Five Genes as Diagnostic Markers of Ischemic Stroke and Their Correlation with Immune Infiltration

Meng Wang  
Hwa Mei Hospital, University of Chinese Academy of Science, Ningbo, Zhejiang, PR China

Ruijie Zhang  
Department of Global Health, Ningbo Institute of Life and Health Industry, University of Chinese Academy of Science, Zhejiang, PR China

Qiongfeng Guan  
Hwa Mei Hospital, University of Chinese Academy of Science, Ningbo, Zhejiang, PR China

Yindan Yao  
Hwa Mei Hospital, University of Chinese Academy of Science, Ningbo, Zhejiang, PR China

Liyuan Han (hanliyuan@ucas.ac.cn)  
Hwa Mei Hospital, University of Chinese Academy of Science, Ningbo, Zhejiang, PR China

Research Article

Keywords: Gene Expression Omnibus, ischemic stroke, diagnostic, CIBERSORT, immune cells

Posted Date: October 22nd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-936536/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. 
Read Full License
Five Genes as Diagnostic Markers of Ischemic Stroke and Their Correlation with Immune Infiltration

Meng Wang, M.M.¹,²*, Ruijie Zhang, M.M.², Qiongfeng Guan, M.M.¹, Yindan Yao, M.D.¹, Liyuan Han, M.D., PhD.¹,²,Δ.

¹Hwa Mei Hospital, University of Chinese Academy of Science, Ningbo, Zhejiang, PR China, 315000; ²Department of Global Health, Ningbo Institute of Life and Health Industry, University of Chinese Academy of Science, Zhejiang, PR China, 315000.

*First author.
ΔCorresponding author.

To whom correspondence should be addressed, Prof. Liyuan Han, Hwa Mei Hospital, University of Chinese Academy of Science, No.159, BeiJiao Road, JiangBei District, Ningbo, Zhejiang, PR China, 315000. Tel: +86-057487609020; Email: hanliyuan@ucas.ac.cn.

Conflict of Interest Statement: The authors who have taken part in this study declared that they do not have anything to disclose regarding funding or conflict of interest with respect to this manuscript.

Author Contributions: Liyuan Han and Meng Wang had full access to the data of study, and takes responsibility for the integrity of the data. Meng Wang and Ruijie Zhang performed the check of the bioinformatic analysis Yindan Yao and Qiongfeng Guan perform the check the integrality of data and wrote the manuscript. Meng Wang performed the submission of manuscript.

Prior Presentations: This work has not been presented previously.

Key words: Gene Expression Omnibus, ischemic stroke, diagnostic, CIBERSORT, immune cells
Five Genes as Diagnostic Markers of Ischemic Stroke and Their Correlation with Immune Infiltration

Abstract: Background: This study aimed to identify potential diagnostic markers of ischemic stroke (IS) and discuss the function of immune cell infiltration during the pathological process. Methods: We used IS datasets from the Gene Expression Omnibus database. Differentially expressed genes (DEGs) were identified, and functional correlation analysis was performed. We then screened and verified the diagnostic markers of IS. We evaluated the infiltration of immune cells in infarcts using CIBERSORT and analyzed the correlation between diagnostic markers and infiltrating immune cells. Results: A total of 366 DEGs were screened in this study. Genes encoding CTSG, F13A1, PABPC1, ECHDC2, BIRC2 and infiltrating monocytes, M0 macrophages, activated dendritic cells, and neutrophils (area under the curve [AUC] = 0.945) were identified as diagnostic markers of IS. Immune cell infiltration analysis suggested that memory B cells, regulatory T cells, M0 macrophages, CD8+ T cells, γδT cells, activated natural killer cells, monocytes, activated mast cells, and neutrophils were involved in the IS process. Analysis of correlations between expressed genes and infiltrating immune cells found that CTSG was positively associated with M0 macrophages, F13A1 was positively associated with monocytes, PABPC1 was positively associated with activated dendritic cells, eosinophils were negatively associated with neutrophils, ECHDC2 was negatively associated with monocytes, and BIRC2 was positively associated with eosinophils. Conclusion: five genes and four types of immune cells were identified as diagnostic markers of IS, and immune cell infiltration may play an important role in the progression of IS.

Keywords: Gene Expression Omnibus; ischemic stroke; diagnostic; CIBERSORT; immune cells

1. Introduction

Stroke is characterized by a high rate of disability (> 80% of patients) and is the second most common cause of mortality worldwide, after cancer [1,2]. As an acute cerebrovascular disease, stroke causes significant social and economic burdens annually [3].

To date, the diagnosis of IS is primarily based on scales, such as the National Institute of Health stroke scale (NIHSS), and imaging findings. However, early alterations in ischemic brain regions may not be detected using these technologies due to limited sensitivity [4, 5]. Thus, there is an urgent need for a new diagnostic method that is either minimally invasive or noninvasive and has high accuracy and specificity for improving the early diagnosis of IS.

