (A) Bar plot of relative percentage of each immune cells in samples; (B) Box plot of immune cell fraction between normal samples and IS samples; (C) Heat map of correlation analysis between DEAGs and immune cells

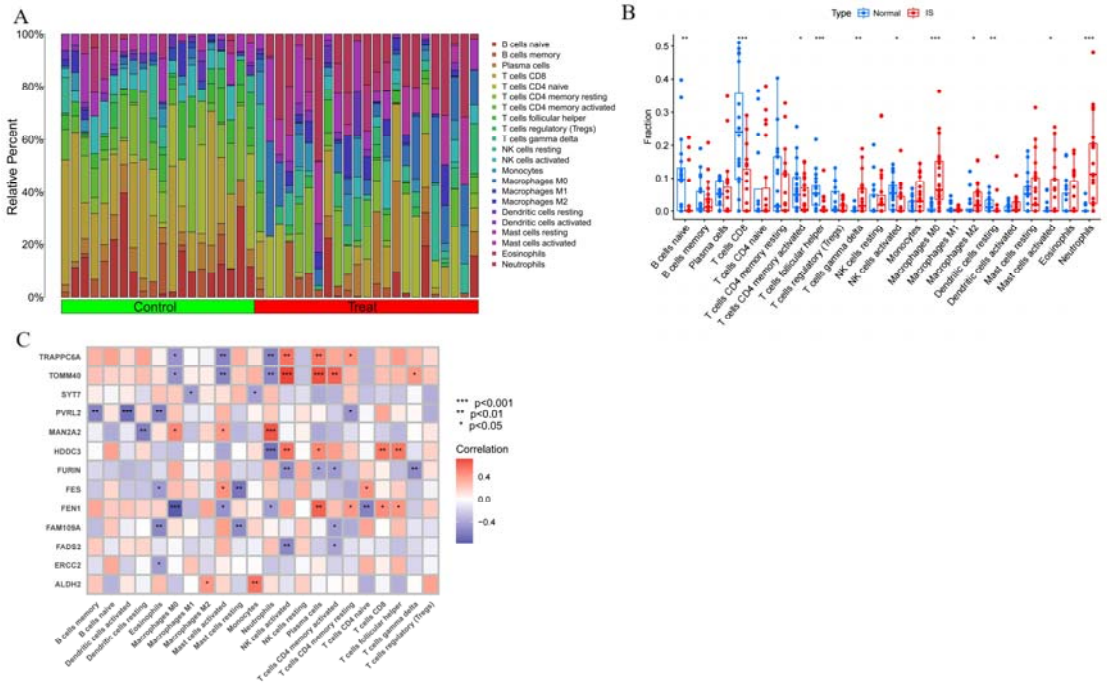


FIGURE 4 | (A) Box plots of residual of the four machine learning models; (B) Reverse cumulative distribution of residual of the four machine learning models; (C) ROC of the four machine learning models; (D) Bar plot of feature importance of the four machine learning models; (E) Nomogram of the feature genes; (F) Calibration curve of feature genes nomogram; (G) Decision curve of feature genes nomogram; (H) ROC of the test GEO dataset

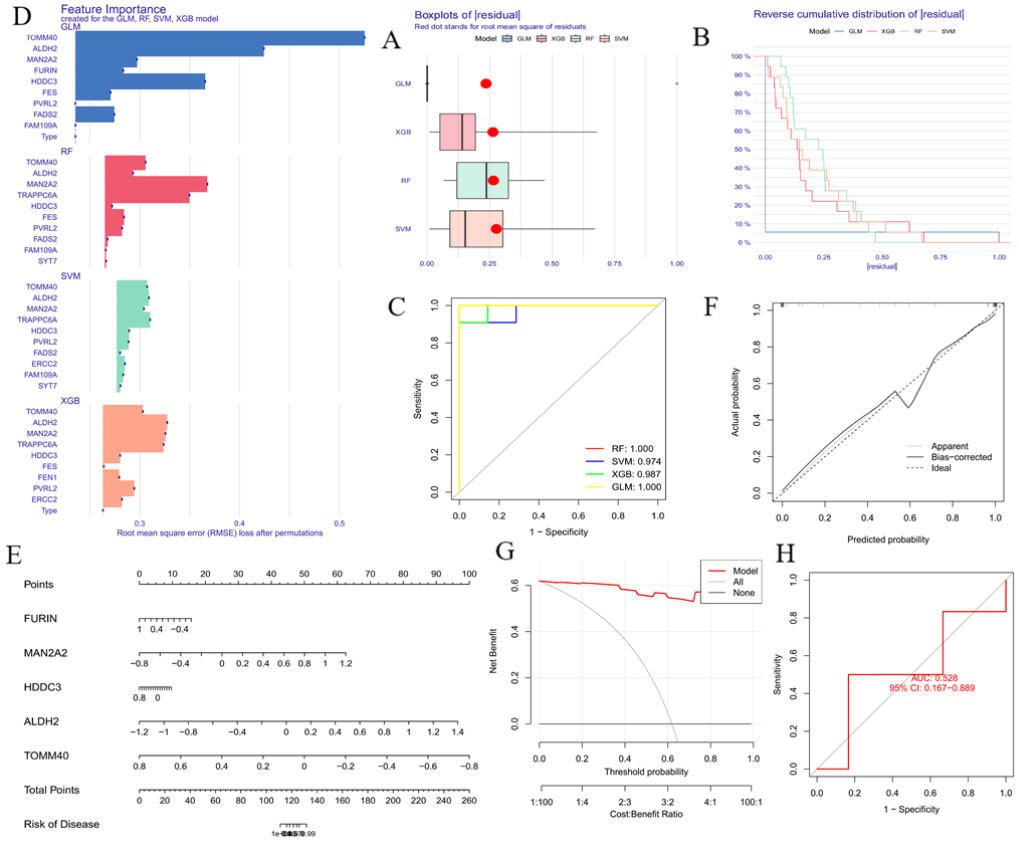


FIGURE 5 | (A) Consensus cumulative distribution plot of DEAG clustering for samples; (B) Consensus CDFs of DEAG clustering; (C) Consensus matrix heat map of DEAG clustering for samples; (D) Heat map of DEAGs expression between DEAG clusters; (E) Box plot of expression difference analysis of DEAG clusters; (F) Scatter plot of PCA between DEAG clusters; (G) Box plot of immune cell fraction between DEAG clusters; (H) Bar plot of relative percentage of each immune cells in samples of DEAG clusters; (I) Bar plot of GO terms of GSVA between DEAG clusters; (J) Bar plot of KEGG terms of GSVA between DEAG clusters

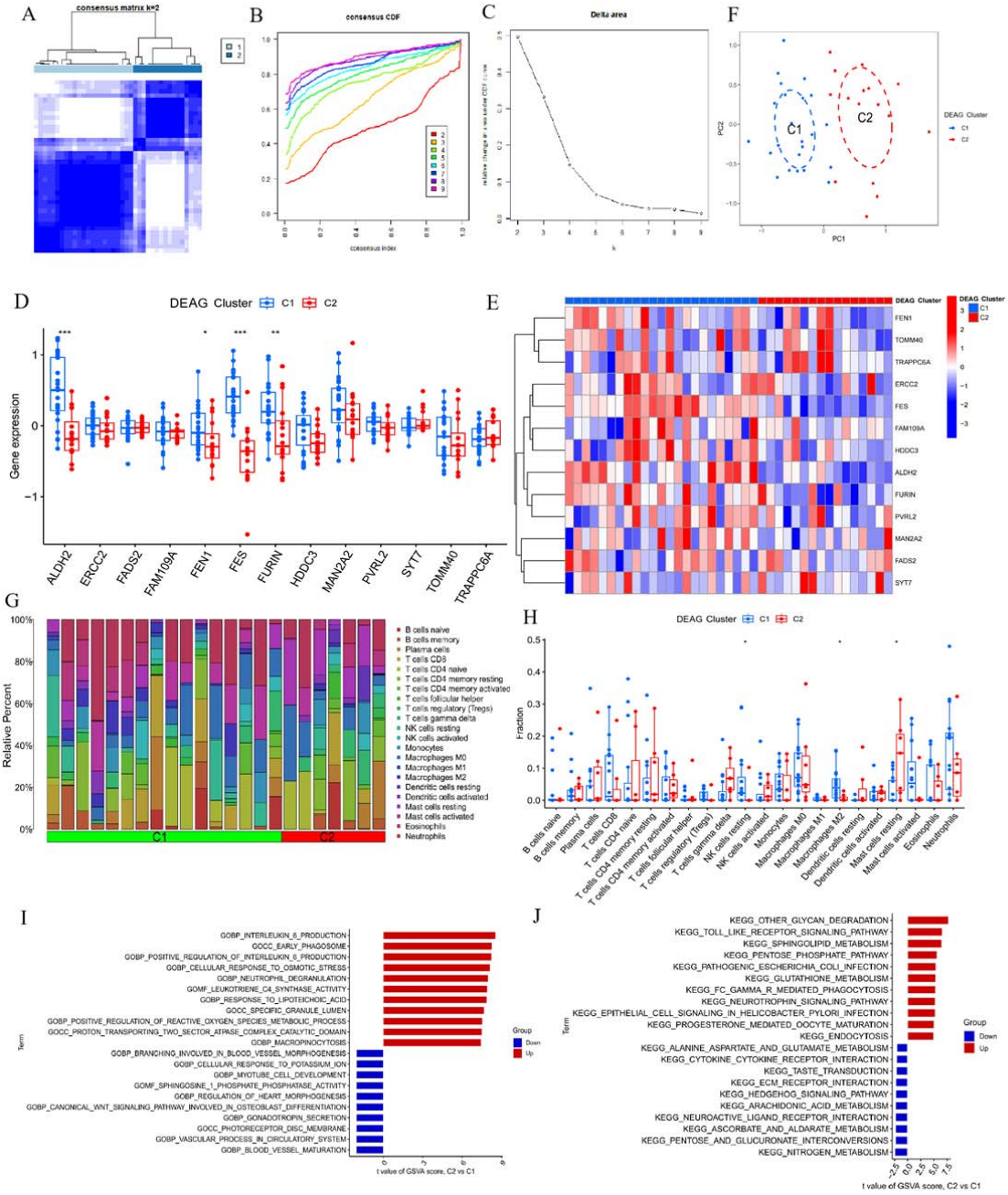


FIGURE 6 | (A) Venn plot of DEGs; (B) Consensus matrix heat map of DEG clustering for IS samples; (C) Heat map of DEGs expression between DEG clusters; (D) Box plot of expression difference analysis of DEG clusters; (E) Box plot of immune cell infiltration between DEG clusters

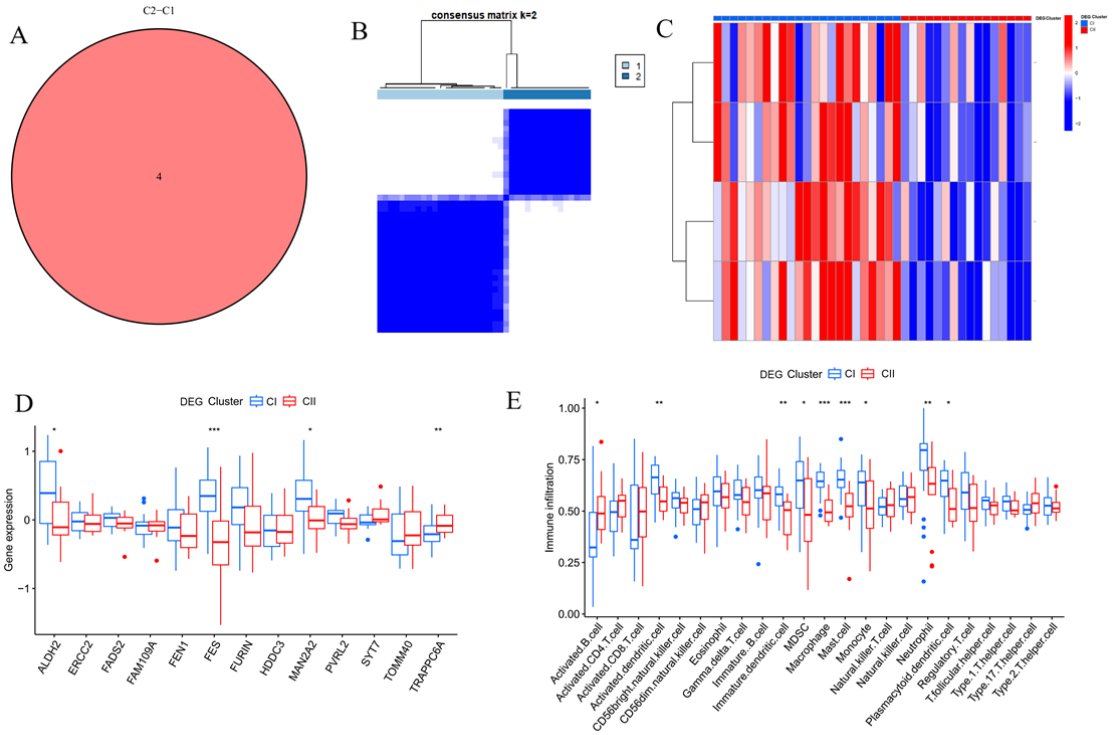
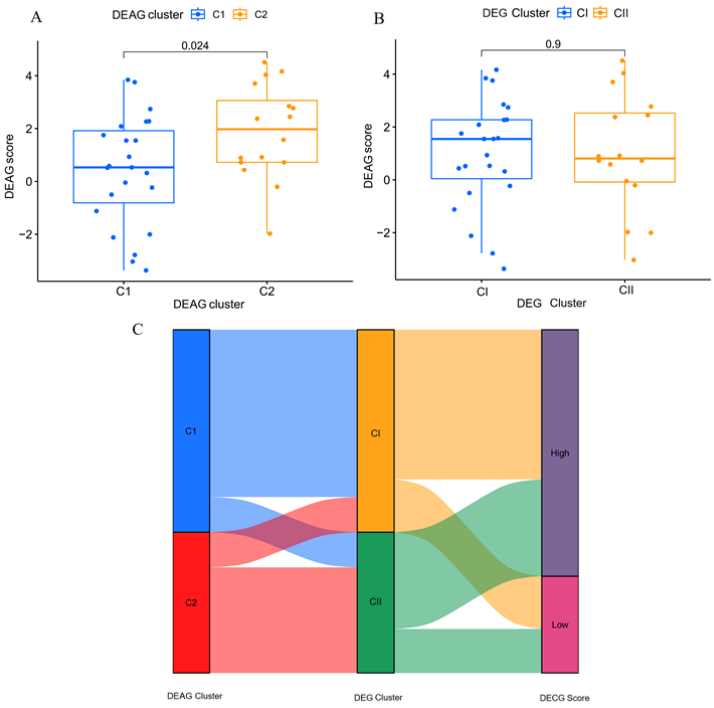


FIGURE 7 | (A) Box plot of different expression analysis of DEAG score between DEAG clusters; (B) Box plot of different expression analysis of DEAG score between DEG clusters; (C) Alluvial plot of the correspondence of the different clustered samples



Exploring Genetic Associations of Three Types of Risk Factors with Ischemic Stroke: An Integrated Bioinformatics Study

Yi Liu[1#], PhD; Weili Wang[1#], MMedSci; Xin Cui[1*], PhD; Yanming Xie[1*], Prof.

[1]Institute of Basic Research in Clinical Medicine, China Academy of Chinese Medical Sciences Beijing, 100700, China;

[#]Yi Liu and Weili Wang contributed equally to this work.

[*]Yanming Xie and Xin Cui are the joint corresponding authors of this article.

Correspondence:

Yanming Xie No. 16, Nanxiaojie, Dongzhimennei, Dongcheng District, Beijing, China

E-mail: ketizu2018@163.com

Xin Cui No. 16, Nanxiaojie, Dongzhimennei, Dongcheng District, Beijing, China

E-mail: Xinroberttcm@hotmail.com

Short Title: Associations of exposures and ischemic stroke

The total number of words: 8817 words

Abstract：

Background: Ischemic stroke (IS) is a primary cause of disability and mortality globally. More and more reports suggest a strong association between blood pressure, blood glucose, and blood lipids and their metabolic products with IS.

Methods: We extracted the genetic tools of blood pressure, blood glucose, and blood lipids and their metabolites as instrumental variables, which were then paired with GWAS data on IS and a Mendelian randomization (MR) analysis was performed to assess the effect of these exposures on the disease. Following the positive results, colocalization analysis was performed to identify shared genes associated with exposures and IS. We then performed differential expression analysis using the GEO dataset to identify the differentially expressed associated genes (DEAGs) from associated shared genes. Additional analyzes were performed on these DEAGs to obtain their importance scores using four machine learning models. A nomogram was created using genes with high importance scores to predict the level of risk assessment between DEAGs and IS.

Results: There is a positive correlation between blood pressure, blood glucose and the risk of IS onset, while blood lipids and their metabolic products are positively or negatively correlated with the risk. There are 64 shared genes of blood pressure, blood lipids and their metabolic products with IS. Thirteen DEAGs were obtained, and among which FURIN, MAN2A2, HDDC3, ALDH2, and TOMM40 were identified as feature genes for creating the nomogram which can quantitatively predict the risk of IS onset with the expression of these feature genes. By cluster analysis, we found that DEAGs expression underlying immune inflammation, angiogenesis and development, lipid metabolism, etc.

Conclusion: This study suggests a significant association between blood pressure, blood glucose, and blood lipids and their metabolic products with IS, and predicts that these exposures mainly regulate the occurrence, development, and prognosis of IS through mechanisms such as DNA repair, DNA methylation, mitochondrial repair, apoptosis, autophagy, etc.

