Supplemantary materials

Methods. ***Medical ethics.*** In this study, no human or animal experiments were involved.

Data download. We downloaded the datasets (GSE58294, GSE16561) from NCBI website [1, 2]. GSE58294 (annotated with GPL570) has ninety-two peripheral blood samples including 69 IS patients and 23 healthy controls. The GSE16561 dataset (annotated with GPL6883) has 39 IS patients, and 24 healthy controls. GSE22255 has 20 IS patients and 20 healthy controls and was used as the external validated dataset.

Data management and cuproptosis-related DEGs identification. The stroke microarray data were processed, normalized and corrected with the Affy and RMA package; then, the batch effect between the two datasets were corrected with the Limma and sva package. The cuproptosis-related genes were summarized from previous literature reviews [3]. Principal component analysis (PCA) was performed to show the mixture of two datasets and eliminate batch effects. Differentially expressed genes (DEGs) between two groups were further identified with the Limma package [4]. DEGs were considered at their $\log_2$FC > 1 and adjusted *p*-value < 0.05.

GSVA analyses. Based on the "c2.cp.all.v7.0. symbols" gene set, we used the R package GSVA to calculate the scores of the relevant pathways according to the gene expression matrix of each sample by the method of ssGSEA, and relevant pathways between the two clusters were listed in a bar-plot figure, and adjusted *p*-value < 0.05 was set as statistically *significant*.

Immune infiltration analysis. CIBERSORT is a tool to evaluate the distribution of immune cells in cellular populations [5]. Then, the relationship between the significant immune cells and key CuDEGs were screened with the *p*-value < 0.05.

Consensus. Clustering is an unsupervised clustering method, which is used for the classification of disease subtypes. Based on these cuproptosis-related genes, the R package ConsensusClusterPlus (version 1.56.0) was used to perform consistent clustering of IS samples with Spearman method. The clustering algorithm is pam (Partitioning Around Medoids). Through matrix heatmap, two clusters were identified. The stroke probability, expression of cuproptosis markers and immune cell infiltration extent were further compared between the two clusters.

Weighted Gene Co-Expression Network Analysis (WGCNA). WGCNA is a computational tool to identify co-expressed genes between different groups. It can identify marker genes based on the non-orientation analysis between the gene set and phenotypes. Here, WGCNA was used to locate the gene modules related to different clusters in stroke.

Machine learning model for hub gene identification. We applied a machine learning (ML) model: general linear model (GLM), support vector machine (SVM), XGBoost (XGB) and recursive feature elimination (RFE) to identify the hub CuDEGs in stroke. The input was CuDEGs in stroke and we listed top 10 genes with high feature importance in each method. The predictive ability of each method was calculated according to the root mean square of residuals, the area under curve (AUC) value and reverse cumulative distribution of residuals. The high AUC value and high residual level indicated the higher predictive ability of the ML model.

Single Cell Sequencing Data obtained and processing. The sc RNA-seq data were downloaded from NCBI website: GEO174574. Then, we carried out quality control and cells with less than 10% of mitochondrial genes were captured, with a total number of genes ranged from 200 to 10000 and in at least three cells. We integrated all samples via SCT correction. Next, an uMAP method was used for the dimension reduction. The single cell-RNA sequencing method was used to map the hub genes and locate their cell source [6]. CellChat R package (1.4.0) was applied to explore the cell-cell interaction between the cell clusters. We also applied the monocle (2.22.0) method for pseudotime analysis to identify the stage and gene changes of microglia. We compared the cuproptosis extent with the AddModuleScore method between each specific cell.

Pseudotime analysis. The R package monocle v2.22.0 was used to do the pseudotime analysis. The cell status in different stages was demonstrated.

Receiver Operating Characteristic (ROC) analysis. To confirm the clinical significance of hub genes, the diagnostic value of the key CuDEGs were assessed with 'pROC', 'DALEX', 'randomForest', 'kernlab' and 'XGBoost' packages to construct the machine learning model. The calibration curve, bootstrap method and DCA were derived based on these analyses.

Statistical analysis. The statistical analysis was carried out in R v4.1.3. A *p*-value less than 0.05 was considered as statistically significant.