Several studies have emphasized that immune cells may play a significant role in ischemic regions during the occurrence and development of stroke [6,7,8,9,10,11]. Neutrophils are attracted by chemokines and reach the ischemic lesions first [12]. Subsequently, secondary injury is mediated by neutrophils cause via releasing matrix metalloproteinases (MMPs), proteases, reactive oxygen species (ROS) and other proinflammatory factors [6]. Furthermore, adherence of neutrophils to the endothelium causes the cerebral no-reflow phenomenon and the simultaneous production of MMPs, ROS, and proteases which
damage vessels and ischemic tissues [7]. T lymphocytes, such as regulatory T (Treg) cells, CD8+ and CD4+ effector T cells and γδT cells, infiltrate the brain parenchyma at a later stage, and Tregs have shown to have a neuroprotective function [8,9]. As a primary component of brain tissue, glial cells also take a part in the post-IS immune reaction.

Several types of immune cells have a profound influence on the development of ischemic injury in brain tissue. Therefore, investigations on the infiltration of these cells and the expression matrices of specific genes can facilitate the early diagnosis of IS [8,9]. In this study, as an initial step toward identifying potential diagnostic biomarkers, we screened hub genes with differential expression between patients and healthy participants from an RNA sequencing (RNA-seq) dataset. For the first time in IS patients, we quantified 22 types of immune cells [13] and conducted the correlative analysis between these cells and hub gene expression. Subsequently, several types of immune cells that exhibited the strongest correlations with gene expression were chosen for the second step in the identification of potential diagnostic biomarkers. We built a diagnostic predictive model based on these diagnostic features to identify IS in an individual and conducted related enrichment and protein–protein interaction (PPI) analyses.

2. Materials and Methods

2.1. Stroke Data acquisition and Preprocessing

GEOquery (version 2.58.0) [14] was used to download the stroke-relate gene expression datasets GSE16561 [15] and GSE22255 [16] from Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/) [17].

Raw data from both datasets were processed using the Affy software package containing Robust Multi-Array Average (RMA) algorithm [18], and gene expression profiling by an array was combined into one dataset using the sva on the R platform [19] to minimize batch effects. Two principal component analysis (PCA) plots were drawn to test batch effects and inter-sample correction using TBtools software [20].

2.2. Display of Differential Gene Expression and Enrichment Analysis

R package limma [21] was used to identify DEGs, between stroke patients and healthy people, via estimating gene expression differences based on an empirical Bayesian approach. The R package ggplot2 [22] was used to draw a volcano plot showing DEGs that met the significance criteria (|log2FC| > 1.3, adjusted p-value < 0.05).

Then Gene Ontology (GO) enrichment analysis was conducted, via using package clusterProfiler [23], as well as Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analysis. The gene set enrichment and PPI analyses were conducted online using Metascape (http://metascape.org) [24].

2.3. Selection and Identification of Stroke Diagnostic Markers

Support vector machine-recursive feature elimination (SVM-RFE) [25] combined with Least absolute shrinkage and selection operator (LASSO) logistic regression [26] were used to single out variables that were significant for diagnosing stroke. The R packages e1071 [27] and glmnet [28] were used respectively to implement LASSO-logistic regression and the SVM-RFE algorithm. We selected variables from both the LASSO and SVM-RFE results as diagnostic markers.
2.4. Quantitative Analysis of Immune Cell Infiltration

CIBERSORT [13], a computational approach, was used to estimate the presence of all types of immune cells in merged stroke transcriptomes according to filtration criteria (p-value < 0.001). A PCA plot was drew, via using package ggplot2 [29], which revealed differences in the expression of genes related to infiltrating immune cells between stroke patients and healthy individuals, and a correlation plot was drew, via using the ggcor [30] package, which can visualize the co-expression of genes related to all types of immune cells. A violin plot was also drawn using the R package ggplot2 [29] to display the quantities of immune cell matrices between the two groups.

2.5. Correlations between Stroke Diagnostic Markers and Immune Cells

We used the R package ggcor [30] to identify correlations between all types of immune cells and stroke diagnostic biomarkers, and ggplot2 [29] was used to draw a lollipop chart to illustrate the results.

3. Results

3.1 Batch Effect Minimization and Differential Gene Expression Analysis

First, we merged two GEO datasets into one matrix following background adjustment and normalization. The effects of the likelihood of batch effects due to non-biological technical bias minimization were displayed in PCA cluster plots. Data from different GEO documents were inequivalent before normalization and batch effect removal (Figure 1A). These data from diverse matrices are comparable depending on the sva package that can support subsequent analyses (Figure 1B). After preliminary data processing, we identified 366 DEGs by comparing gene expression between stroke patients and healthy individuals, which are shown in the volcano diagram in Figure 1C.