Key words: Mendelian Randomization; Colocalization analysis; Bioinformatics; Ischemic stroke; Blood pressure; Blood glocose; blood lipids and their metabolic productes;

Abbreviations: Area Under the Curve, AUC; Deoxyribonucleic acid, DNA; Differentially expressed associated genes, DEAGs; differentially expressed genes, DEGs; Extreme Gradient Boosting model, XGB; false discovery rate, FDR; Gene Expression Omnibus , GEO; Gene ontology, GO; gene set variation analysis, GSVA; Generalized Linear, GL; genome-wide association study, GWAS; high-density lipoprotein, HDL; Instrumental variables, IVs; Integrated Epidemiology Unit, IEU; intermediate-density lipoprotein, IDL; inverse variance weighted, IVW; Ischemic stroke, IS; Kyoto Encyclopedia of Genes and Genomes, KEGG; low-density lipoprotein, LDL; Mendelian randomization, MR; Natural killer, NK; one-way analysis of variance, ANOVA; posterior probability, PP; Principal Component Analysis, PCA; Random Forest model, RF; receiver operating characteristic curves, ROC; Single nucleotide polymorphisms, SNPs; Single-sample gene set enrichment analysis, ssGSEA; Support Vector Machine model, SVM; very low-density lipoprotein, VLDL

Ischemic stroke (IS) is a disease caused by the blockage of blood vessels in the brain, leading to local cerebral hypoxia and ischemia, resulting in the death of brain cells. This blockage is usually caused by thrombosis or embolism, causing neurological deficits in the brain area with insufficient blood supply.[1-3] In 2019, the number of patients who died from stroke worldwide reached 6.55 million, making it the main cause of adult disability and death worldwide.[4,5] Among newly diagnosed stroke cases, IS accounted for 62.4%.[6] Therefore, identifying the risk factors for IS is crucial for preventing this disease. Blood pressure, blood glucose, and blood lipids and their metabolic products have received widespread attention as potential factors affecting the risk of IS. A large number of clinical and epidemiological studies have shown that controlling blood pressure, blood glucose, and blood lipids can reduce the risk of IS.[7-9] In recent years, with a deeper understanding of the human genome, we have come to realize that the expression of these biomarkers is largely influenced by genetic information.[10] Therefore, exploring high-risk groups from a genetic perspective and identifying shared genes related to blood pressure, blood glucose, and blood lipids and their metabolic products with IS is of great significance for the precise prevention and treatment of IS.

Mendelian randomization (MR) is a genetic epidemiological research tool that is widely used to assess the potential causal relationship between exposure and disease.[11,12] The working mechanism of MR analysis is similar to that of a natural randomized controlled trial. It uses genetic instruments (single nucleotide polymorphisms, SNPs) that are strongly correlated with exposure factors and are not affected by confounding factors under the principle of MR distribution as instrumental variables (IVs).[13,14] This study used a

two-sample MR analysis to explore the potential causal effects of blood pressure, blood glucose, and blood lipids and their metabolic products on IS. The selected datasets are all from the Integrated Epidemiology Unit (IEU) database (https://gwas.mrcieu.ac.uk/). This database is public and contains nearly 2.45 trillion genetic associations from 42,334 summary datasets of genome-wide association study (GWAS).

Colocalization analysis is a method for assessing whether two correlated traits are driven by the same genetic mechanism. This method often uses the Bayesian model coloc.[15] The coloc model assumes that in each tested region, each trait has at most one association point, and calculates the posterior probability (PP) of all possible association patterns of the two traits through approximate Bayesian factors: 1) H0: no association; 2) H1: only related to trait 1; 3) H2: only related to trait 2; 4) H3: related to both trait 1 and trait 2, but through two independent SNPs; 5) H4: related to both trait 1 and trait 2, through a shared SNP. The posterior probabilities of each association pattern are denoted as PP0 to PP4.[15] A higher PP4 value (e.g., PP4 > 50%) provides colocalization support, indicating the existence of shared genetic variation between the two traits.[16] This study used colocalization analysis to find shared genes associated with blood pressure, blood glucose, blood lipids and their metabolic products, and IS. These shared genes may provide important clues for revealing the pathogenesis of IS.

The Gene Expression Omnibus (GEO) database collects genomic datasets from different species and tissues, which can be used in conjunction with bioinformatics methods to discover evidence of specific gene expression, identify disease prediction factors, and ultimately help us understand the mechanisms by which the genome regulates physiological

and pathological states.[17-19] Therefore, based on the associated shared genes of blood pressure, blood glucose, blood lipids and their metabolic products with IS, this study performed differential analysis on normal samples and IS samples in the GEO dataset, and constructed a machine learning model to screen out feature genes among the associated shared genes. On the one hand, it verifies the MR results through real human samples, and on the other hand, it further screens out biomarkers with higher importance, to further explore the mechanisms of blood pressure, blood glucose, blood lipids and their metabolic products on IS.

1. Study design

In this study, aligned with our research objective, we selected GWAS summary datasets for blood pressure, blood glucose, and blood lipids and their metabolic products with IS. We conducted a two-sample MR analysis and chose positive results for further gene colocalization analysis to identify genes associated with the exposure factors and IS. Subsequently, we retrieved two datasets of IS (training and validation sets), encompassing normal and IS groups, from the GEO database (https://www.ncbi.nlm.nih.gov/geo/). Differential analysis, correlation analysis, and immune infiltration analysis were carried out on the normal and IS groups of the training set. With the help of machine learning, we derived feature genes related to blood pressure, blood glucose, blood lipids and their metabolic products, and IS. Following this, the results of the machine learning model were validated using the validation set, and two rounds of clustering analysis on the IS group samples were performed to further investigate the functional mechanisms of the exposure-related genes (Figure 1).

1.1 Data sources

In accordance with the goals of this study, we utilized the IEU database, conducting searches with the keywords of "VLDL (very low-density lipoprotein)", "LDL (low-density lipoprotein)", "IDL (intermediate-density lipoprotein)", "HDL (high-density lipoprotein)", "apoliprotein", "triglyceride", "fatty acid class", "systolic pressure", "diastolic pressure", "blood glucose", "glycosylated hemoglobin", and "IS". This can identify genetic instruments for blood pressure, blood glucose, and blood lipids and their metabolic products as exposure variables. Concurrently, IS-related genetic instruments were selected as outcome variables for an MR analysis, aiming to explore potential causal relationships between exposures and outcomes. By using "ischemic stroke" as the keyword, setting the data type as array-based expression profiles, and specifying the species as homo sapiens, we retrieved samples from the GEO database (https://www.ncbi.nlm.nih.gov/geo/) to obtain gene expression and clinical data of IS patients and healthy individuals. Perl code was employed for gene symbol annotation and data adjustment, thus deriving expression levels of genes related to blood pressure, blood glucose, blood lipids and their metabolites, and IS, in both the normal and IS groups.

1.2 Mendelian Randomization analysis

We utilized the "TwoSampleMR" package to conduct a two-sample MR analysis on the relationship between blood pressure, blood glucose, and blood lipids and their metabolic products with IS.[20] In the analysis, we set the parameters as follows: 1) The selected genetic

instruments need to have a strong correlation with the exposure factors, and the corresponding P-value should be less than 5×10^{-8}; 2) In conducting Linkage Disequilibrium (LD) clumping, we set the r^2 threshold at 0.01; 3) When performing clumping analysis, we set the window size as 10,000 kilobase pairs.[21] These stringent parameter settings help us to accurately reveal the causal relationship between exposures and outcomes.

1.3 Colocalization analysis

After completing the two-sample MR analysis, we selected positive results that demonstrated significant associations between the exposure factors and outcome in the MR analysis. We used the "ieugwasr" and "coloc" packages to conduct colocalization analysis on these positive results to determine if the exposure factors and outcome might be associated within the same gene region. We set a posterior probability (PP4) greater than 50% as the standard, indicating that the genes in that region are associated with both exposure and outcomes[16]. Finally, we employed the "gassocplot2" package to visualize the colocalization results, clearly demonstrating the associated shared genes of blood pressure, blood glucose, and blood lipids and their metabolic products with IS.

1.4 Identification of differently expressed associated genes (DEAGs) and analyses of DEAGs

Expression levels of genes associated with blood pressure, blood glucose, blood lipids and their metabolic products, and IS were extracted from both the normal and IS groups. Differential expression analysis was performed using the "limma", "pheatmap", "ggpubr" and other R packages, and results were presented as box plots and heatmaps. Genes with a p-value less than 0.05 were defined as differentially expressed associated genes (DEAGs).

Using Perl coding, DEAGs were located on chromosomes and their positions were displayed on a circos plot created with the "Rcircos" package. In addition, the "cor" command was used to calculate the correlation coefficient between each two DEAGs and the results were visualized to illustrate the correlation among DEAGs.

1.5 Analysis of immune cells in IS samples

We performed 1000 simulations using the CIBERSORT command in R to obtain a total relative amount of immune cells equal to 1, then visualized the content of immune cells in each sample using a bar plot. Single-sample gene set enrichment analysis (ssGSEA) was performed using the "GSVA" and "GSABase" packages to compare the differences in immune cell contents between the normal group and the IS group, and the results of ssGSEA are presented as box plot. We matched the differentially expressed associated genes (DEAGs) with the ssGSEA scores, performed a correlation test to obtain the correlation coefficient, and then visualized the results as a heat map.

1.6 Identification of feature genes and construction and validation of nomogram based on ML

Expression data from DEAGs was utilized to construct four predictive models: Random Forest model (RF), Support Vector Machine model (SVM), Generalized Linear Model (GL), and Extreme Gradient Boosting model (XGB). The results from these four models were calculated based on the prediction function. Feature genes within DEAGs were then selected through comprehensive analysis of reverse cumulative distribution plots, residual box plots, and receiver operating characteristic (ROC) curves. After selecting the optimal model, line plots were constructed using the expression levels of the feature genes in both the normal and

IS groups. Lastly, decision and calibration curves were drawn to assess the accuracy of the line plots. Another dataset that includes both a normal and IS group was obtained from the GEO database, and a machine learning model was constructed using the same R language approach as before. The ROC was plotted to validate the machine learning model constructed in the test dataset.

1.7 Clustering of DEAGs and analysis between DEAG clusters.

The "ConsensusClusterPlus" package in R was used for clustering, employing Euclidean distance type and allowing up to nine clusters. Expression levels between clusters were compared using heatmaps and box plots, and principal component analysis (PCA) was applied to assess differences between clusters. Following this, a ssGSEA was performed on the DEAG clusters, generating bar plots of the amount of individual immune cells in each sample within the different clusters and comparing the differences in immune cell content among different clusters. Gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analyses were performed using gmt files downloaded from the GSEA platform (http://www.gsea-msigdb.org/), and gene set variation analysis (GSVA) was conducted in R language to analyze the expression of enrichment items between clusters. Finally, under the filtering conditions of $|logFC| > 1$ and an adjusted P-value < 0.05, differential expression analysis was performed on the gene expression of DEAG clusters. Intersection of DEAG clusters through a Venn diagram yielded differentially expressed genes (DEGs).

1.8 Clustering of DEGs and analysis between DEG clusters

The same clustering method as in section 1.7 was applied to cluster DEGs, and the DEG cluster with the highest precision was selected. Based on the DEGs clustering results, we compared the expression levels of DEGs, the differences in DEAGs expression, and the immune cell content in different clusters. These results were visualized using heatmaps and box plots.

1.9 Construction of DEAG scores

The PCA method was employed to calculate the expression levels of DEAGs for each sample, yielding DEAG scores (The formula is shown below).[22] Using R packages such as "limma" and "ggpubr", we performed differential analysis on the scores of DEAGs in both significantly differentially expressed core gene clusters and DEG clusters. We created box plots to illustrate the scores of significantly differentially expressed core genes in samples within the significantly differentially expressed core gene and DEG clusters. In addition, an alluvial diagram was drawn using the package "ggalluvial" to visualize the relationships and overall processes among DEAG clusters, DEG clusters, samples with higher scores, and samples with lower scores of DEAGs.

$$\text{DEAG Score} = \sum (\text{PC1}_i + \text{PC2}_i)$$

1.10 Statistical analysis

This study carried out MR analysis and colocalization analysis using R V4.2.1. In the MR analysis, if the number of SNPs for the exposure factor is less than 5 after filtering, we will use the Wald ratio method for analysis; otherwise, we will opt for the inverse variance

weighted (IVW) method for analysis. In colocalization analysis, to maximize the acquisition of shared genes between "exposure" and "outcome", we set the chromosomal locus parameter range from 75 kilobase pairs to 500 kilobase pairs, aiming to include a wider area around the SNP to enhance the likelihood of discovering shared genes related to both exposure and outcome. In the bioinformatics part, we used Strawberry Perl 5.32.1.1 for GEO dataset extraction and data annotation, and all other statistical analyses were performed with R V4.2.1. For two independent samples, we applied a t-test, while for two paired samples we used the Wilcoxon paired-rank sum test, and for three or more groups of data, we employed one-way analysis of variance (ANOVA) and the Kruskal-Wallis rank sum test. The Spearman rank correlation test was used for correlation analysis. We set a P value < 0.05 or a false discovery rate (FDR) < 0.05 after correction by the Benjamini-Hochberg method as being statistically significant.

2 Results

2.1 Results of collection of GWAS and GEO datasets

We retrieved 516 relevant exposure datasets from the ieu database based on the set keywords, and the IS dataset came from 440,328 participants of European ancestry, including 34,217 cases and 406,111 controls. These datasets were used in two-sample MR analysis for IS to explore their potential causal relationships. By searching the GEO database for IS datasets, we filtered out two datasets that met the requirements, namely GSE16561 and GSE22255, using the former as the test set and the latter as the validation set. The former is

total RNA data from human peripheral blood samples, including 39 IS samples (diagnosed by MRI) and 24 normal samples; the latter is RNA expression data from human peripheral blood monocytes, including 20 IS samples and 20 normal samples.

2.2 Results of Mendelian randomization analysis

Among the 516 exposure datasets, 114 showed a positive association with IS. These positive results further suggested that increased levels of Apolipoprotein A-I and HDL may reduce the risk of IS, while elevated levels of VLDL, LDL, IDL, Apolipoprotein B, triglycerides, fatty acids, systolic blood pressure, diastolic blood pressure, fasting blood sugar, and glycated hemoglobin might increase the risk of IS. (Please see Attachment 1 for detailed analysis results)

2.3 Results of colocalization analysis

Upon completion of the two-sample MR analysis, we further subjected the 114 positive results to colocalization analysis to study the potential shared genes between the exposure factors and IS. The results showed that the PP4 was greater than 50% for 70 of the positive results, including exposure factors such as LDL, IDL, VLDL, Apolipoprotein B, fatty acids, systolic and diastolic blood pressure. This suggests that these exposure factors may have genes shared with IS. The colocalization results were visualized using the "gassocplot2" package, and we finally identified 64 shared genes between these exposure factors and IS. These genes may play a key role in the association between the selected exposure factors and

IS. (Please see Attachment 1 for detailed analysis results and Attachment 2 for specific shared genes situations)

2.4 Results of DEAGs identification and analysis of DEAGs.

Through colocalization analysis, we identified 64 associatied shared genes between blood pressure, blood sugar,and blood lipids and their metabolites with IS. Differential analysis between the IS group and the normal group in the test dataset revealed that 13 associatied shared genes exhibited statistically significant differences. Among them, ERCC2, FEN1, HDDC3, TOMM40, TRAPPC6A were highly expressed in the normal group, while ALDH2, FADS2, FES, FURIN, MAN2A2, PVRL2, SYT7, and FAM109A were highly expressed in the IS group, as shown in Figures A and B. The specific positions of these significantly differentially expressed association genes on the chromosome are shown in Figure C. The correlation analysis between each pair of DEAGs in the IS samples showed that there is a certain correlation between DEAGs, which are mainly positive correlations, as shown in Figures D and E.

2.5 Results of Immune cell infiltration analysis, immune cell difference analysis and correlation analysis of IS samples

Through immune cell infiltration analysis, we obtained the types and content of immune cells expressed in each sample, as shown in Figure A. ssGSEA analysis revealed (Figure B) that the expression differences of 11 types of immune cells in the control group and IS group were statistically significant. Among them, B cells naive, T cells CD8, T cells CD4 memory

activated, T cells follicular helper, Natural killer (NK) cells activated, and Dendritic cells activated were highly expressed in the control group, while T cells gamma delta, Macrophages M0, Macrophages M2, Mast cells activated, and Neutrophils were highly expressed in the IS group. The correlation analysis between DEAGs and immune cells showed that there is a certain correlation between DEAGs and immune cells, and the levels of positive and negative correlations are comparable, as shown in Figure C.

2.6 Results of selection of machine learning models, construction of nomogram, and verification

The results of building SVM, RF, XGB, and GL machine learning prediction models using DEAGs data were obtained. From the residual boxplots, inverse cumulative distribution plots, and ROC curves (Figure A, C, and D), it can be observed that the GL method had the highest area under the ROC curve, the lowest residual values, and the lowest inverse cumulative values. Therefore, the GL method was considered the most accurate and chosen as the best model for further analysis. The GL model provided importance scores for the selected feature genes, as shown in Figure B, revealing nine feature genes. Among them, the top five genes with the highest importance scores were used to construct the column line graph (including FURIN, MAN2A2, HDDC3, ALDH2, and TOMM40). Then, separate scoring scales were obtained for these five feature genes. The risk rate of the co-located feature genes associated with blood pressure, blood glucose, and blood lipids and their metabolic products in the occurrence of IS was assessed by calculating the sum of the feature gene expression scores (Figure E). The predictive accuracy of the model was evaluated by the

distance between the solid line and the dotted line in the calibration curve (Figure F) and the distance between the red and gray lines in the decision curve (Figure G), indicating high accuracy. The GL model was validated using the validation dataset (GSE22255) and the top five feature genes with the highest importance scores were included in the model validation and ROC curve analysis. The results showed an Area Under the Curve (AUC) value of 0.528, with a 95% confidence interval of 0.167-0.889. Therefore, based on the validation of the model, it can be concluded that the model based on the GEO dataset has good accuracy.

2.7 Results of Clustering of DEAGs, Analysis of DEAGs Expression between Clusters, Immune Cell Analysis, and GO and KEGG Enrichment Analysis

Clustering analysis based on the expression of DEAGs revealed two distinct clusters with the highest accuracy, as shown in Figure A. Consequently, the IS samples were divided into two groups: C1 and C2, as depicted in Figures B and C. Subsequently, DEAGs expression analysis was performed between the two DEAG clusters (Figure D and E), indicating that ALDH2, FEN1, FES, and FURIN were significantly upregulated in C1. PCA analysis (Figure F) demonstrated that DEAGs can differentiate between C1 and C2. Moreover, ssGSEA analysis (Figure G) identified three statistically significant immune cell types between C1 and C2. Specifically, NK cells resting and Macrophages M2 were upregulated in C1, while Mast cells resting were upregulated in C2. The respective content of different immune cells in each C1 and C2 sample is illustrated in Figure H.