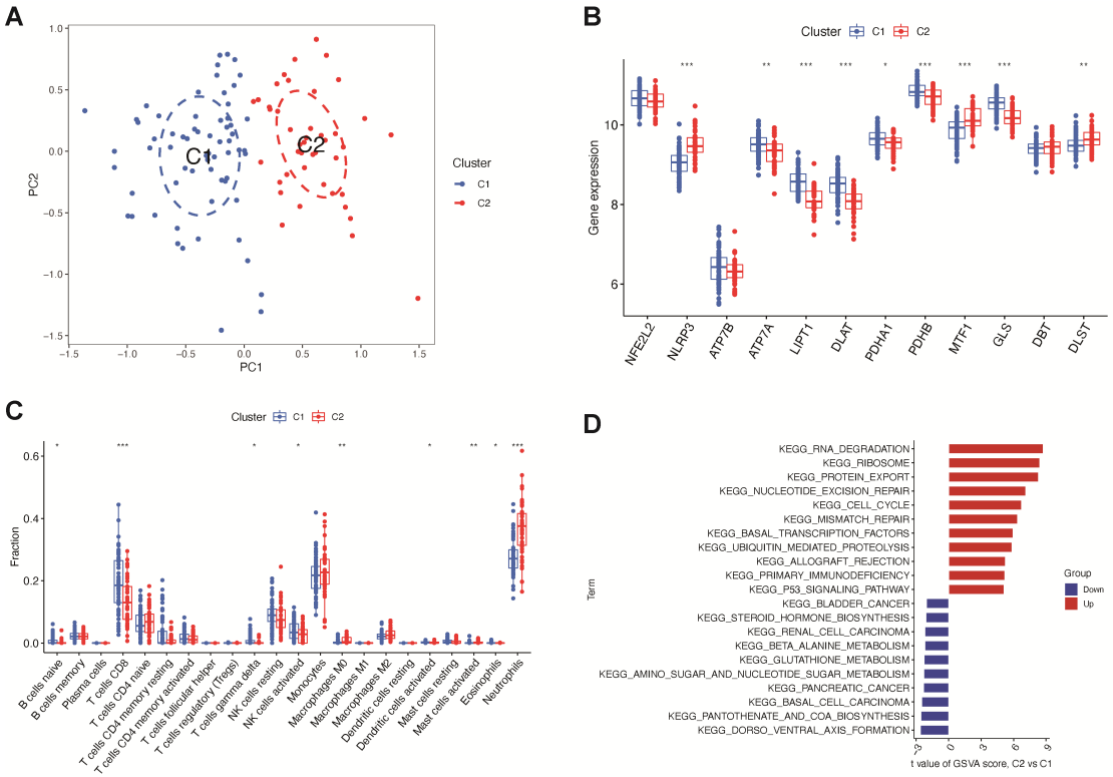
Discusion. In this study, four CuDEGs (LIPT1, PDHA1, DBT, DLST) might be associated with the probability of IS. LIPT1 transfers lipid salts to the E2 subunit of two-keto acid dehydrogenase and is involved in lipoic acid regulation, both PDHA1 and LIPT1 deficiency were related to the impaired TCA cycle [7], and the TCA cycle is thought to be truncated under hypoglycemic condition in stroke and glutamate oxaloacetate transaminase (GOT) can refill the TCA cycle and reduce the infarction lesion in both *in-vitro* and *in-vivo* stroke models [8].

DBT was also found to be a diagnostic gene in Parkinson's disease with an AUC at 0.717 [9]. Another study identified both DBT and DLST are significantly correlated with disease-free survival (DFS) and high expression of these genes predicted longer DFS in clear cell renal carcinoma as well [10]. However, the exact role of these CuDEGs in both neurological and non-neurological diseases need to be validated in further experiments.

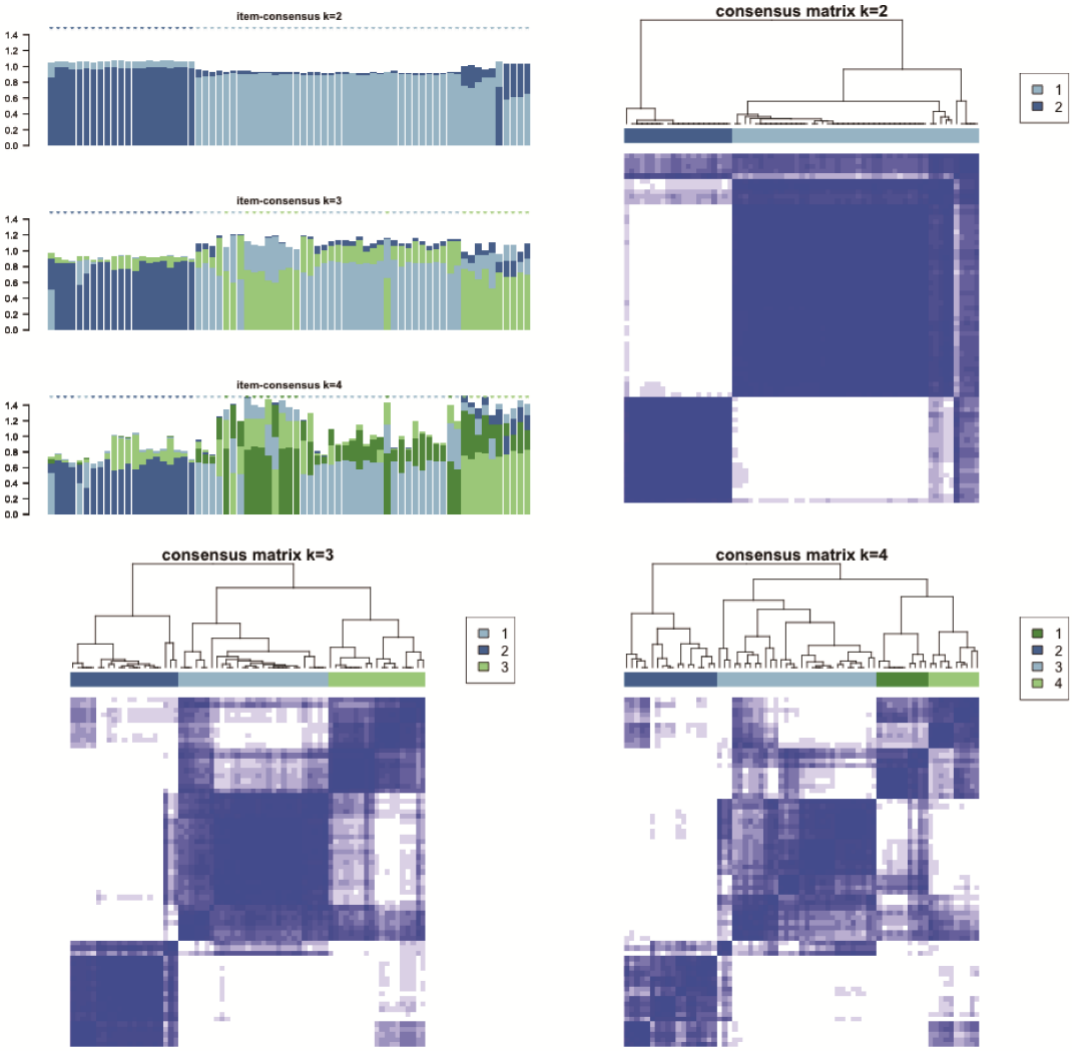
Grammer *et al.* reported that the plasma copper level is positively correlated with the occurrence of coronary heart diseases and related mortality [11]. In addition, Bagheri *et al*. also reported plasma copper was positively correlated with the severity of atherosclerosis [12]. A large study of 9588 participants concluded that higher serum copper levels have an increased incidence rate of hypertension with an odd ratio at 1.99 [13].