**Figure 1.** PCA diagram before and after batch effect removal and volcano map of DEGs. (A) PCA cluster diagram of the GSE16561 and GSE22255 datasets before batch effect removal. (B) PCA diagram of the GSE16561 and GSE22255 datasets after batch effect removal. (C) Volcano plot of DEGs; red, grey and blue respectively represents up-regulated DEGs, no significant difference genes and down-regulated DEGs.

3.2. Enrichment Analysis

The GO enrichment analysis results showed that the DEGs were mainly concentrated in the annotations of neutrophil degranulation, neutrophil activation involved in immune responses to biological process, the secretory granule membrane, the secretory granule lumen, the cytoplasmic vesicle lumen, the vesicle lumen of the cellular component, and phosphoric ester hydrolase activity of molecular function (Figure 2A and B). The most significant functional pathways were all biological processes: myeloid cell differentiation, neutrophil activation involved in immune responses, neutrophil degranulation, response to lipopolysaccharide, and response to molecules of bacterial origin (Figure 2C).

The KEGG analysis revealed that the DEGs were mainly associated with hematopoietic cell lineages, the interleukin 17 (IL-17) signaling pathway, the nuclear factor (NF)-kappa B signaling pathway, rheumatoid arthritis, and transcriptional misregulation in cancer (Figure 3A, B, and C).

PPI network analysis was carried out via using MCODE algorithm that included in Metascape. The
sub-network of PPI which ranked first was made up of 30 genes including CXCL1, PPBP, CXCL2, ELANE, CXCL8, CCR7, ITGAM, MCEMP and so on (Supplementary Figure 1). It supported that neutrophil activation, granulocyte activation and neutrophil degranulation is of great importance in stroke pathological process as functional pathways. Summarily, immune activation may participates in stroke pathological progress.

**Figure 2.** Results of GO enrichment analysis. (A) Bubble plots where horizontal axis represents generation of each pathways included in three modules (BP, CC and MF). (B) Bar plots where horizontal axis represents numbers of genes under each pathways that can be classified into three modules (BP, CC and MF). (C) Circle diagram which shows the general situation of enrichment processes.

**Figure 3.** Results of KEGG enrichment analysis. (A) Bubble plots where horizontal axis represents gene ratio of each pathways. (B) Bar plots where abscissa displays the number of genes under each pathways. (C) Circle diagram which shows the general situation of enrichment processes.

### 3.3. Identification of Potential Diagnostic Markers

The LASSO-logistic algorithm screened out 20 genes to be used for subsequent validation (Figure 4A and B), seven genes, of which five were included in the former 20 genes, were identified via SVM-REF (Figure 4C). For both methods, the training set was composed of samples from both the stroke patient and healthy individual groups with even-numbered GSM codes, whereas the testing set contained samples with odd-numbered codes. The Venn diagram shows that the overlap of the results from the two methods contained five genes (cathepsin G [CTSG], F13A1, PABPC1, ECHDC2, and BIRC2), which were selected as our potential diagnostic markers (Figure 4D).

**Figure 4.** Screening process of potential diagnostic markers. (A) The LASSO coefficient profile of mRNAs. (B) The tuning parameters (log λ) of IS related mRNAs were used to cross-validate the error curve. Perpendicular imaginary lines were plotted at the optimal value According to the minimal criterion and 1-se criterion. (C) Process of screening potential diagnostic markers via SVM-RFE algorithm. (D) The common part of diagnostic markers screened by two algorithms is shown by venn diagram.

### 3.4. Immune Cells Quantization and Correlation Analysis

A three-dimensional PCA diagram was used to visually reveal the different features of immune infiltration between IS patients and healthy ones. By using this method, we observed significantly different patterns of immune cells infiltration between the groups (Figure 5A and Figure 5B). The correlations between 22 types of immune cells are presented in a heatmap diagram (Figure 5C) which shows that the strongest positive correlation between follicular helper T cells and activated mast cells. The strongest negative correlation was observed between inactive CD4+ memory T cells and inactive natural killer (NK) cells. A violin plot of horizontal comparisons of the immune cell expression matrices between the groups showed higher levels of CD8+ T cell, memory B cell, Treg cell, and inactive mast cell infiltration and lower levels of γδT cell, M0 macrophage, activated mast cell, and neutrophil infiltration in the stroke patient group compared with the healthy individual group (Figure 5D).
Figure 5. Analysis of characteristics of immune infiltration. (A) Three-dimensional PCA cluster diagram of immune infiltration between IS group and normal group. (B) Three-dimensional PCA diagram of two groups that each one has a distribution ellipse. (C) Correlation plot of immune cells. The size of the eccentricity of each ellipse represents the strength of the correlation; red represents a negative correlation, blue represents a positive correlation. The lighter the color, the weaker the correlation. (D) Violin plot of the quantities of 22 types of immune cells between two groups.