In GSVA analysis (Figures I and J), it can be observed that compared to C1, C2 showed upregulation of several GO biological processes, including Interleukin 6 production, positive

regulation of interleukin 6 production, cellular response to osmotic stress, neutrophil degeneration, response to lipoteichoic acid, positive regulation of reactive oxygen species metabolic process, and macropinocytosis. On the other hand, C2 exhibited downregulation of branching involved in blood vessel morphogenesis, cellular response to potassium ion, myotube cell development, regulation of heart morphogenesis, canonical Wnt signaling pathway involved in osteoblast differentiation, gonadotropin secretion, vascular process in the circulatory system, and blood vessel maturation in GO biological processes. In GO molecular functions, C2 showed upregulation of leukotriene C4 synthase activity and downregulation of sphingosine 1-phosphate phosphatase activity. In GO cellular components, C2 exhibited upregulation of early phagosome, specific granule lumen, and proton-transporting two-sector ATPase complex catalytic domain, while downregulation of photoreceptor disc membrane. In KEGG pathways, C2 demonstrated upregulation of other glycan degradation, toll-like receptor signaling pathway, sphingolipid metabolism, pentose phosphate pathway, pathogenic Escherichia coli infection, glutathione metabolism, Fc gamma R-mediated phagocytosis, neurotrophin signaling pathway, cell signaling in Helicobacter pylori infection, progesterone-mediated oocyte maturation, and endocytosis. Conversely, C2 showed downregulation of alanine, aspartate, and glutamate metabolism, cytokine-cytokine receptor interaction, taste transduction, ECM receptor interaction, hedgehog signaling pathway, arachidonic acid metabolism, neuroactive ligand-receptor interaction, ascorbate and aldarate metabolism, pentose and glucuronate interconversions, and nitrogen metabolism in KEGG pathways.

2.8 Results of DEGs screening, DEGs clustering, and analyzes of DEG clusters

After filtering for significantly differentially expressed genes (DEGs) between C1 and C2 samples, a total of 4 DEGs were identified (Figure A). These 4 DEGs were further subjected to clustering analysis within the DEAGs cluster, resulting in two distinct clusters, namely CI and CII (Figure B). Analysis of DEAGs expression in these two clusters in IS samples revealed that RNASE2, RNASE3, MMP9, and CAMP exhibited statistically significant differential expression, with higher expression in CI and lower expression in CII (Figure C).

The differential expression analysis of DEAGs based on DEGs clustering was performed on the samples, as shown in Figure D. The results revealed that these genes were mainly upregulated in the CI group and downregulated in the CII group. Among them, the significantly differentially expressed genes included ALDH2, FES, MAN2A2 (upregulated in CI), and TRAPPCA (upregulated in CII). Additionally, using ssGSEA based on DEGs clustering (using the same gene set file as above), statistically significant differences in immune cells were observed, including activated B cells (downregulated in CI), activated dendritic cells, immature dendritic cells, myeloid-derived suppressor cells (MDSCs), macrophages, mast cells, monocytes, neutrophils, and plasmacytoid dendritic cells (upregulated in CI), as shown in Figure E.

2.9 Results of DEAGs scoring, differential analysis of DEAGs score between clusters, and construction of the alluvial plot

The differential analysis of DEGs clustering based on the principal component analysis

(PCA) scores showed statistically significant differences between the clusters (Figure A). Specifically, C1 had lower scores, while C2 had higher scores. However, there was no statistically significant difference observed in the DEGs clustering (Figure B). The scatter plot in Figure C displayed that the C1 and C2 clusters of DEAGs corresponded mainly to the CI and CII clusters of DEGs, respectively. However, there was no clear correspondence observed between the high and low scores of DEAGs and the DEGs clustering.

3 Discussion

3.1. The relationships of blood pressure, blood glucose, and blood lipids and their metabolic products with IS

Clinical studies have demonstrated that effective management of various physiological indicators can significantly impact the risk of IS. For example, blood pressure control has been shown to significantly reduce the risk of IS recurrence.[23] Elevated fasting blood glucose levels are significantly associated with an increased risk of IS.[24] Furthermore, increased levels of glycated hemoglobin, regardless of diabetes status, are associated with an increased risk of IS.[25] Lipid management also plays a significant role in IS risk. There are significant differences in VLDL levels between the control and IS groups.[26] However, the role and mechanisms of VLDL level changes in IS pathogenesis require further investigation. Medications that lower LDL levels can significantly reduce the risk of IS.[27] Increased levels of IDL and apolipoprotein B are associated with an increased risk of IS.[28,29] Controlling triglyceride levels can lower the risk of IS.[30] Both polyunsaturated fatty acids and saturated

fatty acids are also associated with IS.[31,32] However, elevated levels of HDL and apolipoprotein A-I are associated with a lower risk of IS.[33,34] Taking all the above clinical studies into consideration, it can be concluded that blood pressure, blood glucose, and blood lipids and their metabolic products play critical roles in influencing the risk of IS.

The results of MR analysis indicate a potential causal relationship between blood pressure (diastolic and systolic), blood glucose (fasting blood glucose and glycated hemoglobin), blood lipids and their various metabolic products (VLDL, LDL, IDL, HDL, apolipoprotein A-I, apolipoprotein B, triglycerides, Eicosapentaenoate, Docosapentaenoate, Stearidonate, Docosahexaenoic acid) with IS. Increasing levels of high-density lipoprotein (HDL) and apolipoprotein A-I are associated with a decreased risk of IS. However, elevated blood pressure (including systolic and diastolic), blood glucose (fasting blood glucose and glycated hemoglobin), various low-density lipoproteins such as VLDL, LDL, IDL, as well as apolipoprotein B, triglycerides, Eicosapentaenoate, Docosapentaenoate, Stearidonate, Docosahexaenoic acid are associated with an increased risk of IS.

3.2 Discussion of mechanisms associated shared genes

Through MR analysis, we found that blood pressure, blood glucose, blood lipids and their metabolic products have a significant impact on the risk of IS. To further explore the specific mechanisms underlying these effects, we conducted a colocalization analysis. The results showed that there are 64 associated shared genes between LDL, IDL, VLDL, apolipoprotein B, fatty acids, systolic blood pressure, and diastolic blood pressure with IS. Although a positive association between blood glucose and IS risk has been established, we

did not identify any associated shared genes due to the PP4 being less than 50%.

The differential analysis of the 64 associated shared genes between normal and IS samples, using the GEO dataset, revealed 13 genes with significant differential expression. Among them, ERCC2, a transcription factor (TF) involved in nucleotide excision repair of damaged DNA,[35] has been clinically associated with increased stroke risk,[36,37] and animal studies have shown its neuroprotective role in preventing ischemia-reperfusion injury.[38] FEN1, another TF encoding a protein involved in rDNA and mitochondrial DNA repair,[39] has been implicated in the pathogenesis of IS and is considered a therapeutic target for IS treatment.[40,41] HDDC3, a protein encoded by HDDC3 gene, participates in the starvation response, and starvation-induced autophagy is known to protect neurons and regulate IS.[42,43] TOMM40, encoding the translocase of the outer mitochondrial membrane 40 homolog, has been identified as a potential susceptibility locus for IS based on a study conducted in Japan.[44] ALDH2, encoding aldehyde dehydrogenase, has shown neuroprotective effects by clearing 4-hydroxy-2-nonenal and reducing mitochondrial-associated cell apoptosis through JNK-mediated cystathionine-β-synthase-3 activation, making it a potential target for IS intervention. Case-control studies have also suggested an association between ALDH2 polymorphisms and IS risk in the Han Chinese population.[45-47] FADS2, encoding fatty acid desaturase 2, has been associated with IS risk and lipid levels in case-control studies, although the specific mechanisms are still under investigation.[48] The product of the FES gene exhibits tyrosine-specific protein kinase activity. Inhibitors of this class of enzymes have shown potential therapeutic value for various diseases, including IS, but the involvement of FES in IS is yet to be confirmed.[49] FURIN, encoding proprotein convertase subtilisin/kexin

type 1, has been found to be upregulated within 24 hours after cerebral ischemia in animal experiments,[50] and case-control studies have revealed lower DNA methylation levels in the IS region, suggesting its association with IS.[51] SYT7, encoding synaptotagmin 7, has been identified to have significant DNA methylation correlation with IS in a large-scale sequencing study.[52] The roles of FAM109A, MAN2A2, PVRL2, and others in IS require further investigation.

These genes primarily function through DNA repair, DNA methylation, mitochondrial repair, cell apoptosis, and autophagy to regulate lipid levels, blood pressure, and neuronal protection, thereby affecting the occurrence, development, and prognosis of IS. The associated shared genes are predominantly upregulated in the IS group, and they exhibit mostly positive regulatory relationships, indicating potential synergistic effects among these genes in IS. Differential expression analysis of immune cells between the IS and normal groups reveals significant expression differences in 50% of the immune cells, suggesting the involvement of immune cells in the regulation of IS by these associated shared genes. Moreover, the comparable levels of positive and negative regulation of immune cells by DEAGs suggest that DEAGs may have bidirectional regulatory effects on the immune system's role in IS.

The feature genes obtained from the GL model constructed through machine learning (FURIN, MAN2A2, HDDC3, ALDH2, TOMM40) have significant importance in regulating the associated shared genes and the risk of developing IS. The column plot provides a quantitative prediction of the aforementioned importance and risk of disease. Furthermore,

the construction of the GL model using the validation dataset indicates that the column plot has high accuracy.

The IS samples were clustered into C1 and C2 groups based on the expression of DEAGs, with C1 showing predominantly high expression of DEAGs and C2 showing low expression. The two clusters were significantly distinguishable according to the PCA results. Furthermore, immune cell infiltration analysis of DEAGs clustering revealed significant differences in immune cell expression between C1 and C2 groups of IS patients. The GSVA results indicated significant expression differences between C1 and C2 groups in various biological processes and pathways, including immune inflammation, vascular development and formation, reactive oxygen species metabolism, carbohydrate metabolism, lipid metabolism, fatty acid metabolism, amino acid metabolism, and neural regulation.

Differential analysis between the C1 and C2 groups identified 4 DEGs, which have been extensively studied and shown to be associated with IS incidence, IS prognosis, and IS complications.[53-55] Subsequently, IS samples were further clustered into CI and CII groups based on the expression of DEGs. Differential analysis revealed that DEGs were highly expressed in the CI group and lowly expressed in the CII group, while DEAGs were mainly upregulated in the CI group and downregulated in the CII group. Immune cell infiltration analysis of DEGs clustering demonstrated significant differences in immune cell expression between the CI and CII groups of IS patients, with upregulation observed in the CI group and downregulation observed in the CII group. PCA-based DEAG scoring and subsequent differential analysis revealed significant differences in DEAG scoring between the C1 and C2 groups, while no significant differences were observed between the CI and CII groups.

Furthermore, the construction of Sankey diagrams provided a visual representation of the correspondence between the two clustering results and the high/low DEAGs\ scoring groups, with C1 and C2 corresponding to CI, CII, and the high/low DEAG groups, while the correspondence between CI and CII was less clear.

3.3 Limitations of this study

Although MR analysis and colocalization analysis provide valuable insights into the risk factors and underlying mechanisms of IS from the perspective of exposure factors and genes, these methods also have their limitations. The pathogenesis of IS is extremely complex and involves intricate interactions among multiple environmental factors, physiological indicators, and genetic factors. Therefore, in order to comprehensively understand and elucidate the risk factors and mechanisms of IS, a holistic and multifactorial research approach is needed. Additionally, the GWAS data included in our study primarily originate from sample populations of European ancestry. Hence, extrapolating these results directly to non-European populations may have limitations. Similarly, bioinformatics methods also have certain limitations. Although they have provided multi-faceted and multi-dimensional analyses of human blood samples, elucidating specific mechanisms of interaction between blood pressure, blood glucose, blood lipids and their metabolic products in relation to IS, further research and clinical applications require careful consideration. Moreover, the blood samples used in this study were only collected from the United States and Portugal, and whether the results can be extrapolated to other countries, ethnicities, and populations remains to be investigated and confirmed. Given the relatively small sample size, there is a greater risk of bias in the conclusions. Therefore, large-scale studies or animal experiments can be conducted to further

validate and deepen the conclusions.

4 Conclusion and prospects

This study first integrated MR analysis, colocalization analysis, and bioinformatics analysis to conduct an in-depth study of the association of blood pressure, blood glucose, and blood lipids and their metabolic products with IS. Firstly, the results of MR analysis showed that there is a positive correlation between blood pressure, blood glucose and the risk of IS onset, and blood lipids and their metabolic products have either a positive or negative correlation with the risk of IS onset.Furthermore, through colocalization analysis, we identified 64 associated shared genes with blood pressure, blood lipids and their metabolic products in relation to IS. Additionally, using the GEO dataset, we identified 13 DEAGs. These genes primarily function through deoxyribonucleic acid (DNA) repair, DNA methylation, mitochondrial repair, cell apoptosis, and autophagy, regulating lipid metabolism, blood pressure, and neuroprotection, thereby influencing the occurrence, progression, and prognosis of IS. Subsequently, through the construction of machine learning models, we selected feature genes. Cluster analysis revealed that the expression of DEAGs in IS may also involve mechanisms such as immune inflammation, vascular development, lipid metabolism, and fatty acid metabolism.

The findings of this study will contribute to a deeper understanding of the pathophysiological mechanisms underlying IS, optimize treatment strategies, and drive the development of new drugs. The methodological approach used in this study, which combines genetic epidemiology with bioinformatics, offers a unique perspective for elucidating causal

relationships between exposure factors and diseases, as well as unraveling genetic mechanisms. This research strategy, along with the identification of the associated shared genes, provides a new theoretical framework and practical approach for the prevention, diagnosis, and treatment of IS.

Article Information

Acknowledgments

The authors would like to thanks the institutions for providing GWAS summary statistics and human gene expression datasets.

Source of Founding

National Administration of Traditional Chinese Medicine 2021 Qihuang Scholar Support Project (National TCM Education Letter 2022 No. 6) ; Yanming Xie's National Renowned Traditional Chinese Medicine Expert Inheritance Studio Construction Project (National TCM Education Letter 2022 No. 75).

Disclosures

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References:

1. Hankey GJ. Stroke. Lancet. 2017;389:641-654. doi: 10.1016/S0140-6736(16)30962-X.

2. Lee RHC, Lee MHH, Wu CYC, Couto E Silva A, Possoit HE, Hsieh TH, Minagar A, Lin HW. Cerebral ischemia and neuroregeneration. Neural Regen Res. 2018;13:373-385. doi: 10.4103/1673-5374.228711.

3. Powers WJ, Rabinstein AA, Ackerson T, Adeoye OM, Bambakidis NC, Becker K, Biller J, Brown M, Demaerschalk BM, Hoh B, et al. Guidelines for the Early Management of Patients With Acute Ischemic Stroke: 2019 Update to the 2018 Guidelines for the Early Management of Acute Ischemic Stroke: A Guideline for Healthcare Professionals From the American Heart Association/American Stroke Association. Stroke. 2019;50:e344-e418. doi: 10.1161/STR.0000000000000211.

4. Romano JG, Rundek T. Expanding Treatment for Acute Ischemic Stroke beyond Revascularization. N Engl J Med. 2023;388:2095-2096. doi: 10.1056/NEJMe2303184.

5. Roth GA, Mensah GA, Johnson CO, Addolorato G, Ammirati E, Baddour LM, Barengo NC, Beaton AZ, Benjamin EJ, Benziger CP, et al. Global Burden of Cardiovascular Diseases and Risk Factors, 1990-2019: Update From the GBD 2019 Study. J Am Coll Cardiol. 2020;76:2982-3021. doi: 10.1016/j.jacc.2020.11.010.

6. GBD 2019 Stroke Collaborators. Global, regional, and national burden of stroke and its risk factors, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet Neurol. 2021;20:795-820. doi: 10.1016/S1474-4422(21)00252-0.

7. Hansson L, Zanchetti A, Carruthers SG, Dahlöf B, Elmfeldt D, Julius S, Ménard J, Rahn KH, Wedel H, Westerling S. Effects of intensive blood-pressure lowering and low-dose aspirin in patients with hypertension: principal results of the Hypertension Optimal

Treatment (HOT) randomised trial. HOT Study Group. Lancet. 1998;351:1755-1762. doi: 10.1016/s0140-6736(98)04311-6.

8. Duckworth W, Abraira C, Moritz T, Reda D, Emanuele N, Reaven PD, Zieve FJ, Marks J, Davis SN, Hayward R,et al. Glucose control and vascular complications in veterans with type 2 diabetes. N Engl J Med. 2009;360:129-139. doi: 10.1056/NEJMoa0808431.

9. Amarenco P, Labreuche J. Lipid management in the prevention of stroke: review and updated meta-analysis of statins for stroke prevention. Lancet Neurol. 2009;8:453-463. doi: 10.1016/S1474-4422(09)70058-4.

10. Kim YS, Leventhal BL. Genetic epidemiology and insights into interactive genetic and environmental effects in autism spectrum disorders. Biol Psychiatry. 2015;77:66-74. doi: 10.1016/j.biopsych.2014.11.001.

11. Smith GD, Ebrahim S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? Int J Epidemiol. 2003;32:1-22. doi: 10.1093/ije/dyg070.

12. Grover S, Del Greco M F, Stein CM, Ziegler A. Mendelian Randomization. Methods Mol Biol. 2017;1666:581-628. doi: 10.1007/978-1-4939-7274-6_29.

13. Verduijn M, Siegerink B, Jager KJ, Zoccali C, Dekker FW. Mendelian randomization: use of genetics to enable causal inference in observational studies. Nephrol Dial Transplant. 2010;25:1394-1398. doi: 10.1093/ndt/gfq098.

14. Davies NM, Holmes MV, Davey Smith G. Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians. BMJ. 2018;362:k601. doi: 10.1136/bmj.k601.

15. Giambartolomei C, Vukcevic D, Schadt EE, Franke L, Hingorani AD, Wallace C, Plagnol

V. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS Genet. 2014;10:e1004383. doi: 10.1371/journal.pgen.1004383.

16. Yarmolinsky J, Bouras E, Constantinescu A, Burrows K, Bull CJ, Vincent EE, Martin RM, Dimopoulou O, Lewis SJ, Moreno V, et al. Genetically proxied glucose-lowering drug target perturbation and risk of cancer: a Mendelian randomisation analysis. Diabetologia. 2023. doi: 10.1007/s00125-023-05925-4.

17. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Marshall KA,et al. NCBI GEO: archive for high-throughput functional genomic data. Nucleic Acids Res. 2009;37(Database issue):D885-890. doi: 10.1093/nar/gkn764.

18. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, et al. NCBI GEO: archive for functional genomics data sets--update. Nucleic Acids Res. 2013;41(Database issue):D991-995. doi: 10.1093/nar/gks1193.