References

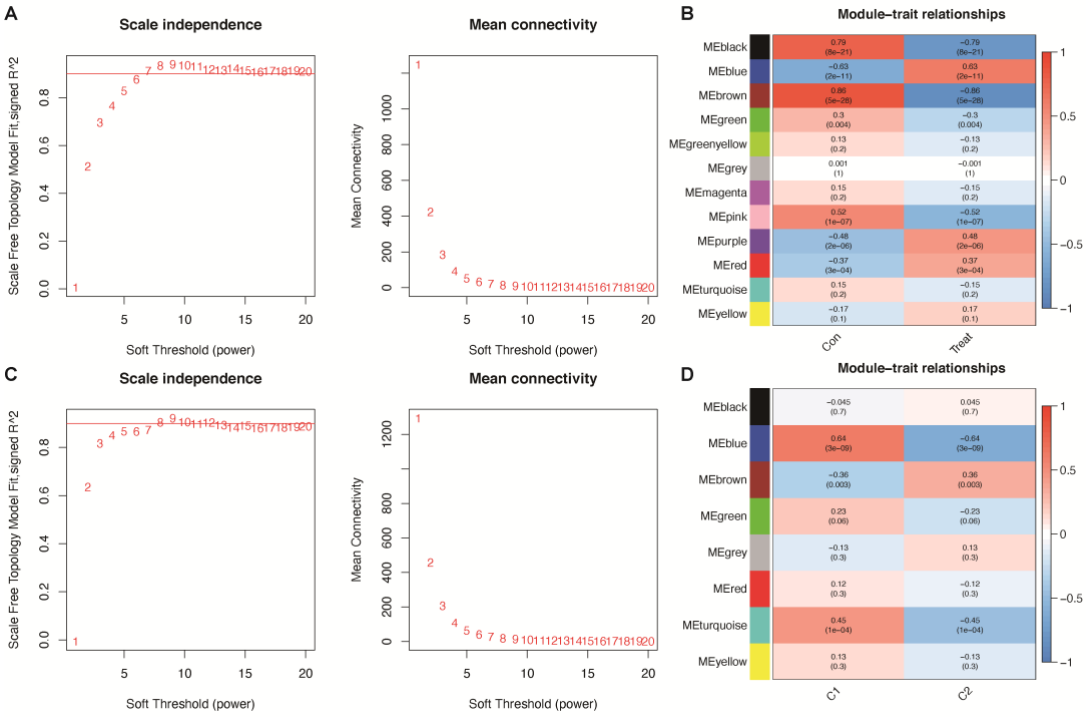
1. Stamova B, Jickling GC, Ander BP, et al. Gene expression in peripheral immune cells following cardioembolic stroke is sexually dimorphic. PLoS One 2014; 9: e102550.
2. O'Connell GC, Petrone AB, Treadway MB, et al. Machine-learning approach identifies a pattern of gene expression in peripheral blood that can accurately detect ischaemic stroke. Npj Genom Med 2016; 1: 16038.
3. Zhao S, Zhang L, Ji W, et al. Machine learning-based characterization of cuprotosis-related biomarkers and immune infiltration in Parkinson's disease. Front Genet 2022; 13: 1010361.
4. Yang W, Wu H, Tong L, et al. A cuproptosis-related genes signature associated with prognosis and immune cell infiltration in osteosarcoma. Frontiers Oncol 2022; 12: 1015094.
5. Lin JZ, Lin N. A risk signature of three autophagy-related genes for predicting lower grade glioma survival is associated with tumor immune microenvironment. Genomics 2021; 113: 767-77.
6. Li X, Liao Z, Deng Z, et al. Combining bulk and single-cell RNA-sequencing data to reveal gene expression pattern of chondrocytes in the osteoarthritic knee. Bioengineered 2021; 12: 997-1007.
7. Solmonson A, Faubert B, Gu W, et al. Compartmentalized metabolism supports midgestation mammalian development. Nature 2022; 604: 349-53.
8. Rink C, Gnyawali, S, Stewart R, et al. Glutamate oxaloacetate transaminase enables anaplerotic refilling of TCA cycle intermediates in stroke-affected brain. FASEB J 2017; 31: 1709-18.
9. Zhao S, Zhang L, Ji W, et al. Machine learning-based characterization of cuprotosis-related biomarkers and immune infiltration in Parkinson's disease. Front Genet 2022; 13: 1010361.
10. Guo L, An T, Wan Z, et al. Identification of the cuproptosis related prognostic gene signature and the associated regulation axis in clear cell renal cell carcinoma. Res Square 2022; https://doi.org/10.21203/rs.3.rs-1815139/v1.
11. Grammer TB, Kleber ME, Silbernagel G, et al. Copper, ceruloplasmin, and long-term cardiovascular and total mortality (The Ludwigshafen Risk and Cardiovascular Health Study). Free Radical Res 2014; 48: 706-15.
12. Bagheri B, Akbari N, Tabiban S, et al. Serum level of copper iin patients with coronary artery disease. Niger Medical J J Niger Medical Assoc 2015; 56: 39-42.
13. Darroudi S, Saberi-Karimian M, Tayefi M, et al. Association between hypertension in healthy participants and zinc and copper status: a population-based study. Biol Trace Elem Res 2019; 190: 38-44.