3.5. Calculation of the Multivariate Fitted Prognostic Factor

The asymmetrical diagram of correlations between the 22 immune cell types and five potential prognostic genes showed that there were 5 kinds of immune cells which were obviously correlated with at least one gene (Figure 6A). The 5 kinds of immune cells identified as diagnostic predictors were M0 macrophages, monocytes, neutrophils, eosinophils and activated dendritic cells. A multivariable combined receiver operating curve analysis of the testing set was conducted by combining the five prognostic genes with one or more immune cell types, and a fitted factor with the highest area under curve (AUC) value, specificity, and sensitivity was determined. When 4 kinds of immune cells (M0 macrophages, neutrophils, activated dendritic cells, and monocytes) and 4 potential genes were fitted into one factor, an AUC value of 0.997 was calculated in the training set (Figure 6B and C), and a maximum AUC of 0.945 was reached in the testing set (Figure 6D and E).

Figure 6. Calculation of the best multi-variate fitted prognostic factor. (A) Asymmetrical diagram of correlations between 27 potential biomarkers (5 genes and 22 kinds of immune cells). (B) ROC curve of the efficacy of 5 diagnostic genes and fitted factor in train set. (C) Evaluation of efficacy of 4 diagnostic prognostic types of cells and fitted factor in train set. (D) Evaluation of efficacy of 5 diagnostic genes and fitted factor in test set. (E) Evaluation of efficacy of 4 diagnostic types of cells and fitted factor in test set.

3.6. Correlative Relationships between Potential Diagnostic Genes and Immune Cells

A correlation analysis revealed that expression of BICR2 was correlated with quantity of activated dendritic cells ($r = 0.289, p = 0.026$) and eosinophils ($r = 0.337, p = 0.0009$; Figure 7A) positively. CTSG expression was correlated with infiltration of activated NK cells ($r = -0.261, p = 0.046$), activated dendritic cells ($r = -0.316, p = 0.015$), and naive B cells ($r = -0.356, p = 0.006$) negatively and correlated with infiltration of M1 ($r = 0.314, p = 0.015$) and M0 macrophages ($r = 0.507, p = 0.00004$; Figure 7B) positively. ECHDC2 expression was correlated with infiltration of inactive dendritic cells ($r = -0.260, p = 0.047$) and monocytes ($r = -0.522, p = 0.00002$) negatively and correlated with infiltration of naive B cells ($r = 0.316, p = 0.015$; Figure 7C) positively. F13A1 expression was correlated with infiltration of CD8$^+$ T cells ($r = -0.267, p = 0.041$), activated NK cells ($r = -0.272, p = 0.037$), activated mast cells ($r = -0.286, p = 0.028$), and follicular helper T cells ($r = -0.336, p = 0.009$) negatively and correlated with infiltration of monocytes ($r = 0.469, p = 0.00002$; Figure 7D) positively. PABPC1 expression was correlated with infiltration of plasma cells ($r = -0.260, p = 0.046$), naive CD4$^+$ T cells ($r = -0.329, p = 0.011$), and neutrophils ($r = -0.336, p = 0.009$) negatively and correlated with infiltration of eosinophils ($r = 0.339, p = 0.009$) and activated dendritic cells positively ($r = 0.409, p = 0.001$; Figure 7E).
Figure 7. Correlation between 5 potential diagnostic genes and 22 types of immune cells. (A) Correlative relationship between BIRC2 and 22 type of immune cells. (B) Correlative relationship between CTSG and 22 type of immune cells. (C) Correlative relationship between ECHDC2 and 22 type of immune cells. (D) Correlative relationship between F13A1 and 22 type of immune cells. (E) Correlative relationship between PABPC1 and 22 type of immune cells. The color of dot as well as the color of dot represents the correlative relationship. The color is closer to deep green, the correlative relationship is stronger and the color is closer to yellow, the correlation is weaker. The bigger the size of dot, the stronger the correlation, the smaller the size, the weaker the correlation.

4. Discussion

In this study, we used an RNA-seq expression matrix to construct a diagnostic prediction model for IS by making use of machine learning algorithms and analyzed the characteristics of associated immune cell infiltration. Following two rounds of screening, we found five potential diagnostic genes between people who have IS and healthy ones and four types of immune cells that contributed to the diagnostic model. Results of subsequent enrichment and PPI analyses indicated that immune response processes and pathways related to neutrophils are involved in the pathophysiological alterations of IS.

Stroke is an acute cerebrovascular disease that is characterized by significant disability and causes considerable social and economic burdens annually [3]. Recently, several studies have implicated glial cells, an intrinsic cell population in brain tissue that includes microglia, astrocytes, and oligodendrocytes, in post-stroke immune alterations. Activated microglia play the same pathophysiological role as macrophages following ischemia and can induce neuronal injury by producing ROS via nicotinamide adenine dinucleotide phosphate oxidase [31]; these cells also produce cytokines [32] and MMPs [33,34]. Another possible mechanism of inflammation is the expression of CD14 on microglia and subsequent activation by toll-like receptor 4, which triggers microglial activation [36,37]; interferon produced via this process induces the expression of microglial genes that contribute to ischemia/reperfusion injury [38]. Surprisingly, microglia may also play a bidirectional role in ischemia/reperfusion by secreting anti-inflammatory factors [39] and preventing the secretion of inflammatory factors related to astrocytes [40].