19. Xu B, Wang L, Zhan H, Zhao L, Wang Y, Shen M, Xu K, Li L, Luo X, Zhou S, et al. Investigation of the Mechanism of Complement System in Diabetic Nephropathy via Bioinformatics Analysis. J Diabetes Res. 2021;2021:5546199. doi: 10.1155/2021/5546199.

20. Hemani G, Zheng J, Elsworth B, Wade KH, Haberland V, Baird D, Laurin C, Burgess S, Bowden J, Langdon R,et al. The MR-Base platform supports systematic causal inference across the human phenome. Elife. 2018;7:e34408. doi: 10.7554/eLife.34408.

21. Xiang M, Wang Y, Gao Z, Wang J, Chen Q, Sun Z, Liang J, Xu J. Exploring causal correlations between inflammatory cytokines and systemic lupus erythematosus: A

Mendelian randomization. Front Immunol. 202;13:985729. doi: 10.3389/fimmu.2022.985729.

22. Zhang B, Wu Q, Li B, Wang D, Wang L, Zhou YL. m6A regulator-mediated methylation modification patterns and tumor microenvironment infiltration characterization in gastric cancer. Mol Cancer. 2020;19:53. doi: 10.1186/s12943-020-01170-0.

23.Biffi A, Anderson CD, Battey TW, Ayres AM, Greenberg SM, Viswanathan A, Rosand J. Association Between Blood Pressure Control and Risk of Recurrent Intracerebral Hemorrhage. JAMA. 2015;314:904-912. doi: 10.1001/jama.2015.10082.

24. Boden-Albala B, Cammack S, Chong J, Wang C, Wright C, Rundek T, Elkind MS, Paik MC, Sacco RL. Diabetes, fasting glucose levels, and risk of ischemic stroke and vascular events: findings from the Northern Manhattan Study (NOMAS). Diabetes Care. 2008;31:1132-1137. doi: 10.2337/dc07-0797.

25. Nomani AZ, Nabi S, Ahmed S, Iqbal M, Rajput HM, Rao S. High HbA1c is associated with higher risk of ischaemic stroke in Pakistani population without diabetes. Stroke Vasc Neurol. 2016;1:133-139. doi: 10.1136/svn-2016-000018.

26. Bansal BC, Sood AK, Bansal CB. Familial hyperlipidemia in stroke in the young. Stroke. 1986;17:1142-1145. doi: 10.1161/01.str.17.6.1142.

27. Schwartz GG, Steg PG, Szarek M, Bhatt DL, Bittner VA, Diaz R, Edelberg JM, Goodman SG, Hanotin C, Harrington RA, et al. Alirocumab and Cardiovascular Outcomes after Acute Coronary Syndrome. N Engl J Med. 2018;379:2097-2107. doi: 10.1056/NEJMoa1801174.

28. Berger JS, McGinn AP, Howard BV, Kuller L, Manson JE, Otvos J, Curb JD, Eaton CB, Kaplan RC, Lynch JK, et al. Lipid and lipoprotein biomarkers and the risk of ischemic stroke in postmenopausal women. Stroke. 2012;43:958-66. doi: 10.1161/STROKEAHA.111.641324.

29. Johannesen CDL, Mortensen MB, Langsted A, Nordestgaard BG. ApoB and Non-HDL Cholesterol Versus LDL Cholesterol for Ischemic Stroke Risk. Ann Neurol. 2022;92:379-389. doi: 10.1002/ana.26425.

30. Das Pradhan A, Glynn RJ, Fruchart JC, MacFadyen JG, Zaharris ES, Everett BM, Campbell SE, Oshima R, Amarenco P, Blom DJ, et al. Triglyceride Lowering with Pemafibrate to Reduce Cardiovascular Risk. N Engl J Med. 2022 ;387:1923-1934. doi: 10.1056/NEJMoa2210645.

31. Saber H, Yakoob MY, Shi P, Longstreth WT Jr, Lemaitre RN, Siscovick D, Rexrode KM, Willett WC, Mozaffarian D. Omega-3 Fatty Acids and Incident Ischemic Stroke and Its Atherothrombotic and Cardioembolic Subtypes in 3 US Cohorts. Stroke. 2017;48:2678-2685. doi: 10.1161/STROKEAHA.117.018235.

32. Venø SK, Schmidt EB, Jakobsen MU, Lundbye-Christensen S, Bach FW, Overvad K. Substitution of Linoleic Acid for Other Macronutrients and the Risk of Ischemic Stroke. Stroke. 2017;48:3190-3195. doi: 10.1161/STROKEAHA.117.017935.

33. Curb JD, Abbott RD, Rodriguez BL, Masaki KH, Chen R, Popper JS, Petrovitch H, Ross GW, Schatz IJ, Belleau GC, et al. High density lipoprotein cholesterol and the risk of stroke in elderly men: the Honolulu heart program. Am J Epidemiol. 2004;160:150-157. doi: 10.1093/aje/kwh177.

34. O'Donnell MJ, McQueen M, Sniderman A, Pare G, Wang X, Hankey GJ, Rangarajan S, Chin SL, Rao-Melacini P, Ferguson J, et al. Association of Lipids, Lipoproteins, and Apolipoproteins with Stroke Subtypes in an International Case Control Study (INTERSTROKE). J Stroke. 2022;24:224-235. doi: 10.5853/jos.2021.02152.

35. National Center for Biotechnology Information (2023). PubChem Gene Summary for Gene 2068, ERCC2 - ERCC excision repair 2, TFIIH core complex helicase subunit (human). Retrieved June 14, 2023 from https://pubchem.ncbi.nlm.nih.gov/gene/ERCC2/human.

36. Shyu HY, Shieh JC, Ji-Ho L, Wang HW, Cheng CW. Polymorphisms of DNA repair pathway genes and cigarette smoking in relation to susceptibility to large artery atherosclerotic stroke among ethnic Chinese in Taiwan. J Atheroscler Thromb. 2012;19:316-325. doi: 10.5551/jat.10967.

37. LD, Dawsey SM, Dong ZW, Taylor PR, Mark SD. A prospective study of polymorphisms of DNA repair genes XRCC1, XPD23 and APE/ref-1 and risk of stroke in Linxian, China. J Epidemiol Community Health. 2007;61:737-741. doi: 10.1136/jech.2006.048934.

38. Zhang J, Guo F, Zhou R, Xiang C, Zhang Y, Gao J, Cao G, Yang H. Proteomics and transcriptome reveal the key transcription factors mediating the protection of Panax notoginseng saponins (PNS) against cerebral ischemia/reperfusion injury. Phytomedicine. 2021;92:153613. doi: 10.1016/j.phymed.2021.153613.

39. National Center for Biotechnology Information (2023). PubChem Gene Summary for Gene 2237, FEN1 - flap structure-specific endonuclease 1 (human). Retrieved June 15, 2023 from https://pubchem.ncbi.nlm.nih.gov/gene/FEN1/human.

40. He Z, Ning N, Zhou Q, Khoshnam SE, Farzaneh M. Mitochondria as a therapeutic target for ischemic stroke. Free Radic Biol Med. 2020;146:45-58. doi: 10.1016/j.freeradbiomed.2019.11.005.

41. Andrabi SS, Parvez S, Tabassum H. Ischemic stroke and mitochondria: mechanisms and targets. Protoplasma. 2020;257:335-343. doi: 10.1007/s00709-019-01439-2.

42. Sun D, Lee G, Lee JH, Kim HY, Rhee HW, Park SY, Kim KJ, Kim Y, Kim BY, Hong JI, et al. A metazoan ortholog of SpoT hydrolyzes ppGpp and functions in starvation responses. Nat Struct Mol Biol. 2010;17:1188-94. doi: 10.1038/nsmb.1906.

43. He S, Wang C, Dong H, Xia F, Zhou H, Jiang X, Pei C, Ren H, Li H, Li R, et al. Immune-related GTPase M (IRGM1) regulates neuronal autophagy in a mouse model of stroke. Autophagy. 2012;8:1621-1627. doi: 10.4161/auto.21561.

44. Yamase Y, Horibe H, Ueyama C, Fujimaki T, Oguri M, Kato K, Arai M, Watanabe S, Yamada Y. Association of *TOMM40* and *SLC22A4* polymorphisms with ischemic stroke. Biomed Rep. 2015;3:491-498. doi: 10.3892/br.2015.457.

45. Guo JM, Liu AJ, Zang P, Dong WZ, Ying L, Wang W, Xu P, Song XR, Cai J, Zhang SQ, et al. ALDH2 protects against stroke by clearing 4-HNE. Cell Res. 2013;23:915-930. doi: 10.1038/cr.2013.69.

46. Xia P, Zhang F, Yuan Y, Chen C, Huang Y, Li L, Wang E, Guo Q, Ye Z. ALDH 2 conferred neuroprotection on cerebral ischemic injury by alleviating mitochondria-related apoptosis through JNK/caspase-3 signing pathway. Int J Biol Sci. 2020;16:1303-1323. doi: 10.7150/ijbs.38962.

47. Sun S, He J, Zhang Y, Xiao R, Yan M, Ren Y, Zhu Y, Jin T, Xia Y. Genetic

polymorphisms in the ALDH2 gene and the risk of ischemic stroke in a Chinese han population. Oncotarget. 2017;8:101936-101943. doi: 10.18632/oncotarget.21803.

48. Yang Q, Yin RX, Cao XL, Wu DF, Chen WX, Zhou YJ. Association of two polymorphisms in the FADS1/FADS2 gene cluster and the risk of coronary artery disease and ischemic stroke. Int J Clin Exp Pathol. 2015;8:7318-7331.

49. Gągało I, Rusiecka I, Kocić I. Tyrosine Kinase Inhibitor as a new Therapy for Ischemic Stroke and other Neurologic Diseases: is there any Hope for a Better Outcome? Curr Neuropharmacol. 2015;13:836-844. doi: 10.2174/1570159x13666150518235504.

50. Yokota N, Uchijima M, Nishizawa S, Namba H, Koide Y. Identification of differentially expressed genes in rat hippocampus after transient global cerebral ischemia using subtractive cDNA cloning based on polymerase chain reaction. Stroke. 2001;32:168-174. doi: 10.1161/01.str.32.1.168.

51. Peng H, Fan Y, Li J, Zheng X, Zhong C, Zhu Z, He Y, Zhang M, Zhang Y. DNA Methylation of the Natriuretic Peptide System Genes and Ischemic Stroke: Gene-Based and Gene Set Analyses. Neurol Genet. 2022;8:e679. doi: 10.1212/NXG.0000000000000679.

52. Zhang H, Mo X, Wang A, Peng H, Guo D, Zhong C, Zhu Z, Xu T, Zhang Y. Association of DNA Methylation in Blood Pressure-Related Genes With Ischemic Stroke Risk and Prognosis. Front Cardiovasc Med. 2022;9:796245. doi: 10.3389/fcvm.2022.796245.

53. Sundström J, Söderholm M, Borné Y, Nilsson J, Persson M, Östling G, Melander O, Orho-Melander M, Engström G. Eosinophil Cationic Protein, Carotid Plaque, and Incidence of Stroke. Stroke. 2017;48:2686-2692. doi: 10.1161/STROKEAHA.117.018450.

54. Zhong C, Yang J, Xu T, Xu T, Peng Y, Wang A, Wang J, Peng H, Li Q, Ju Z, et al. Serum matrix metalloproteinase-9 levels and prognosis of acute ischemic stroke. Neurology. 2017;89:805-812. doi: 10.1212/WNL.0000000000004257.

55. Ponsaerts L, Alders L, Schepers M, de Oliveira RMW, Prickaerts J, Vanmierlo T, Bronckaers A. Neuroinflammation in Ischemic Stroke: Inhibition of cAMP-Specific Phosphodiesterases (PDEs) to the Rescue. Biomedicines. 2021;9:703. doi: 10.3390/biomedicines9070703.

TABLE 1| Gene name conversion table

Gene description	Gene symbol
ERCC excision repair 2, TFIIH core complex helicase subunit	ERCC2
flap structure-specific endonuclease	FEN1
HD domain containing 3	HDDC3
translocase of outer mitochondrial membrane 40	TOMM40
trafficking protein particle complex subunit 6A	TRAPPC6A
aldehyde dehydrogenase 2 family member	ALDH2
fatty acid desaturase 2	FADS2
FES proto-oncogene, tyrosine kinase	FES
furin, paired basic amino acid cleaving enzyme	FURIN
mannosidase alpha class 2A member 2	MAN2A2
nectin cell adhesion molecule 2	PVRL2
synaptotagmin 7	SYT7
PH domain containing endocytic trafficking adaptor 1	FAM109A

FIGURE 1 | Flow diagram of this study design

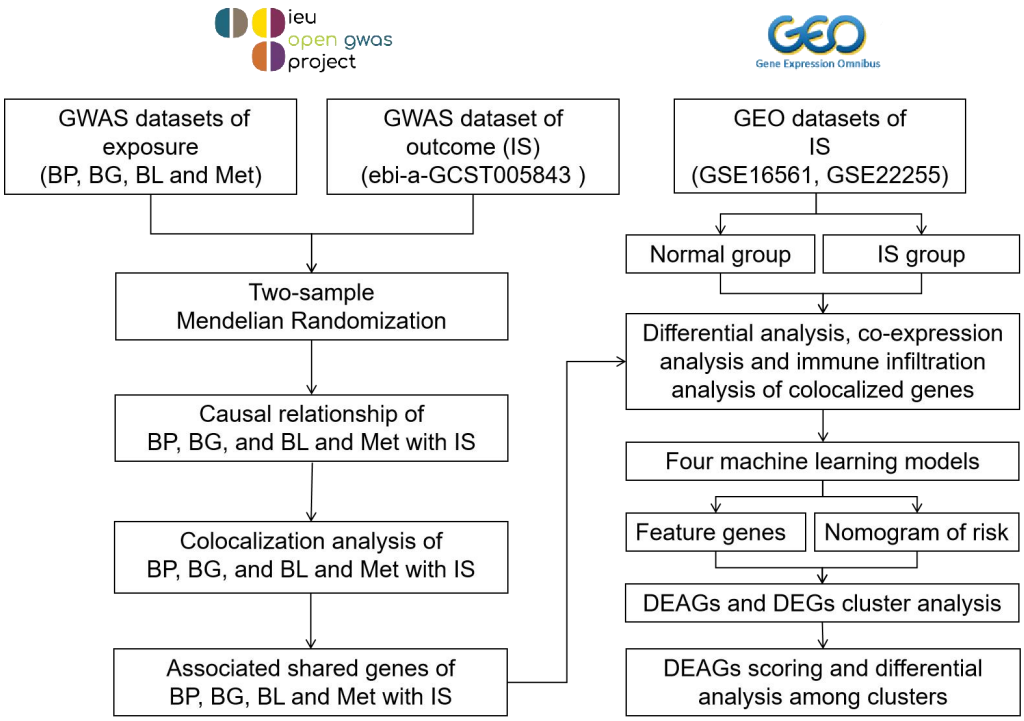


FIGURE 2 | (A) Box plot of expression difference analysis of associated shared genes between normal samples and IS samples; (B) Heat map of DEAGs expression in normal and IS samples; (C) Circle plot of chromosome location of DEAGs; (D) DEAGs correlation network; (E) Correlation analysis between the two DEAGs

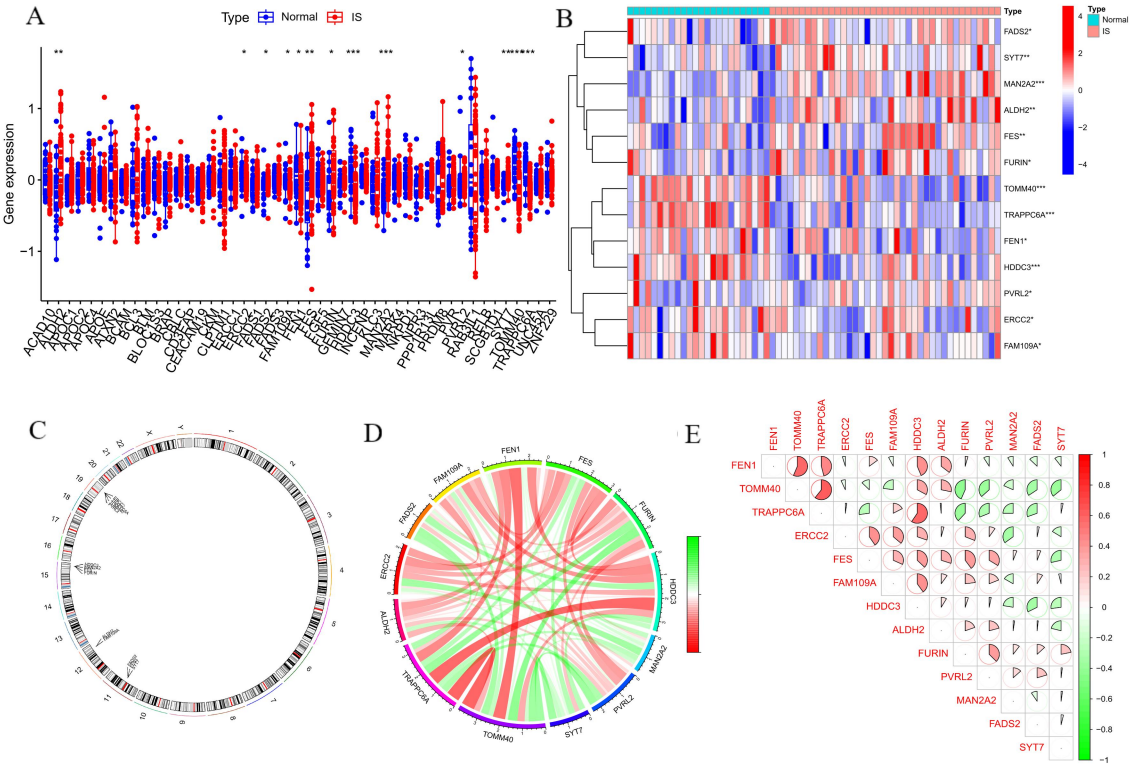


FIGURE 3 | (A) Bar plot of relative percentage of each immune cells in samples; (B) Box plot of immune cell fraction between normal samples and IS samples; (C) Heat map of correlation analysis between DEAGs and immune cells