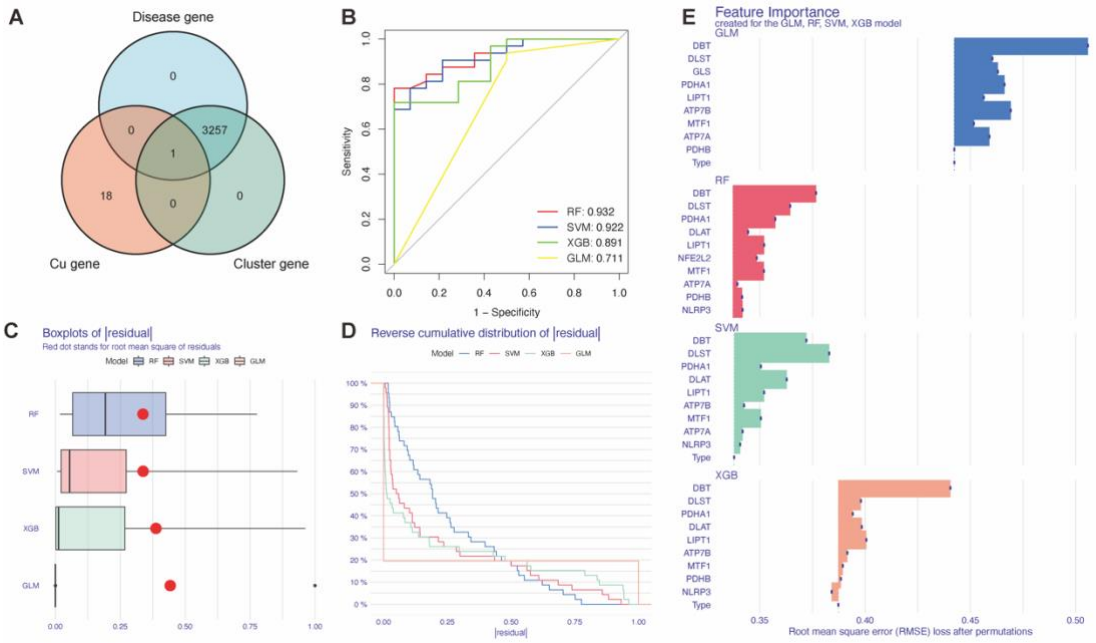
Supplementary Figure S1. The clusters of IS shows the heterogeneous cuproptosis. (**A**) The PCA shows the two cluster in IS groups. (**B**) Box plot shows the expression of hub CuDEGs between the two groups. (**C**) Box plot shows the immune infiltration status between the two groups. (**D**) The GSVA analysis of the enriched pathways between the two clusters