In the acute stage of IS, astrocytes participate in diverse functional processes, which include limiting brain tissue damage caused by ischemia, reducing the inflammatory response, maintaining nervous system homeostasis in the tissue around infarct region, and facilitating repair of the blood–brain barrier [41,42]. Furthermore, astrocytes are crucial for axonal regeneration and functional recovery during the chronic post-stroke stage. Through their interactions with other adaptive immune cells including NK cells and CD8+ T cells, astrocytes not only increase the infiltration of these immune cell populations but also enhance their function, which aggravates cerebral injury.

Another major part of innate immune system in the brain is the oligodendrocyte, which participates primarily in the cerebral infarction repair and significantly influences recovery. In contrast to the other two types of glial cells, the proliferation of oligodendrocytes cannot solely rely on itself without the existence of oligodendrocyte progenitor cells [43]. Research has found that various types of immune cells, such as macrophages, neutrophils, and T lymphocytes, are also of significance in the pathological
alterations of IS [6,8,9,10,11], and this information has aided in the identification of new biomarkers (e.g., from RNA-seq data) for the early diagnosis of IS. In this study, we searched for genes that could act as key diagnostic biomarkers and investigated their functional features related to pathways in IS.

The results that majority of DEGs were enriched in neutrophil degranulation, neutrophil activation of immune response, myeloid cell differentiation, and T cell activation as a biological process are produced by GO enrichment analysis. The results that the NF−kappa B signaling pathway, IL−17 signaling pathway, hematopoietic cell lineage, rheumatoid arthritis, and transcriptional misregulation in cancer were significant pathways in IS are produced by KEGG enrichment analysis. The results of a PPI analysis using MCODE suggested that pathways involving neutrophils and B cell receptors were important [44]. These results confirm the importance of immune cell infiltration in IS.

We used both the LASSO-logistic model and SVM-RFE to screen potential biomarkers, which yielded in five genes: CTSG, F13A1, PABPC1, ECHDC2, and BIRC2. CTSG takes part in the development of dermatomyositis myoidem, which is possibly mediated by inflammatory cytokines [45].

CTSG, which is released by activated neutrophils, was shown to be positively correlated with the development of cardiovascular and cerebrovascular diseases by mediating platelet activation, as a serine protease [46]. CTSG cleaves of the glycoprotein (GP) Ib [47] and mediates the releasing of plasminogen activator inhibitor-1 [48], which promotes intravascular thrombosis. A previous animal experiment reported that CTSG-like antigen was highly expressed in vivo around cerebral wound margins and analogous phenomenon was also observed in cultured primary astrocytes. Astrocytes may take part in a series of anti-inflammatory biological processes by secreting CTSG-like antigens [49]. Meanwhile, a separate study reported that serine protease inhibitor alpha 1-antichymotrypsin, which is mainly released by astrocytes, can form complex with CTSG that will limit the deterioration of cerebrovascular diseases [50]. Another study found that CTSG can suppress the thrombin-induced highly-expression of inducible nitric oxide, via upregulating release of NO, which is the activation signal of astrocytes [51].