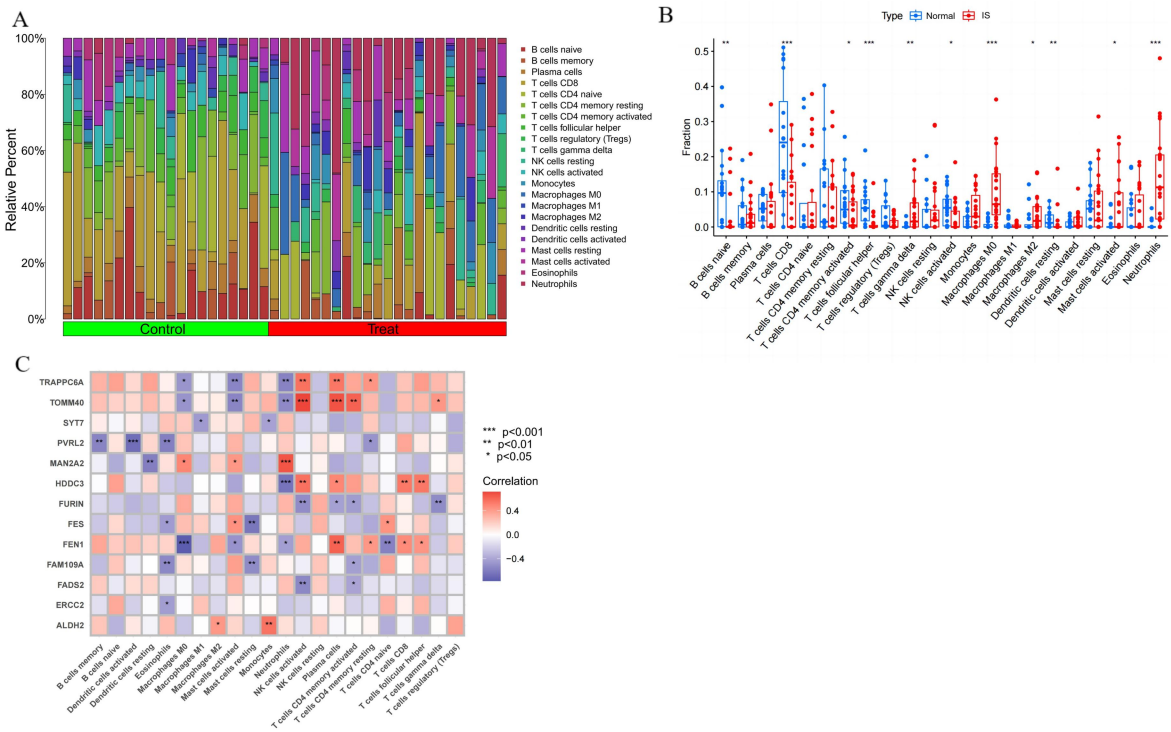


FIGURE 4 | (A) Box plots of residual of the four machine learning models; (B) Reverse cumulative distribution of residual of the four machine learning models; (C) ROC of the four machine learning models; (D) Bar plot of feature importance of the four machine learning models; (E) Nomogram of the feature genes; (F) Calibration curve of feature genes nomogram; (G) Decision curve of feature genes nomogram; (H) ROC of the test GEO dataset

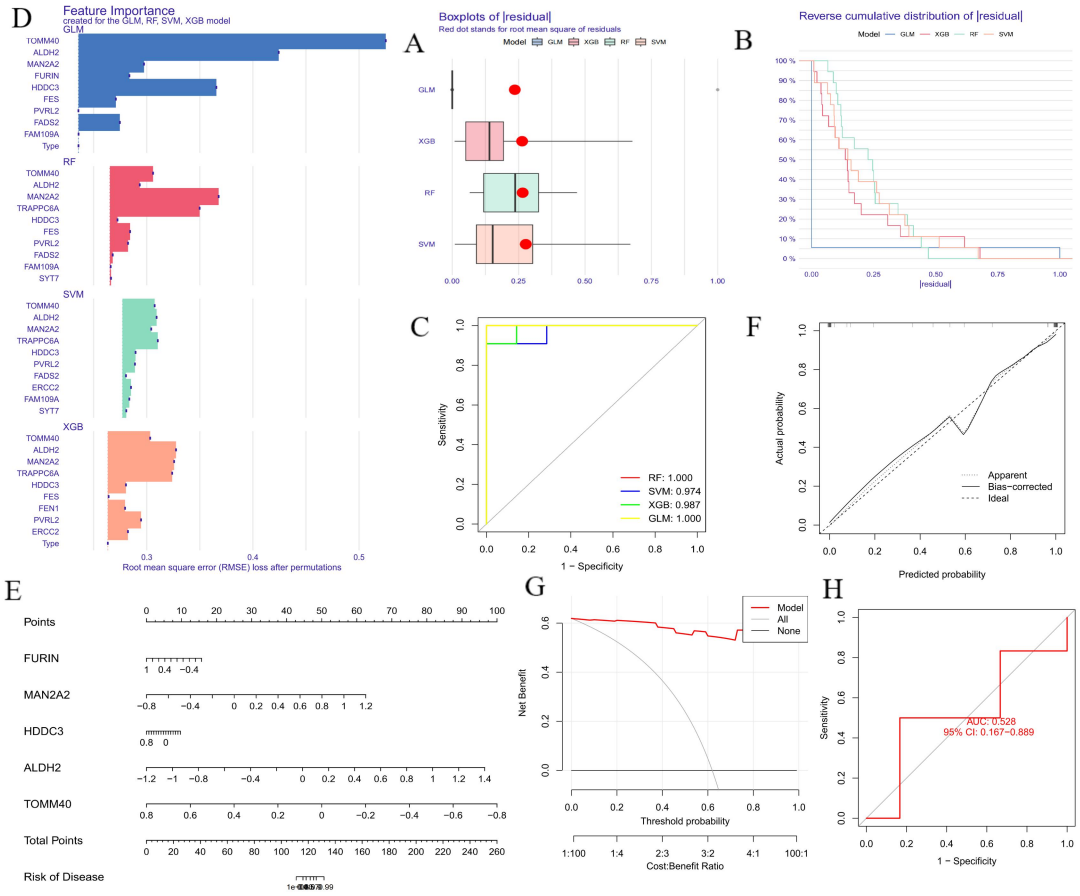


FIGURE 5 | (A) Consensus cumulative distribution plot of DEAG clustering for samples; (B) Consensus CDFs of DEAG clustering; (C) Consensus matrix heat map of DEAG clustering for samples; (D) Heat map of DEAGs expression between DEAG clusters; (E) Box plot of expression difference analysis of DEAG clusters; (F) Scatter plot of PCA between DEAG clusters; (G) Box plot of immune cell fraction between DEAG clusters; (H) Bar plot of relative percentage of each immune cells in samples of DEAG clusters; (I) Bar plot of GO terms of GSVA between DEAG clusters; (J) Bar plot of KEGG terms of GSVA between DEAG clusters

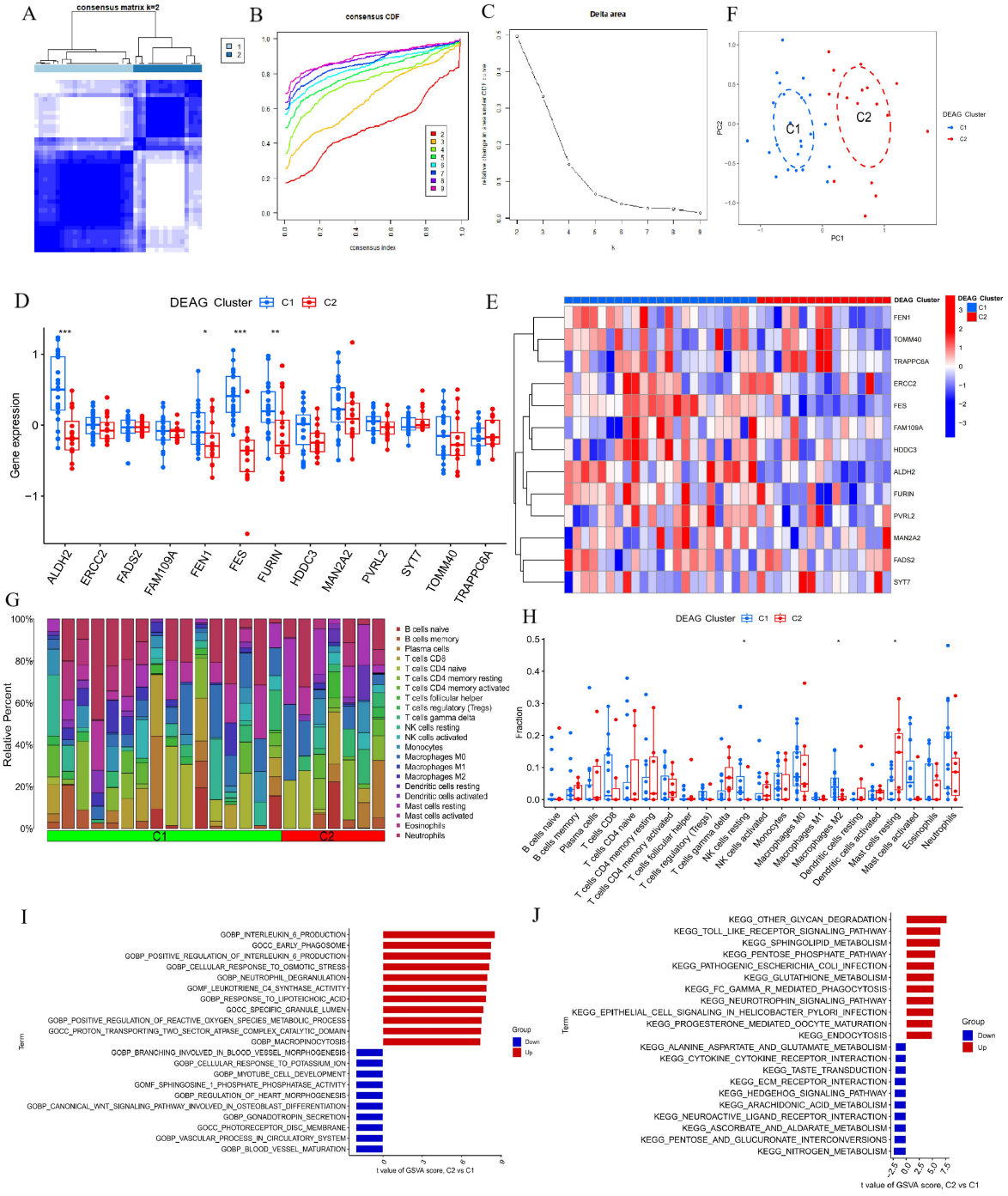


FIGURE 6 | (A) Venn plot of DEGs; (B) Consensus matrix heat map of DEG clustering for IS samples; (C) Heat map of DEGs expression between DEG clusters; (D) Box plot of expression difference analysis of DEG clusters; (E) Box plot of immune cell infiltration between DEG clusters

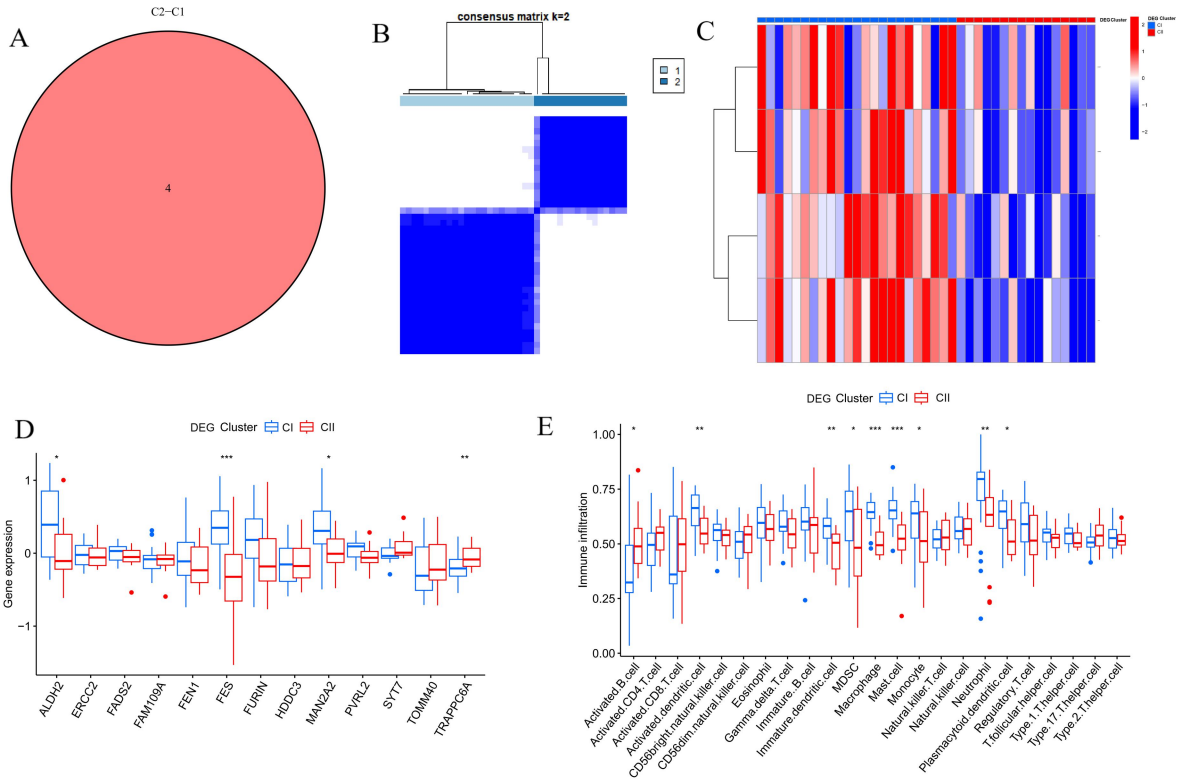
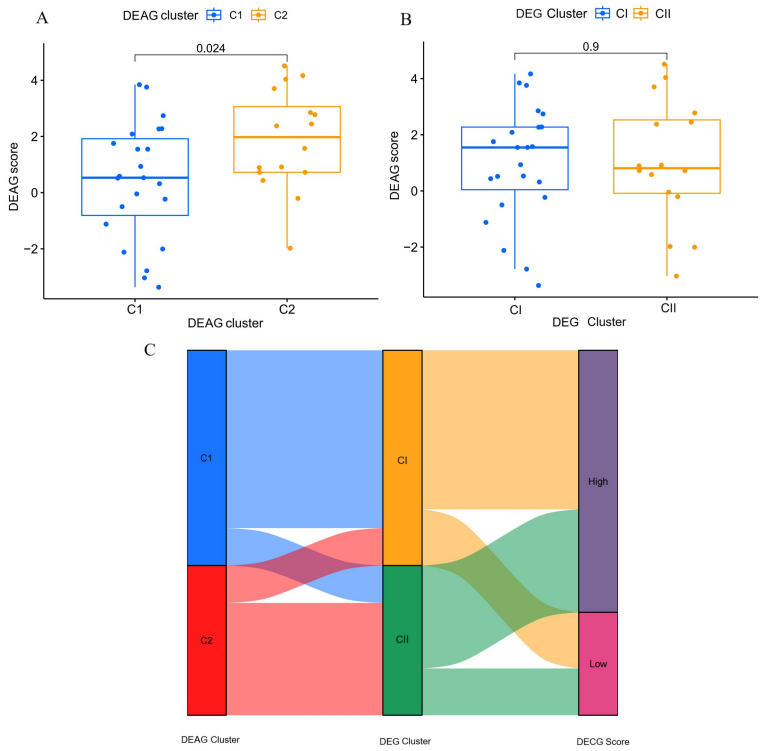
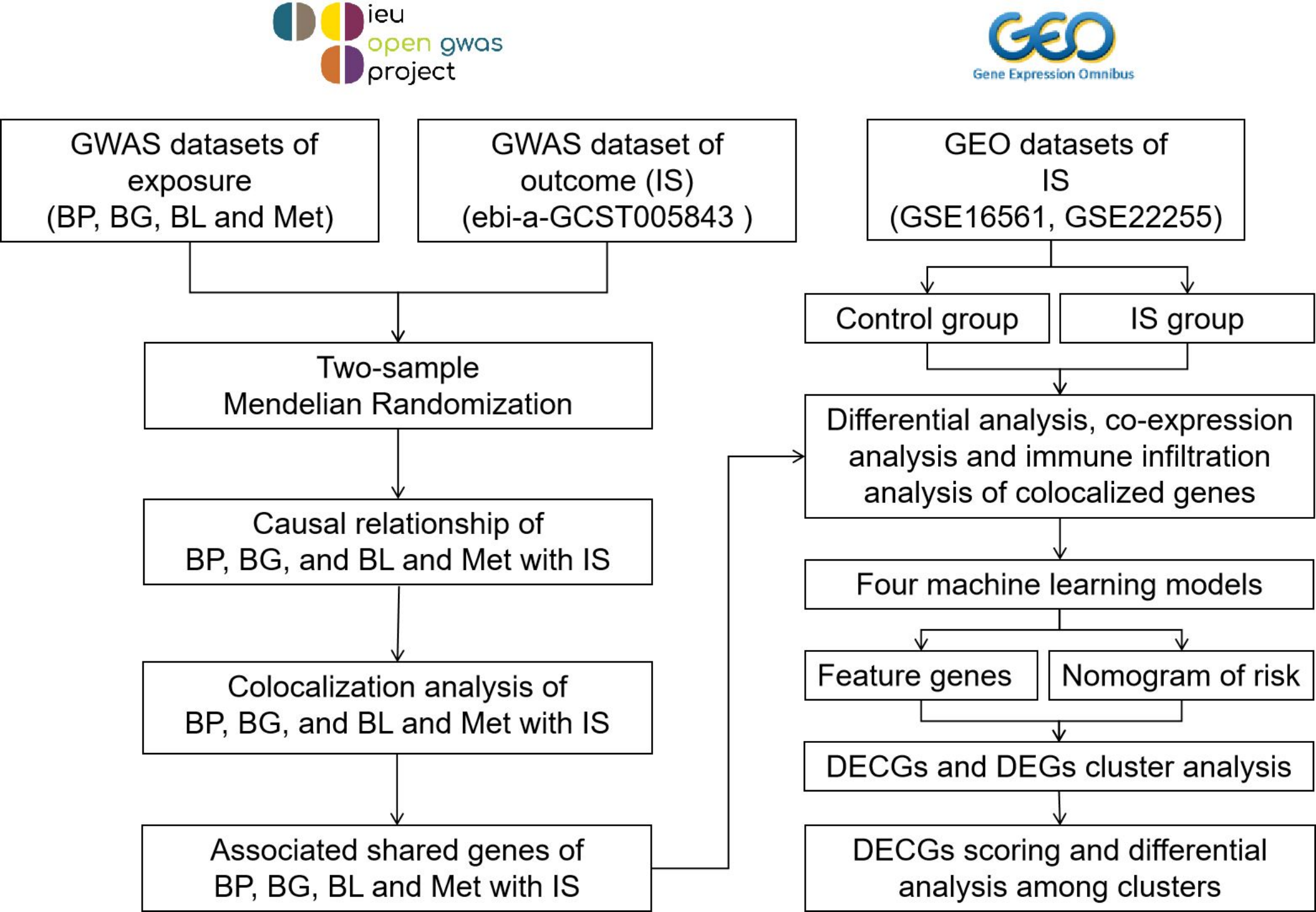
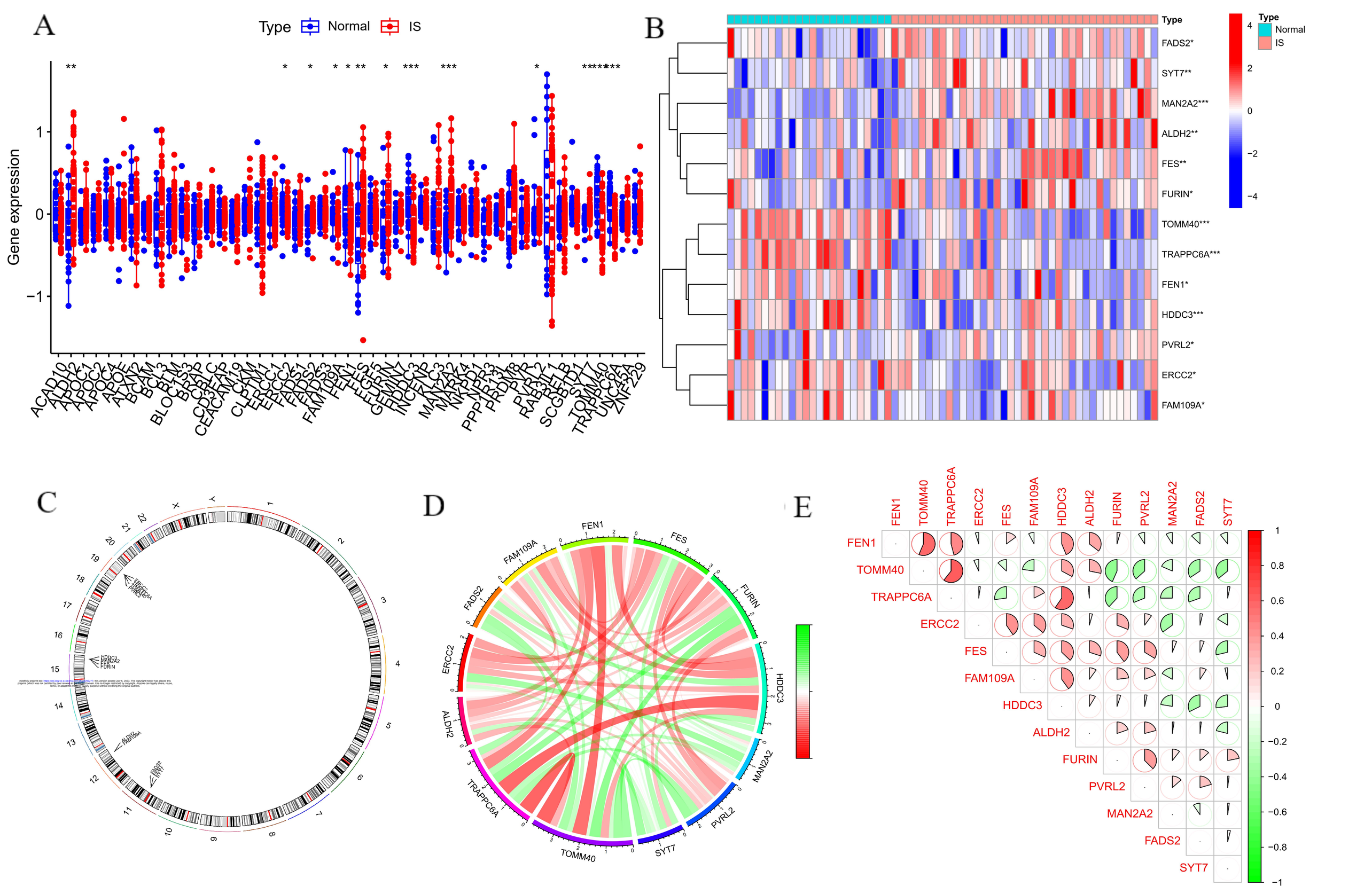