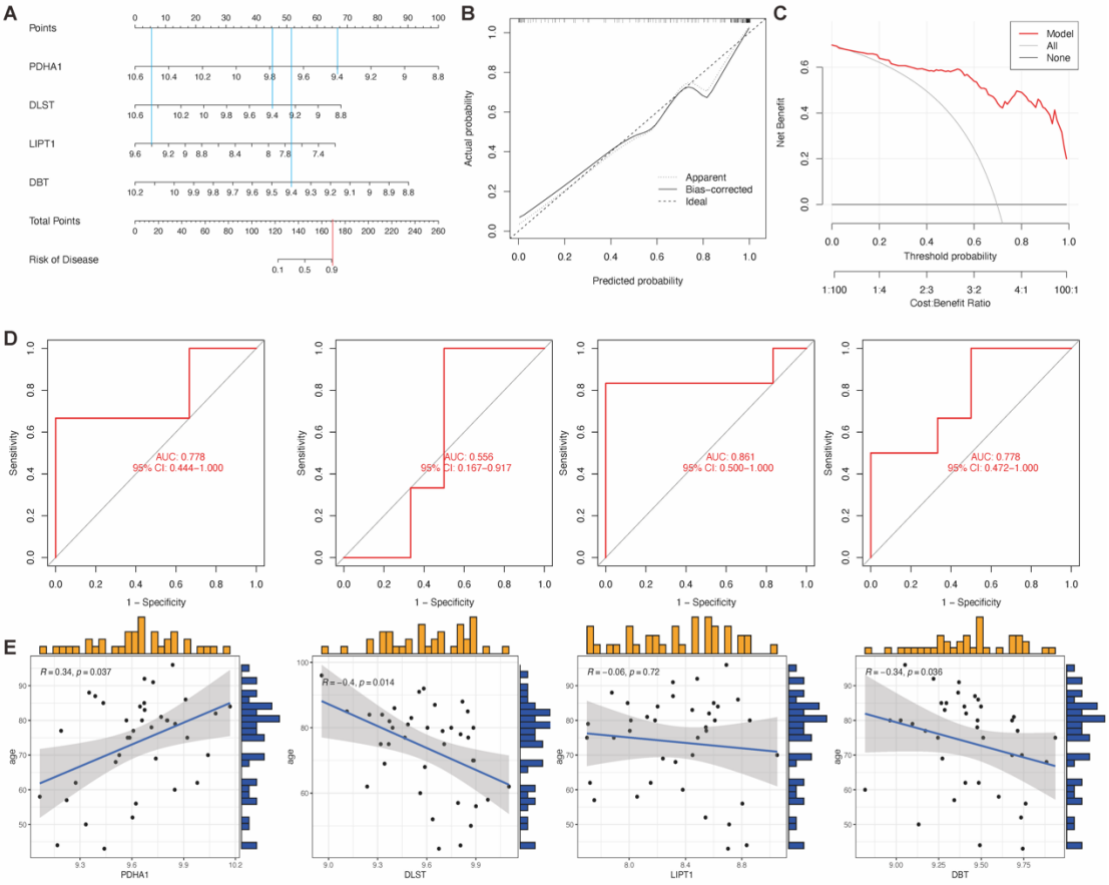
Supplementary Figure S2. The clustering maps and consensus matrix of k = 2,3,4 calculated by ConsensusClusterPlus R package.



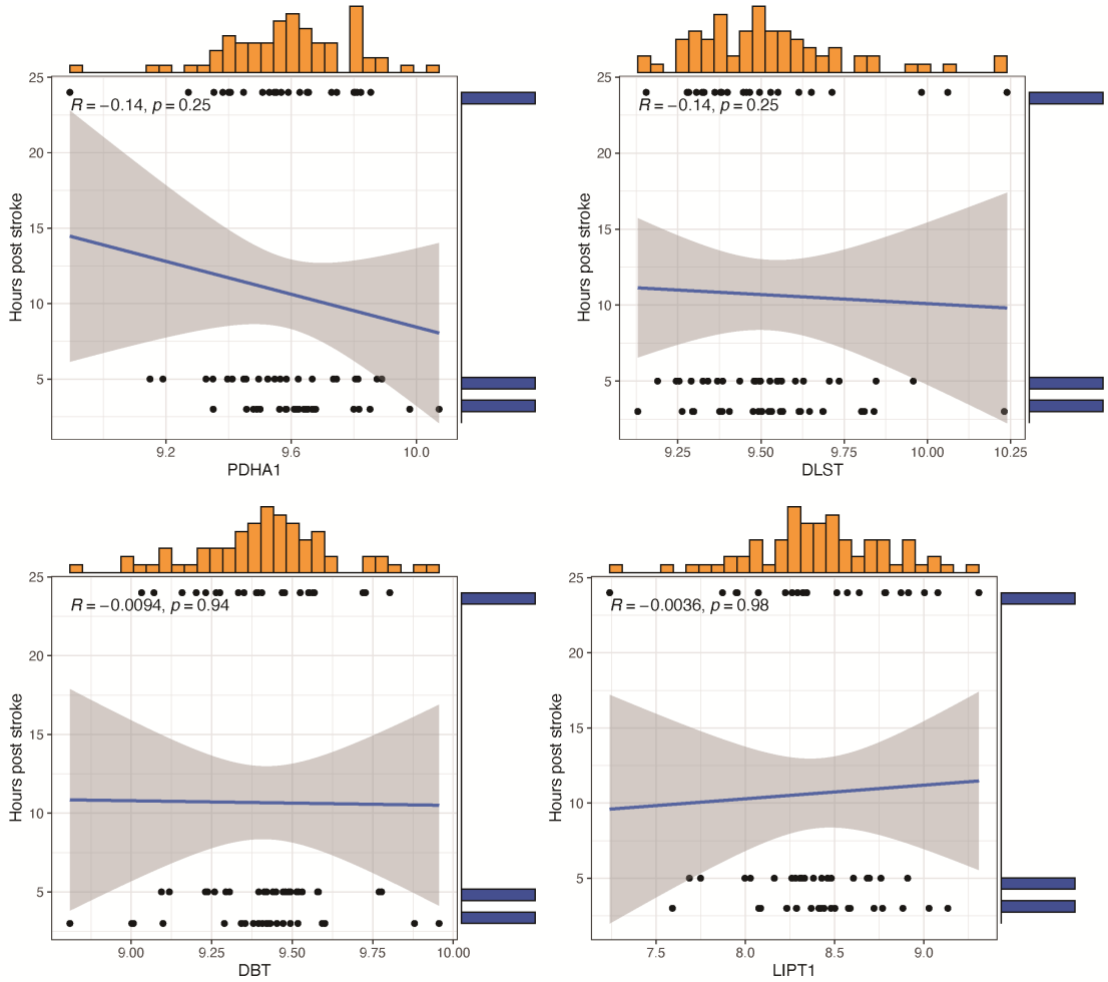
Supplementary Figure S3. The scale independence, mean connectivity and module trait relationships for the disease-based WGCNA and cluster-based WGCNA. The red line shows the selected threshold.