F13A1, which is also known as coagulation factor XIII subunit A, not only promotes thrombosis but also limits both thrombus growth by suppressing embolus growth [49] after coronary angioplasty and the risk of stroke in young female [52]. Another study discovered a novel variant of F13A1 in a child with hemorrhagic stroke [53]. PABPC1 contributes to various physiological processes, such as by facilitating the binding of the mechanistic target of the rapamycin-regulated RNA-binding protein La-related protein 1 to mRNA [54], thus participating in microRNA-mediated gene silencing [55], and by regulating immunoglobulin secretion in plasma cells [56]. One study reported that ECHDC2, increases susceptibility to ischemic injury, whereas resistance to ischemic/reperfusion injury is enhanced when ECHDC2 expression is knocked down. Mechanistically, ECHDC2 may increase the cellular levels of branched-chain amino acids without altering mitochondrial oxygen consumption [57]. Finally, in an animal experiment about stroke, fastigial nucleus stimulation was shown to lead to neuroprotection by targeting BIRC2 [58]. What is more is that BIRC2, induced by glial cell line-derived neurotrophic factor A, as well as BIRC3 and Gadd45b can suppress the process of neuron apoptosis, as reported in a previous in vitro experiment [59]. Therefore, four potential gene biomarkers—F13A1, ECHDC2, CTSG, and BIRC2—may have direct or indirect relationships with the pathological processes of stroke, whereas PABPC1 may affect both gene silencing and plasma cell activation.
We also evaluated specific immune cell patterns in IS and found that monocytes, activated NK cells, γδT cells, CD8+ T cells, inactive and activated mast cells and neutrophils were detected at significantly higher levels in stroke patients compared with healthy individuals, whereas memory B cells, Treg cells, and M0 macrophages were detected at a lower level in person with stroke compared with healthy individuals. After acute stroke, immune cells reach the ischemic region in an orderly and continuous process [60]. The quantity of infiltrating Treg cells, which are characterized by forkhead box protein P3 expression and secrete interleukin 10 to inhibit excessive immune responses [61], was lower in stroke group, whereas the quantities of infiltrating γδT cells, neutrophils, and activated NK cells [62], which damage ischemic tissues [63], were higher in stroke patients than in healthy individuals, indicating more severe inflammatory immune responses in stroke patients. Because the samples of the two GEO datasets were both collected from peripheral blood, the quantity of macrophages identified in the study was not reflective of the number of macrophages in the ischemic cerebral region. Furthermore, peripheral blood monocytes are recruited into the ischemic brain tissue via chemokines, whereas both microglia-derived and monocyte-derived macrophages are produced in cerebral ischemic tissues [60]. Consequently, our results revealed a high monocyte level accompanied by a low macrophage level.

The results of a correlative analysis between potential gene biomarkers and immune cells showed that eosinophils were positively correlated with \( PABPC1 \) and \( BIRC2 \) expression, activated dendritic cells were positively correlated with \( PABPC1 \) expression, M0 macrophages were positively correlated with \( CTSG \) expression, monocytes were positively correlated with \( F13A1 \) and negatively correlated with \( ECHDC2 \) expression, and neutrophils were negatively correlated with \( PABPC1 \) expression. \( ECHDC2 \) was expressed at low levels, whereas \( F13A1 \) was expressed at a higher levels compared with healthy group, which suggests that monocytes differentiate into both activated macrophages and dendritic cells. Moreover, \( CTSG \) may increase the number of macrophages, especially M0 macrophages and regional M1 macrophages in ischemic tissues, which would aggravate the inflammatory response. Further research is required to ascertain the complex immune cell infiltration process in cerebral ischemic regions, especially regarding the three subtypes of macrophages.

Several limitations do exist in our research. First of all, the datasets chosen from the GEO database were all obtained from peripheral blood samples; therefore, biological processes that occur in the brain, such as the differentiation of the three subtypes of macrophages, could not be investigated. RNA-seq data based on cerebral tissue samples should be jointly analyzed in future studies. Secondly, the feature fitted for the five genes did not achieve an optimal AUC, specificity, or accuracy, and the detailed mechanisms underlying the effects of these gene products on the pathological processes of IS via an immune response or other process remain unknown. Therefore, further basic clinical studies are essential to clarify the molecular biological mechanisms in the development of IS.

In conclusion, we identified five potential diagnostic genes for IS, namely, \( CTSG \), \( F13A1 \), \( PABPC1 \), \( ECHDC2 \), and \( BIRC2 \). We also identified four types of immune cells as potential diagnostic biomarkers, namely M0 macrophages, neutrophils, monocytes, and activated dendritic cells, and the combination of these cell types with the five gene biomarkers improved the diagnostic efficiency. Further research on specific immune cell infiltration patterns and detailed inflammatory immune responses will facilitate the development of immunotherapies for IS. Moreover, further examination of the five potential gene
biomarkers identified in this study may aid development of a new highly specific and sensitive diagnostic method for IS that is noninvasive and convenient.

**Declaration**

**Ethics approval and consent to participate**

Not applicable (Original data of this study is from GEO which is a public functional genomics data repository).

**Consent for publication**

Not applicable.

**Availability of data and materials**

The datasets generated during and/or analysed during the current study are available in the GEO repository, web links to datasets are as following:

- “https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=gse16561”;
- “https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=gse22255”.

**Competing interests**

The authors declare that they have no conflict of interests.

**Funding**

This study is supported by the National Natural Science Foundation of China(82173648), Innovative Talent Support Plan of the Medical and Health Technology Project in Zhejiang Province(2021422878), Internal Fund of Ningbo Institute of Life and Health Industry, University of Chinese Academy of Sciences(2020YJ0212),Ningbo Clinical Research Center for Digestive System Tumors (Grant No.2019A21003), Sanming Project of Medicine in Shenzhen (SZSM201803080).

**Authors’ Contributions**

Li-Yuan Han and Meng Wang had full access to the data of study, and takes responsibility for the integrity of the data. Meng Wang and Rui-Jie Zhang performed the check of the bioinformatic analysis Yin-Dan Yao and Qiong-Feng Guan perform the check the integrality of data and wrote the manuscript. Meng Wang performed the submission of manuscript.