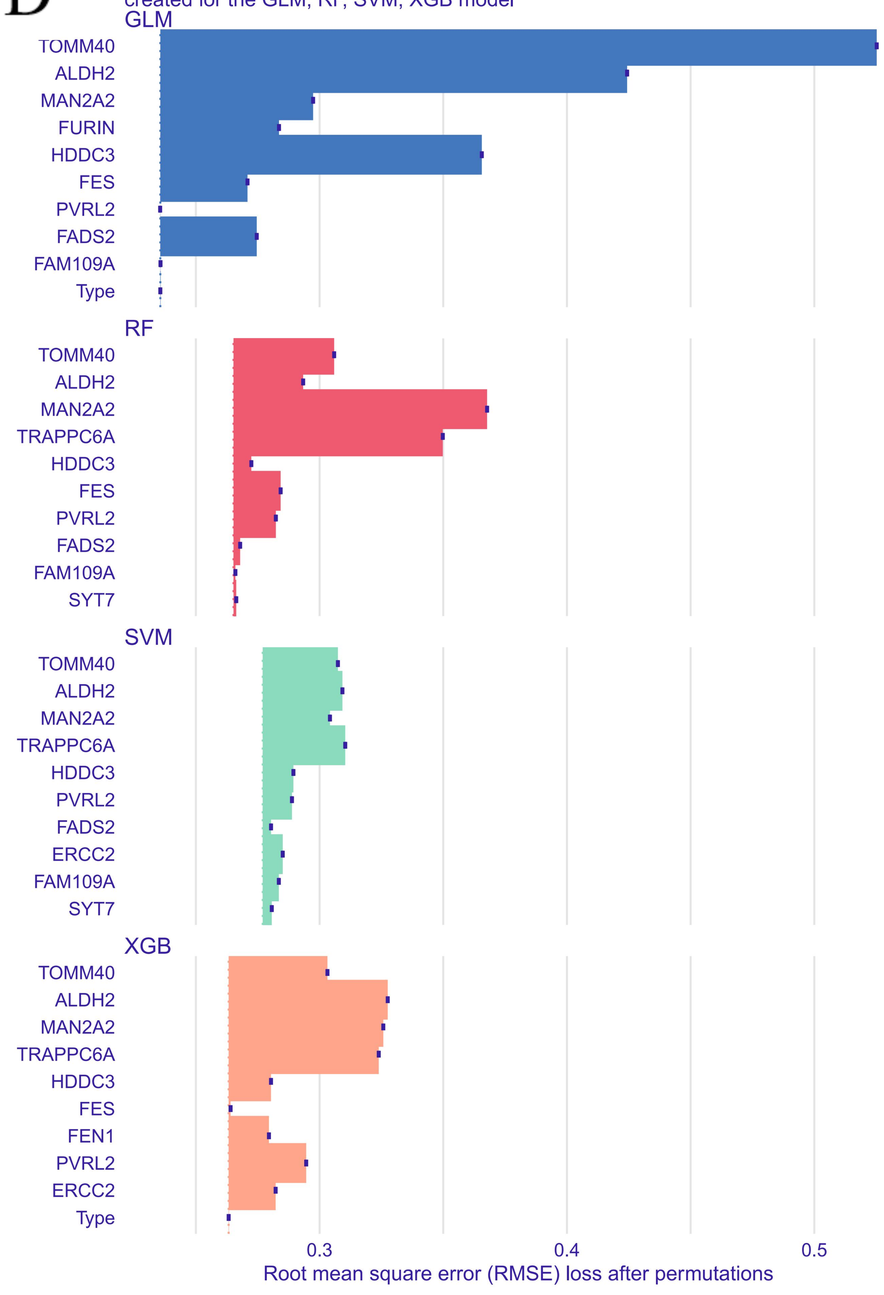
FIGURE 7 | (A) Box plot of different expression analysis of DEAG score between DEAG clusters; (B) Box plot of different expression analysis of DEAG score between DEG clusters; (C) Alluvial plot of the correspondence of the different clustered samples



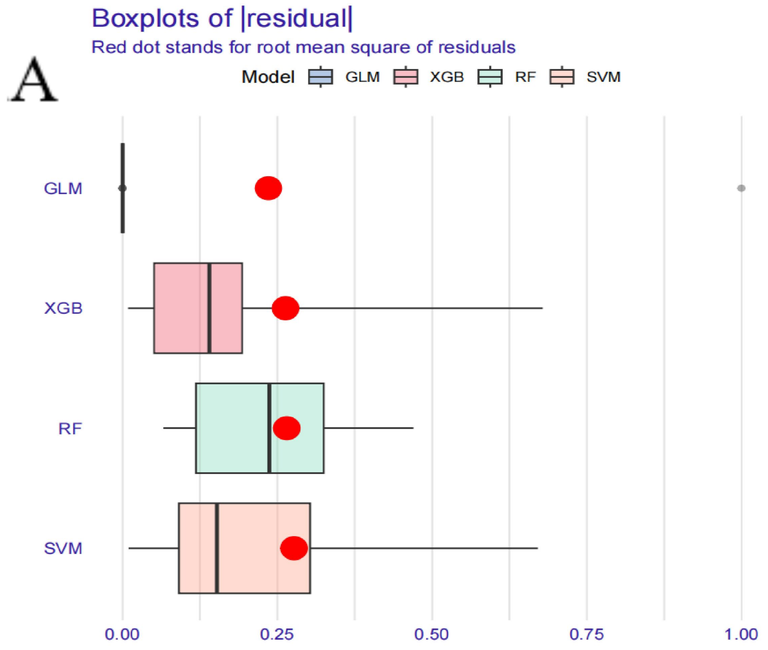

ieu
open gwas
project
Gene Expression Omnibus
GWAS datasets of exposure (BP, BG, BL and Met)
GWAS dataset of outcome (IS) (ebi-a-GCST005843)
GEO datasets of IS (GSE16561, GSE22255)
Two-sample Mendelian Randomization
Causal relationship of BP, BG, and BL and Met with IS
Colocalization analysis of BP, BG, and BL and Met with IS
Associated shared genes of BP, BG, BL and Met with IS
Control group
IS group
Differential analysis, co-expression analysis and immune infiltration analysis of colocalized genes
Four machine learning models
Feature genes
Nomogram of risk
DECGs and DEGs cluster analysis
DECGs scoring and differential analysis among clusters


A
Type
Normal
IS
Gene expression
ACADL10
ALDH2
APOC1
APOC2
APOC4
APOE
ATXN2
BCAM
BCL3
BLOC1S3
BLM
BRAP
CBLC
CD3EAP
CEACAM19
CKM
CLPTM1
ERCC2
ERCC2
FADS1
FADS2
FAM109A
FEN1
FES
FURIN
GEMIN7
HDDC3
INCENP
KLC3
MAN2A2
MARK4
NKPD1
NPR3
PPP1R13
PRDM8
PVR
PVRL2
RAB3IL1
RELB
SCGB1D1
SYT7
TOMM40
TRAPPC6A
UNC45A
ZNF229
B
Type
FADS2*
SYT7**
MAN2A2***
ALDH2**
FES**
FURIN*
TOMM40***
TRAPPC6A***
FEN1*
HDDC3***
PVRL2*
ERCC2*
FAM109A*
Type
Normal
IS
C
D
E
FEN1
TOMM40
TRAPPC6A
ERCC2
FES
FAM109A
HDDC3
ALDH2
FURIN
PVRL2
MAN2A2
FADS2
SYT7