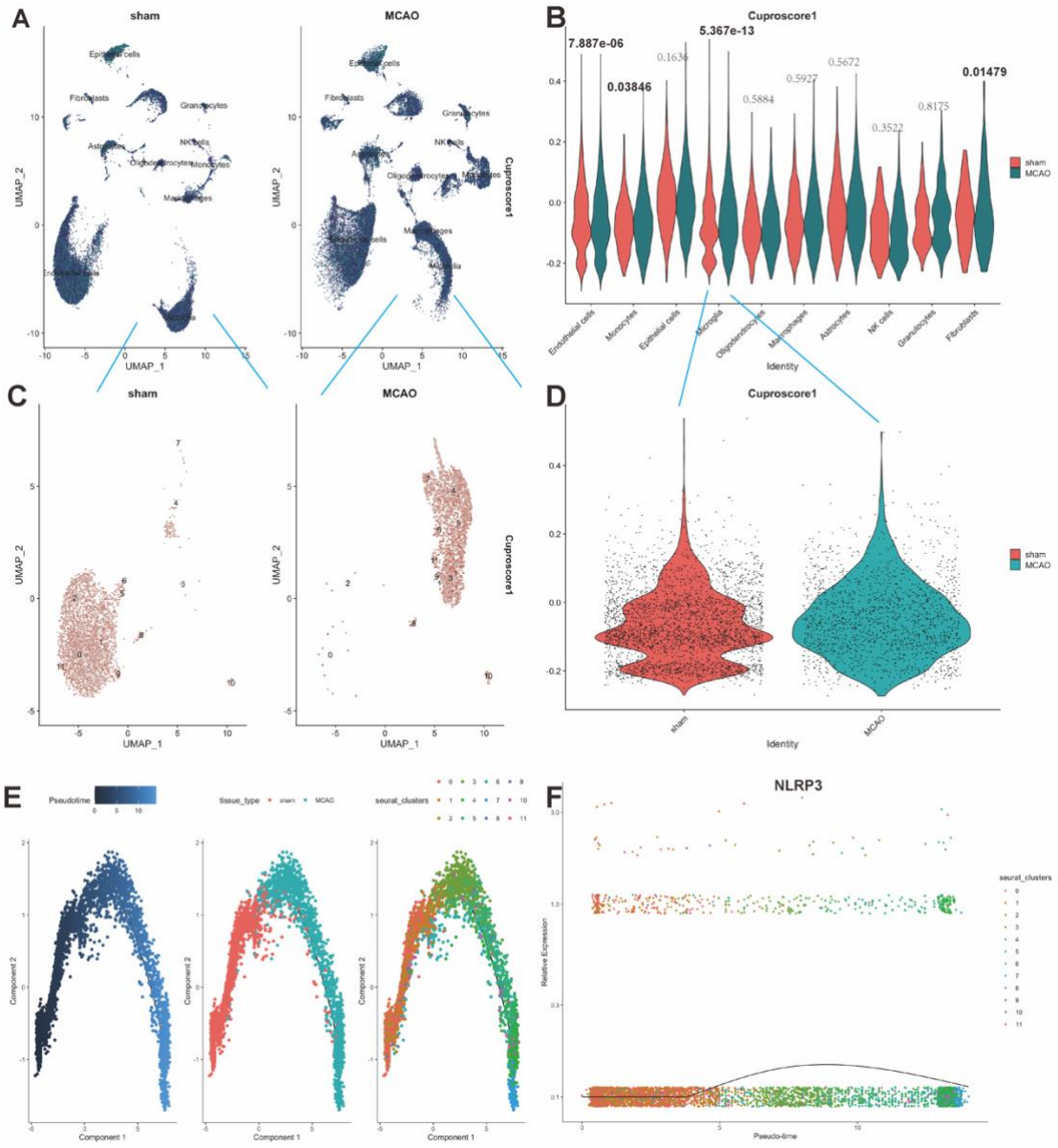
Supplementary Figure S4. The machine learning model for CuDEGs in stroke. According to the criteria for screening, WGCNA was used to locate the co-expressed genes and clinic characteristics. Results indicated that twelve modules were obtained. The correlation between the 12 modules and two groups was investigated by calculating the correlation co-efficiency between ME values and clinical features, which indicated that the blue module was positively associated with stroke phenotype (r = 0.63, p = 2e – 11), and negative relationships were found in black and brown modules (r = -0.79, p = 8e – 21; r = -0.86, p = 5e-28). We also applied the WGCNA analysis in the cluster analysis and found eight different modules between the two clusters. These findings suggested that the brown module was positively associated with cluster two (r = 0.36, p = 0.003), and negative relationships were found between blue and turquoise modules as well (r = -0.64, p = 3e – 9; r = -0.45, p = 1e-04). (**A**) The intersected genes among the disease-related, cluster-related and cuproptosis -related genes. (**B**) The ROC of the risk model constructed by CuDEGs. (**C**, **D**) The box plot and curve show the root mean square of residuals. (**E**) - The bar plots show the feature importance of top 10 genes in each ML method.