**Acknowledgements**

Not applicable.

**References**

1. Moskowitz MA, Lo EH, Iadecola C. The science of stroke: mechanisms in search of treatments. Neuron. 2010; 67(2): 181–198.
2. Mendis, S.; Davis, S.; Norrving, B. Organizational update: The world health organization global status report on noncommunicable diseases 2014; one more landmark step in the combat against stroke and vascular disease. Stroke 2015, 46, e121–e122.
3. Tu WJ, Zeng XW, Deng A, et al. Circulating FABP4 (Fatty Acid—Binding Protein 4) is a novel prognostic biomarker in patients with acute ischemic Stroke. Stroke 48 (6): 1531-1538.
4. Catanese L, Tarsia J, Fisher M. Acute ischemic stroke therapy overview. Circ Res, 2017; 120 (3):541-58.
5. Bösel J. Blood pressure control for acute severe ischemic and hemorrhagic stroke. Curr Opin Crit Care,
6. Martynov MY, Gusev EI. Current knowledge on the neuroprotective and neuroregenerative properties of citicoline in acute ischemic stroke. *J Exp Pharmacol.* 2015;7:17–28.

7. Kim JY, Park J, Chang JY, Kim SH, Lee JE. Inflammation after ischemic stroke: the role of leukocytes and glial cells. *Experimental neurobiology.* 2016;25(5):241–251.

8. Xie L, Yang SH. Interaction of astrocytes and T cells in physiological and pathological conditions. *Brain Res.* 2015;1623:63–73.

9. Liesz A, Suri-Payer E, Veltkamp C, Doerr H, Sommer C, Rivest S, et al. Regulatory T cells are key cerebroprotective immunomodulators in acute experimental stroke. *Nat Med.* 2009;15(2):192–199.

10. Zhou M, Wang CM, Yang WL, Wang P. Microglial CD14 activated by iNOS contributes to neuroinflammation in cerebral ischemia. *Brain Res.* 2013;1506:105–114.

11. Facci L, Barbierato M, Marinelli C, Argentini C, Skaper SD, Giusti P. Toll-like receptors 2, -3 and -4 prime microglia but not astrocytes across central nervous system regions for ATP-Dependent interleukin-1β release. *Sci Rep.* 2014;4:6824.

12. Ceulemans AG, Zgavc T, Kooijman R, et al. The dual role of the neuroinflammatory response after ischemic stroke: modulatory effects of hypothermia. *J Neuroinflammation.* 2010;7(1):74.

13. Binbin C, Michael SK, Chih LL et al. Profiling Tumor Infiltrating Immune Cells with CIBERSORT Methods. *Mol Biol.* 2018;1711:243-259.

14. Davis S, Meltzer PS. GEOquery: A bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics* 2007, 23:1846–1847.

15. 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.

16. Krug T, Gabriel JP, Taipa R, et al. TTC7B emerges as a novel risk factor for ischemic stroke through the convergence of several genome-wide approaches. *J Cereb Blood Flow Metab* 2012, 32(6):1061-72.

17. Barrett T, Wilhite SE, Ledoux P, et al. NCBI GEO: Archive for functional genomics data sets-update. *Nucleic Acids Res.* 2013, 41:D991-D995.

18. Gautier L, Cope L, Bolstad BM et al. Afy-analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* 2004, 20, 307-315.

19. Parker HS, Leek JT, Favorov AV et al. Preserving biological heterogeneity with a permuted surrogate variable analysis for genomics batch correction. *Bioinformatics* 2014, 30, 2757-2763.

20. Chen C, Chen H, Zhang Y et al. TBtools: An Integrative Toolkit Developed for Interactive Analyses of Big Biological Data. Mol Plan 2020, 13(8):1194-1202.

21. Ritchie ME, Phipson B, Wu D et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 2015, 43, e47.

22. Ginestet C. ggplot2: Elegant Graphics for Data Analysis. *J. R. Stat. Soc. Ser. A Stat. Soc.* 2011,174, 245.

23. Yu G, Wang LG, Han Y et al. clusterProfiler: An R package for comparing biological themes among gene clusters. *OMICS.* 2012, 16, 284-287.

24. Yingyao Z, Bin Z, Lars P, et al. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat Commun.* 2019; 10(1):1523.

25. Suykens, J.A.K.; Vandewalle, J. Least Squares Support Vector Machine Classifiers. *Neural Process. Lett.* 1999,
9, 293-300.

26. Tibshirani, R. Regression Shrinkage and Selection Via the Lasso. J. R. Stat. Soc. 1996, 58, 267-288.

27. Huang ML, Hung YH, Lee WM et al. SVM-RFE based feature selection and Taguchi parameters optimization for multiclass SVM classifier. Sci. World J. 2014, 795624.