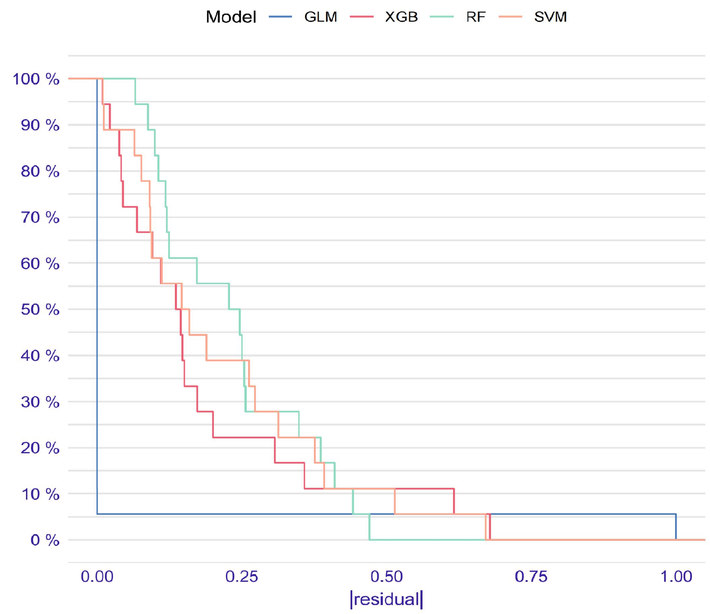
D



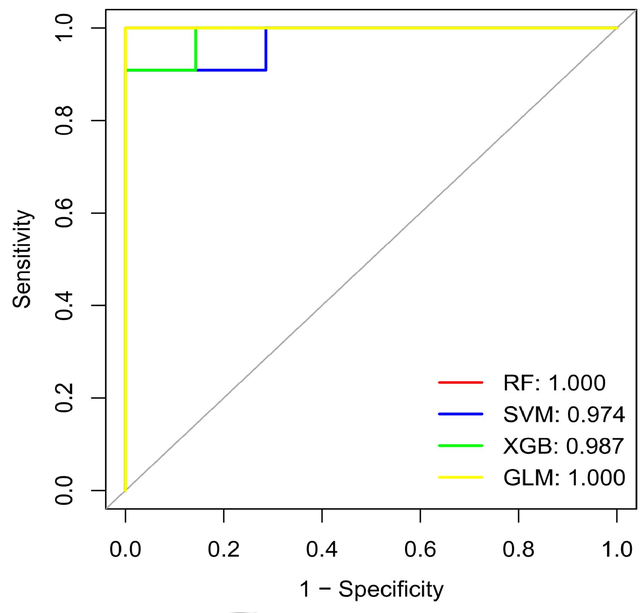
A



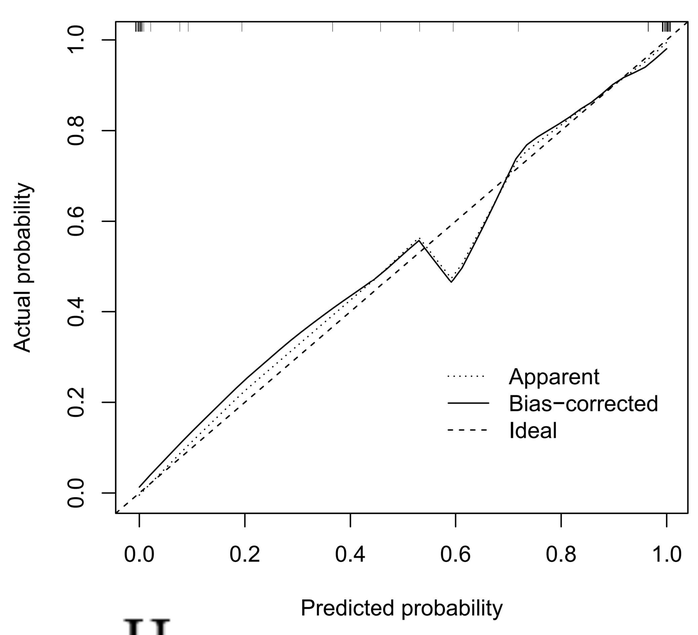
B



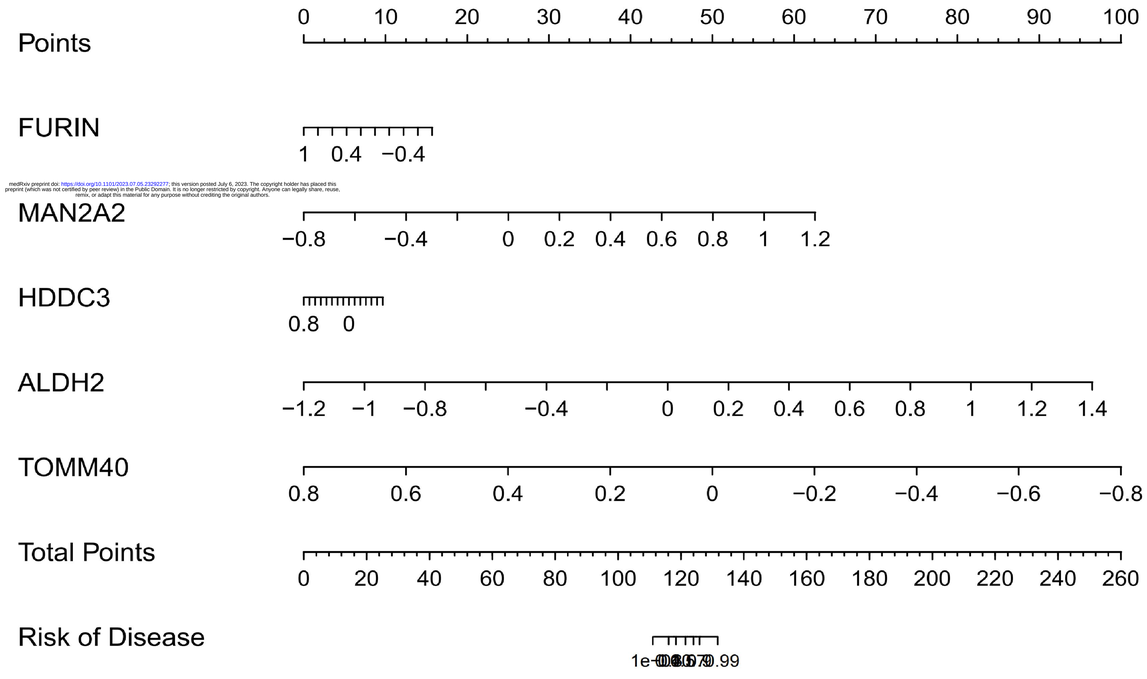
C



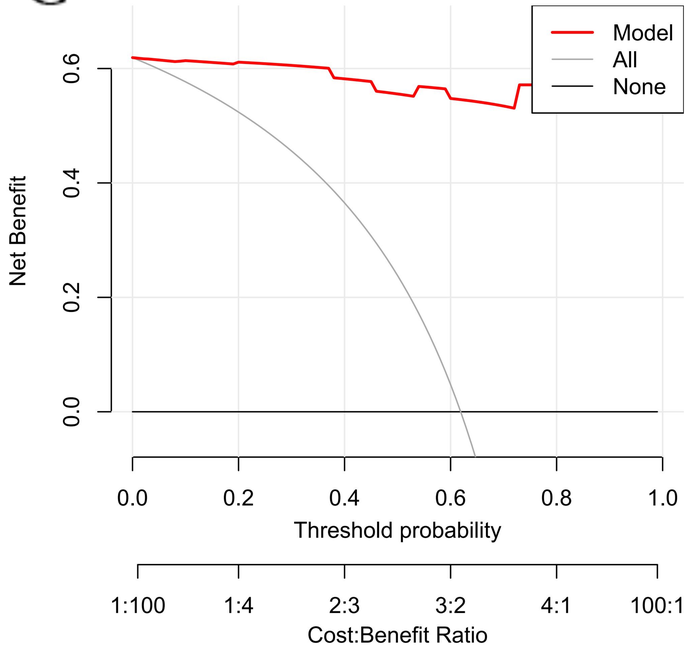
F



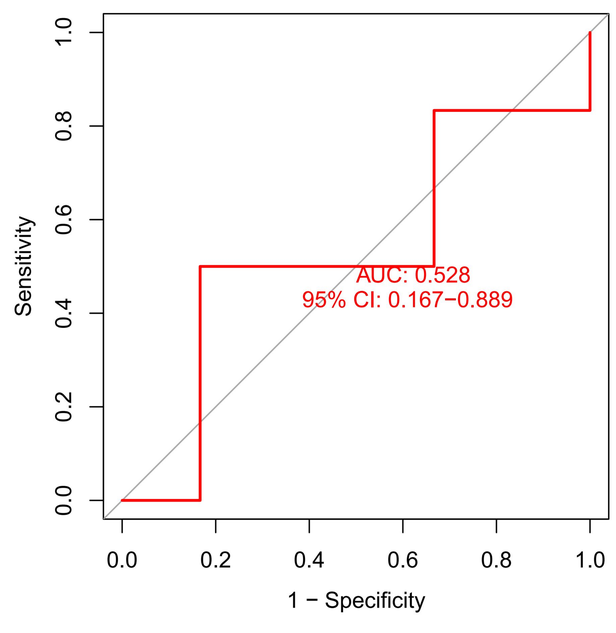
E

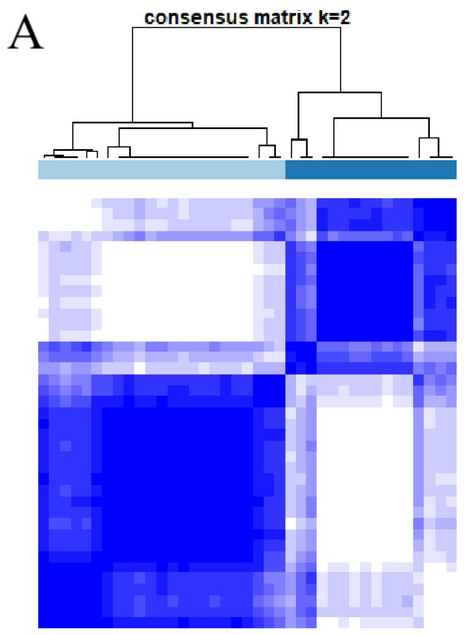


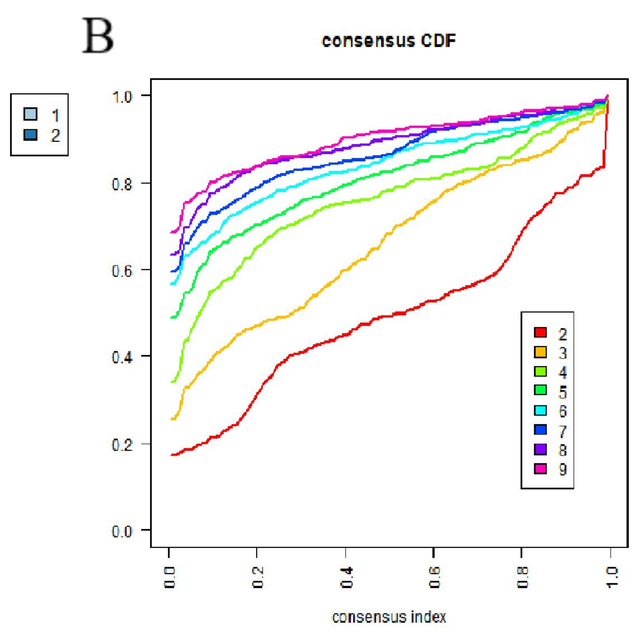
G

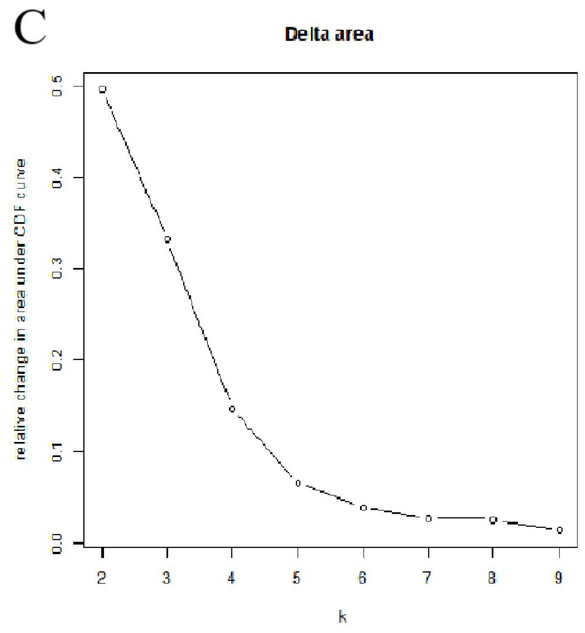


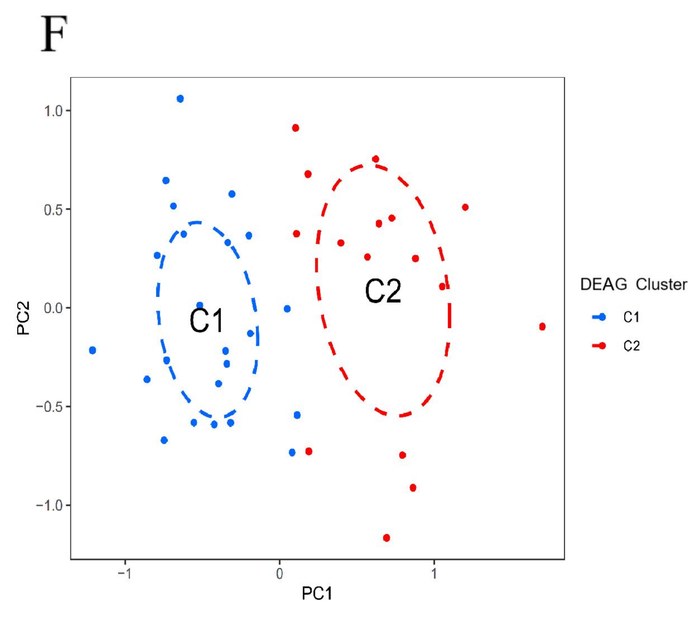
H