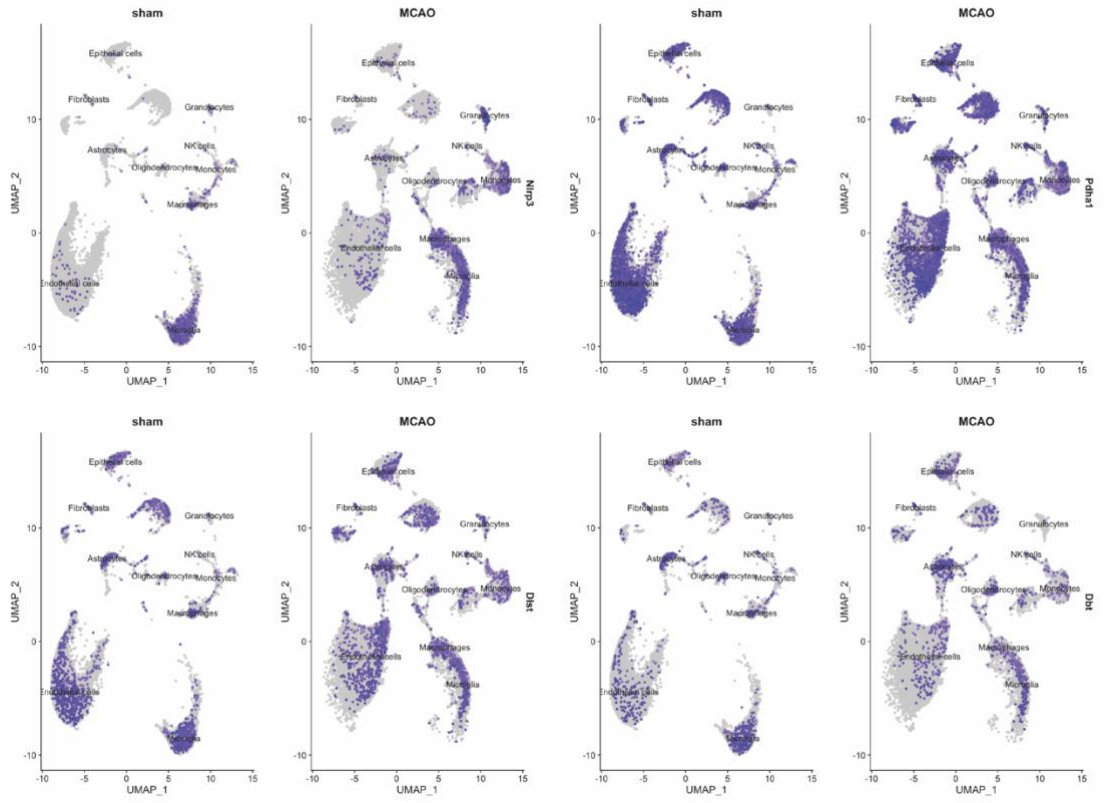
Supplementary Figure S5. The risk model constructed by CuDEGs in stroke. (**A**) The nomogram for diagnostic value of stroke patients was developed in accordance with four CuDEGs in stroke. (**B**, **C**) The calibration curve and DCA of the risk model constructed by four CuDEGs. (**D**) The validation of the risk model in the external validation dataset (GSE22255) for RF, SVM, XGB and GLM. (**E**) The scatter plot for the gene expression of four CuDEGs and patients' age