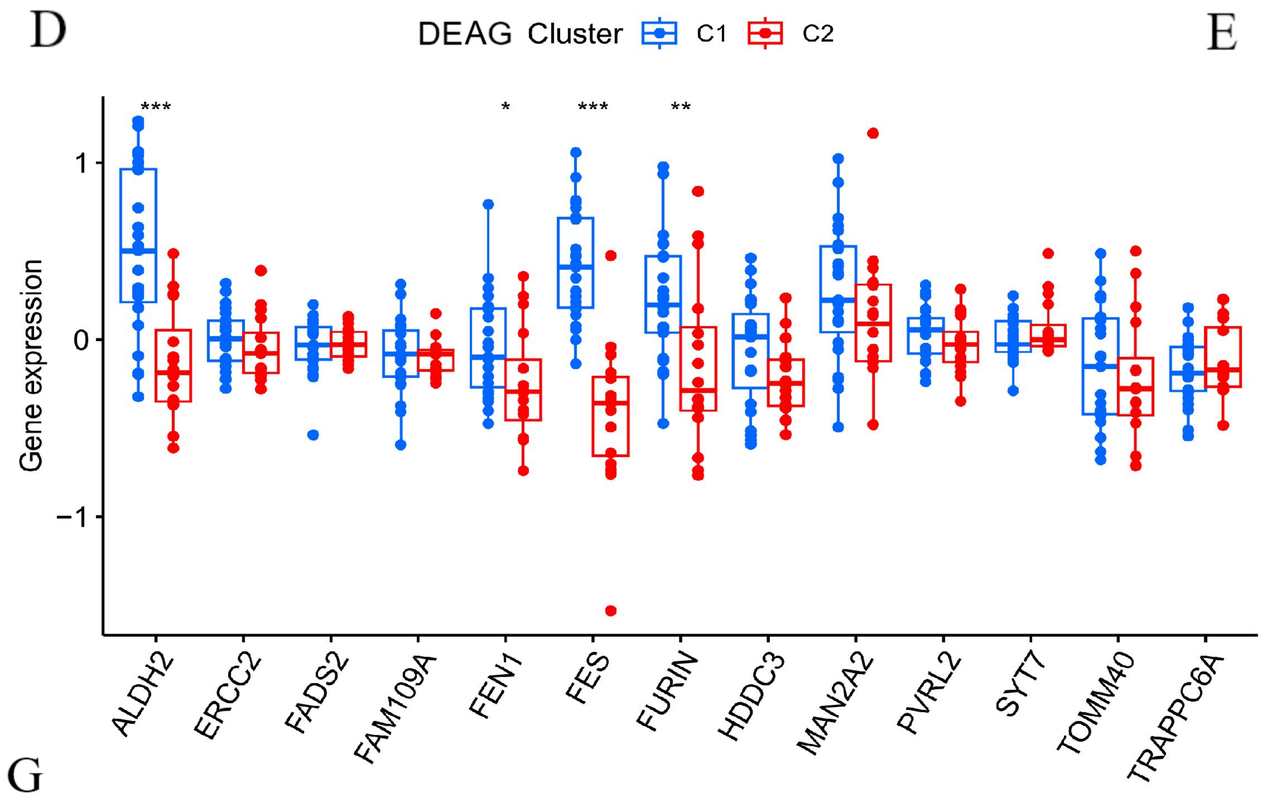


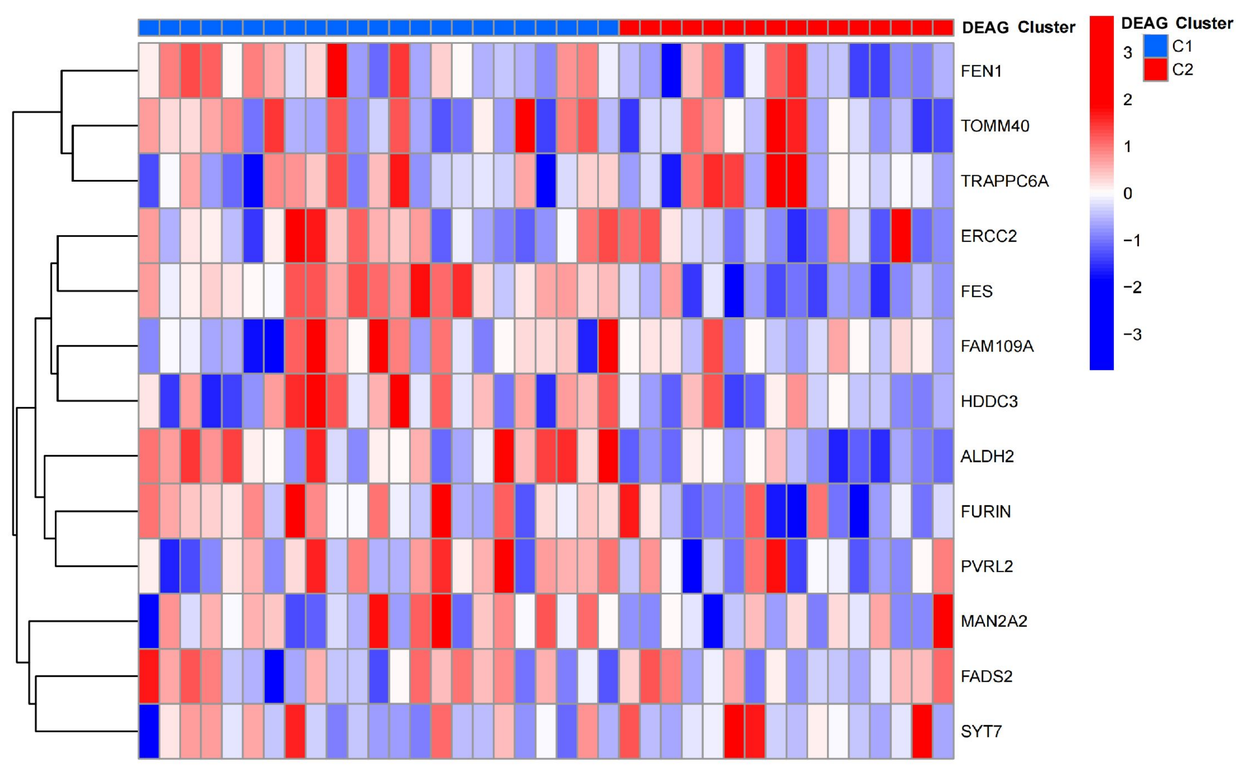


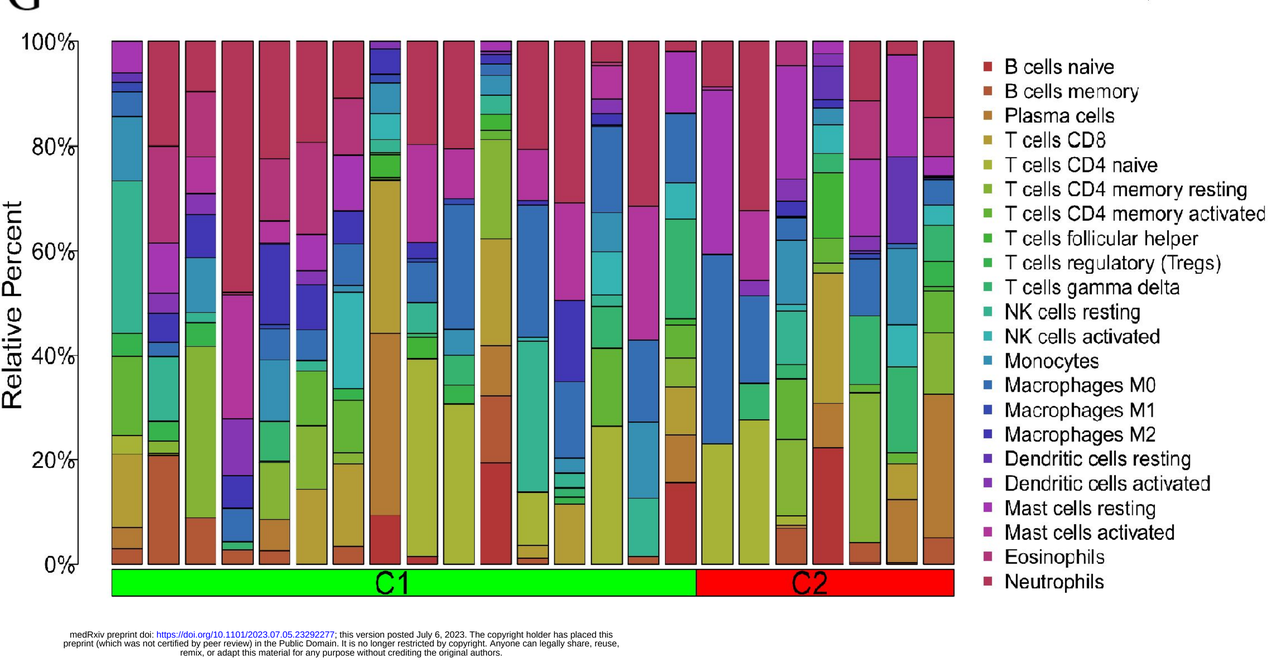


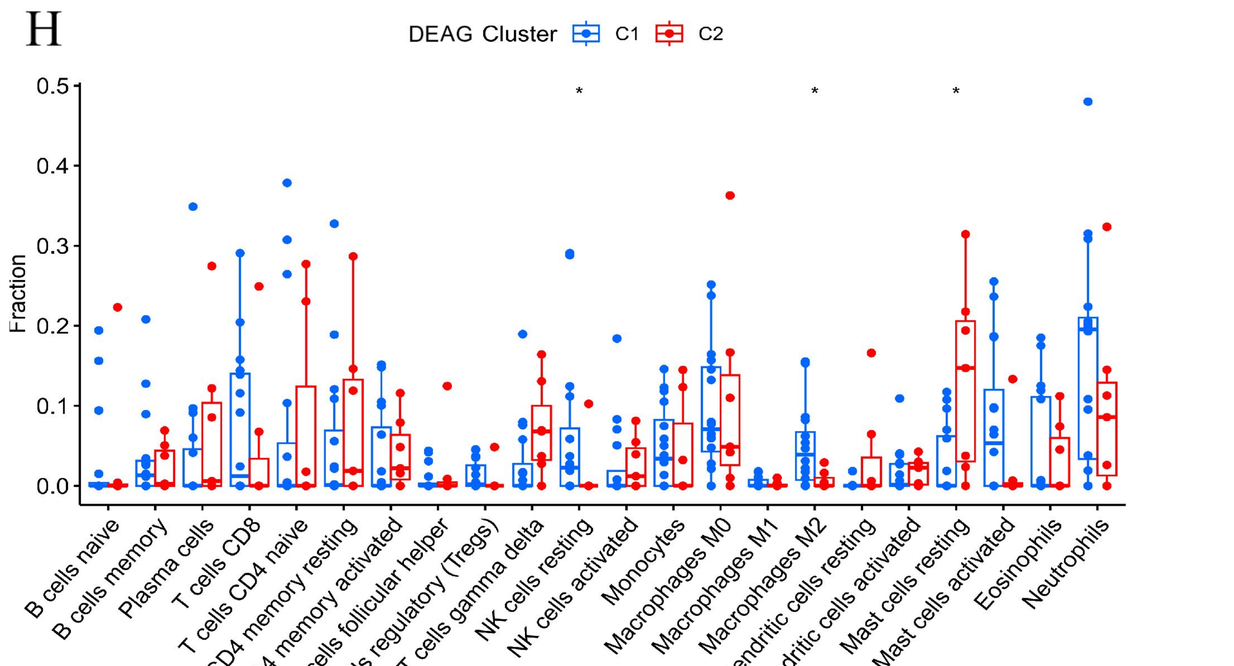


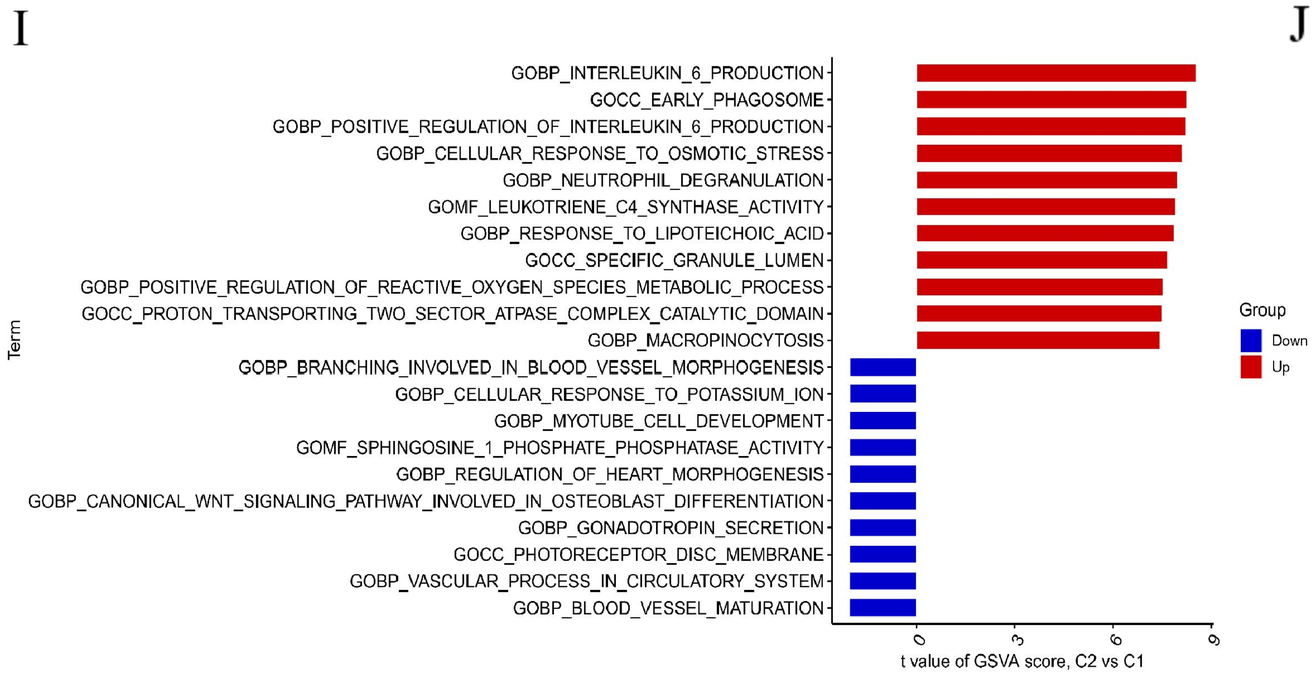


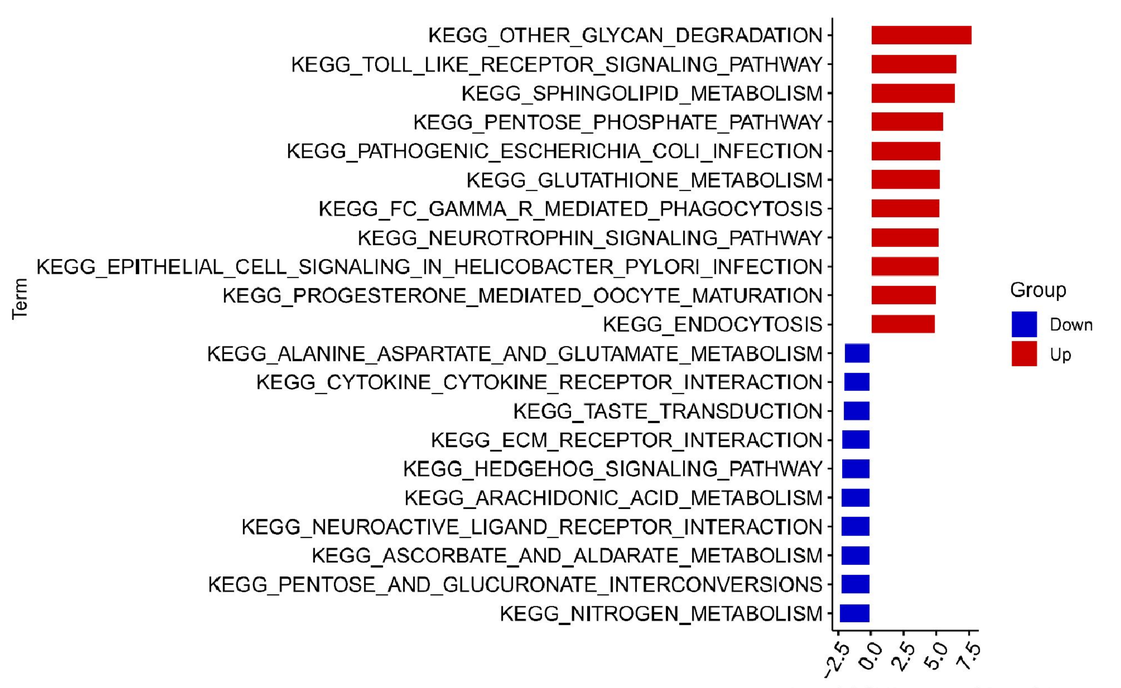


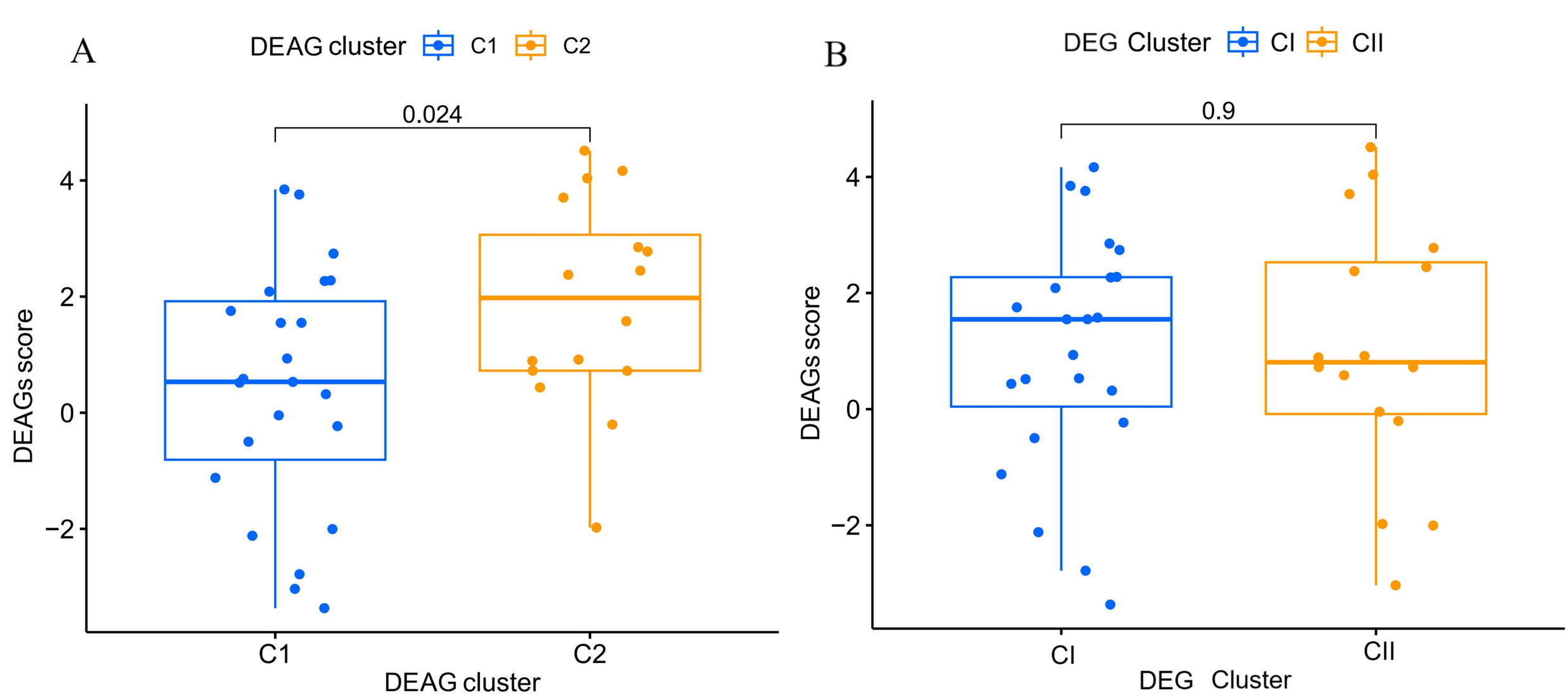


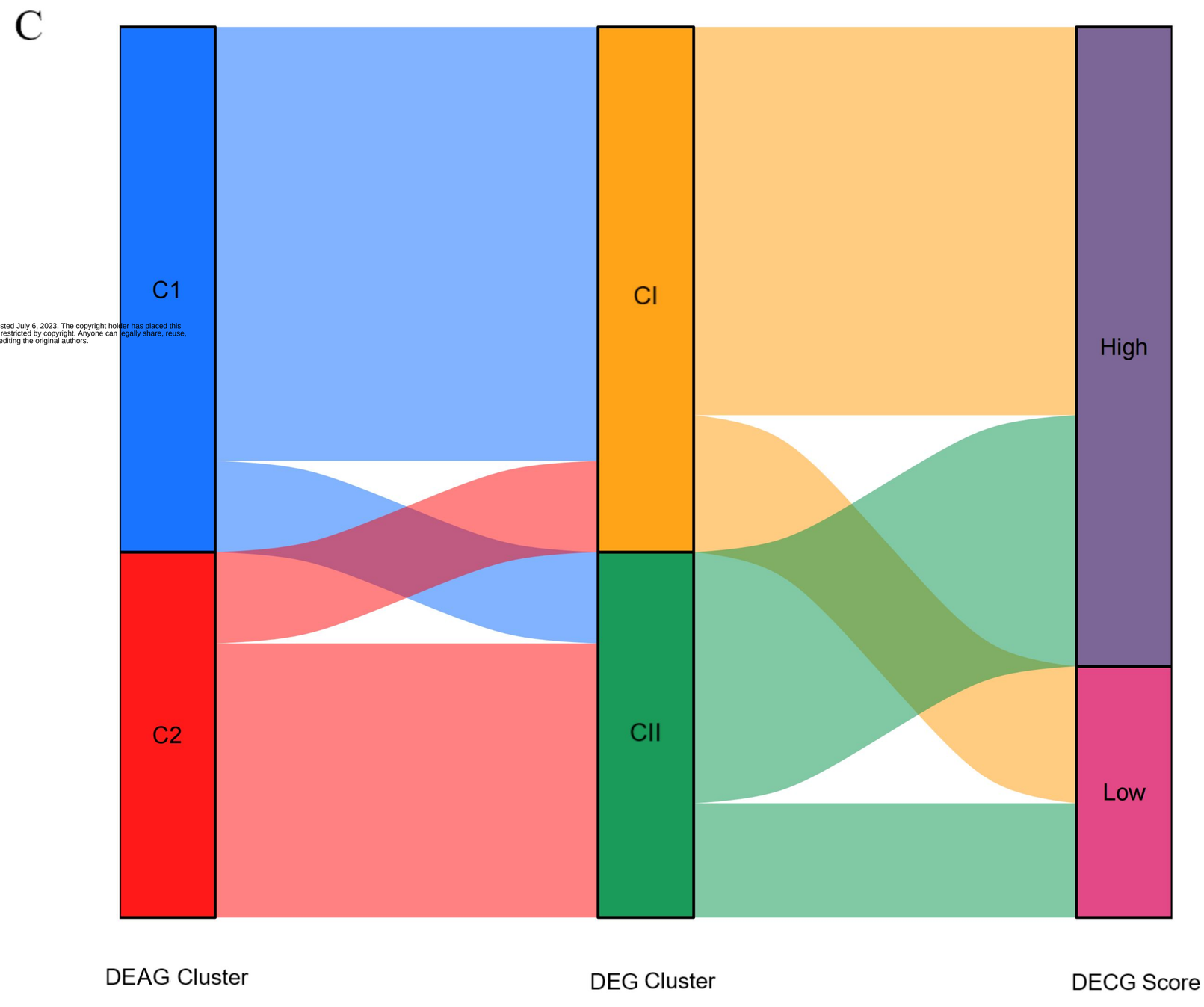




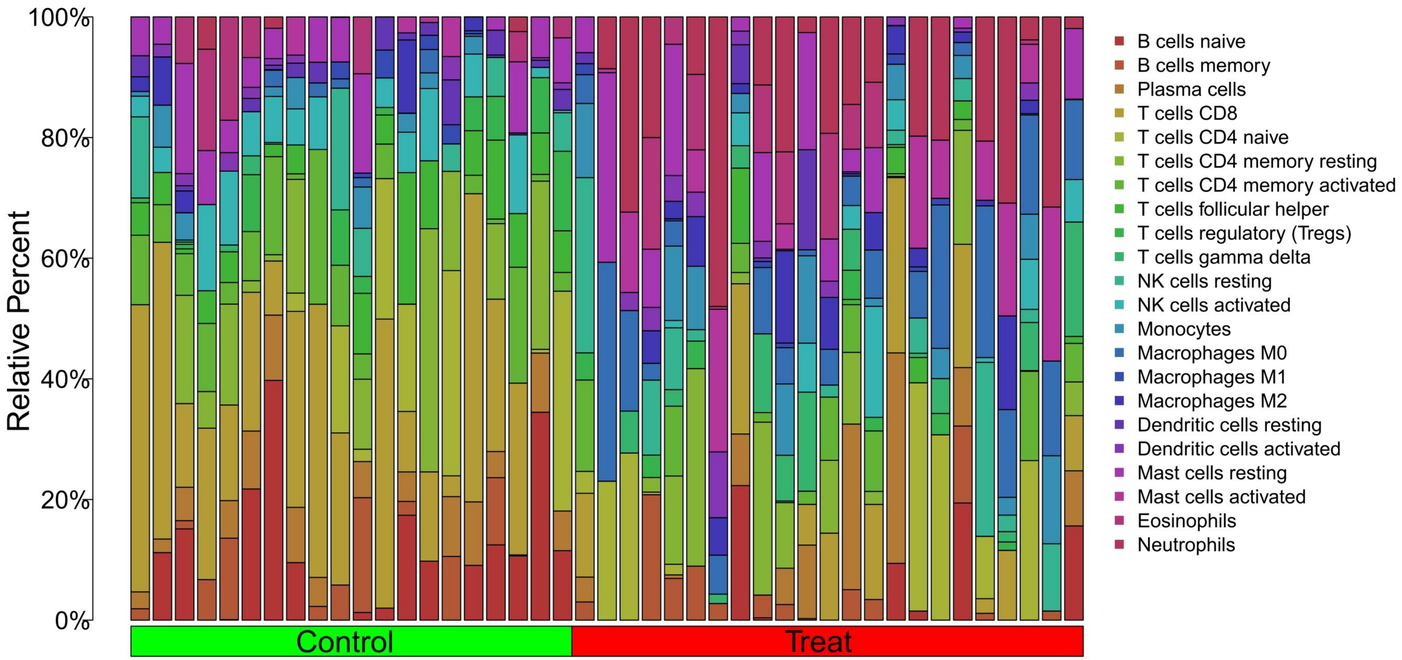




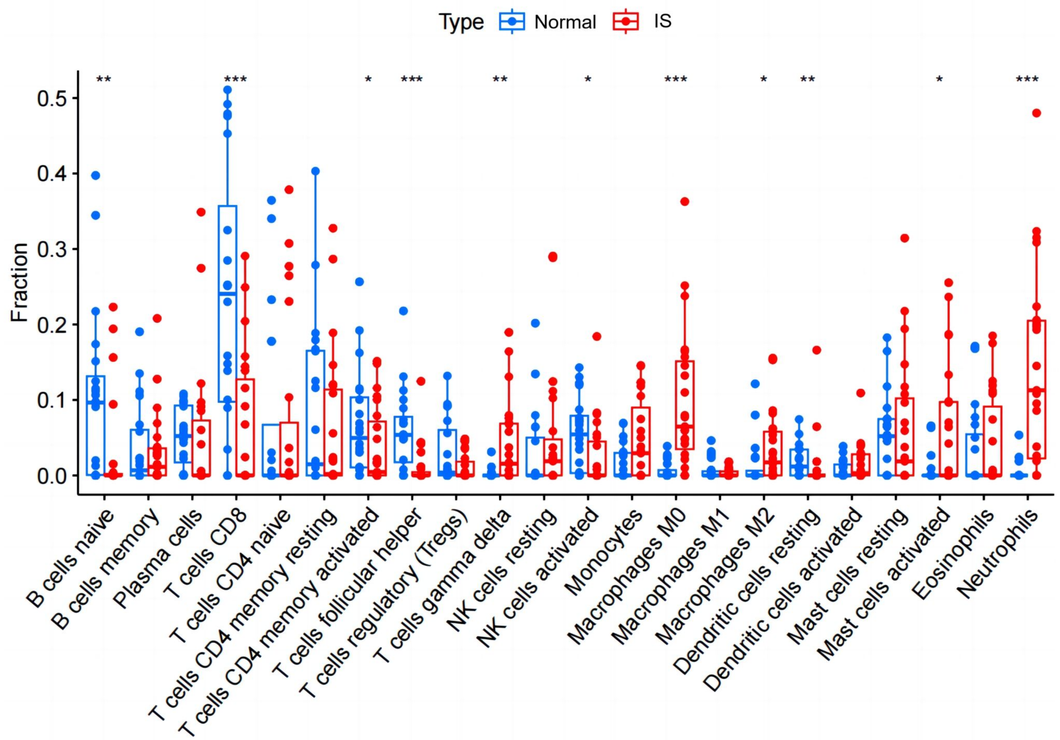




A



B



C