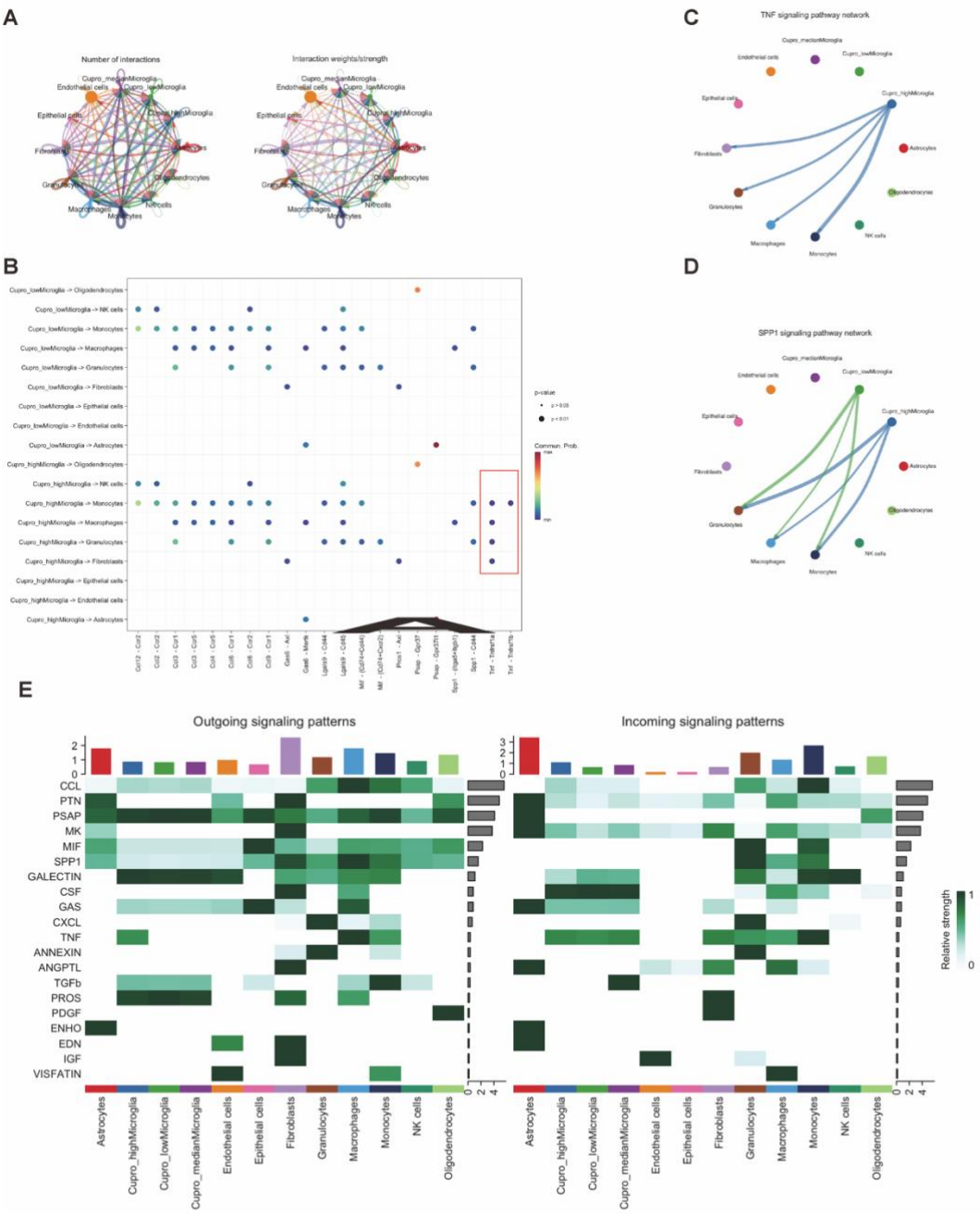
Supplementary Figure S6. The scatter plot for the gene expression of four CuDEGs and patients' age.



Supplementary Figure S7. The cuproptosis scores between MCAO mouse and sham mouse. (**A**) The cuproptosis score between the two groups assessed by addmodulescore. (**B**) Violin plot shows the quantitative comparison of cuproptosis score in each cell cluster between MCAO and sham group. The p-value of each comparison by Wilcox test was listed and bold meant the statistical difference. (**C**) The sub-cluster for microglia between the two groups. (**D**) The comparison of cuproptosis in microglia. (**E**) The pseudo-time analysis of the microglia sub-cluster. (**F**) The pseudo-time analysis shows the hub gene NLRP3 in the stage transition in microglia



Supplementary Figure S8. The expression of Nlrp3, Lipt1, Dbt and Dlst in sham and MCAO mouse in sc-RNA seq data.



Supplementary Figure S9. Cell-cell interaction at single-cell level in MCAO. (**A**) The number of interactions and interaction weights between several cell types. (**B**) The bubble plot shows the relationship between cell types and ligand-receptor. (**C**, **D**) TNF and SPP1 signaling pathway network between several cell types. (**E**) The outgoing and incoming signaling patterns in different cell clusters with related signaling pathways. The relative strength is shown from shallow to deep color