28. Friedman, J.; Hastie, T.; Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J. Stat. Softw. 2010, 33,1-22.

29. Wickham H: ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016; ISBN 978-3-319-24277-4. 10.1007/ 978-3-319-24277-4.

30. Hou YH: ggcor: Extended tools for correlation analysis and visualization. https://github.com/houyunhuang/ggcor.

31. Ma MW, Wang J, Zhang Q, et al. NADPH oxidase in brain injury and neurodegenerative disorders. Mol Neurodegener. 2017;12(1): 7.

32. Dugue R, Nath M, Dugue A, Barone FC. Roles of pro-and anti-inflammatory cytokines in traumatic brain injury and acute ischemic stroke. Mechanisms of Neuroinflammation: IntechOpen. 2017.

33. Iadecola C, Anrather J. The immunology of stroke: from mechanisms to translation. Nature medicine. 2011;17(7):796. doi: 10.1038/nm.2399.

34. Guruswamy Revathy, ElAli Ayman. Complex Roles of Microglial Cells in Ischemic Stroke Pathobiology: New Insights and Future Directions. International Journal of Molecular Sciences. 2017;18(3):496.

35. McDonough A, Lee RV, Noor S, et al. Ischemia/reperfusion induces interferon-stimulated gene expression in microglia. J Neurosci. 2017;37(34):8292–8308.

36. Zhou M, Wang CM, Yang WL, Wang P. Microglial CD14 activated by iNOS contributes to neuroinflammation in cerebral ischemia. Brain Res. 2013;1506:105–114.

37. Facci L, Barbierato M, Marinelli C, Argentini C, Skaper SD, Giusti P. Toll-like receptors 2, -3 and -4 prime microglia but not astrocytes across central nervous system regions for ATP-Dependent interleukin-1β release. Sci Rep. 2014;4:6824.

38. McDonough A, Lee RV, Noor S, Lee C, Le T, Iorga M, et al. Ischemia/reperfusion induces interferon-stimulated gene expression in microglia. J Neurosci. 2017;37(34):8292–8308.

39. Lalancette-Hébert M, Swarup V, Beaulieu JM et al. Galectin-3 is required for resident microglia activation and proliferation in response to ischemic injury. J Neurosci. 2012;32(30):10383–10395.

40. Jin WN, Shi SX, Li Z, Li M, Wood K, Gonzales RJ, et al. Depletion of microglia exacerbates posts ischemic inflammation and brain injury. J Cereb Blood Flow Metab. 2017;37(6): 2224–2236.

41. Li L, Lundkvist A, Andersson D, Wilhelmsson U, et al. Protective role of reactive astrocytes in brain ischemia. J Cereb Blood Flow Metab. 2008(28): 468–481.

42. Cekanaviciute E, Fatihali N, Doyle KP, et al. Astrocytic transforming growth factor-beta signaling reduces subacute neuroinflammation after stroke in mice. Glia. 2014(62): 1227–1240.

43. McTigue DM, Tripathi RB. The life, death, and replacement of oligodendrocytes in the adult CNS. J Neurochem. (2008) 107:1–19.

44. Bandettini WP, Kellman P, Mancini C, et al. MultiContrast Delayed Enhancement (MCODE) improves detection of subendocardial myocardial infarction by late gadolinium enhancement cardiovascular magnetic resonance: a clinical validation study. J Cardiovasc Magn Reson. 2012; 14(1): 83.
45. Yue Peng, Xiaofeng Sun, Ying Liang. Role of DNA methylation on human CTSG in dermatomyositis myoideum. *Cell Biol Int*. 2020. 44 (12): 2409-2415.

46. Herrmann SM, Funke-Kaiser H, Schmidt-Petersen K, et al. Characterization of polymorphic structure of cathepsin G gene: role in cardiovascular and cerebrovascular diseases. *Arterioscler Thromb Vasc Biol*. 2001. 21(9):1538-1543.

47. Iacoviello L, Kolpakov V, Salvatore L, et al. Human endothelial cell damage by neutrophil-derived cathepsin G. Role of cytoskeleton rearrangement and matrix-bound plasminogen activator inhibitor-1. *Arterioscler Thromb Vasc Biol*. 1995. 15: 2037–2046.

48. Pidard D, Renesto P, Berndt MC, et al. Neutrophil proteinase cathepsin G is proteolytically active on the human platelet glycoprotein Ib-IX receptor: characterization of the cleavage sites within the glycoprotein Ib alpha subunit. *Biochim J*. 1994. 303(pt2): 489–498.

49. C R Abraham, K Kanemaru, L Mucke. Expression of cathepsin G-like and alpha 1-antichymotrypsin-like proteins in reactive astrocytes. *Brain Res*. 1993.10;621(2): 222-232.

50. R Meli, G M Raso, C Cicala, et al. Thrombin and PAR-1 activating peptide increase iNOS expression in cytokine-stimulated C6 glioma cells. *J Neurochem*. 2001.79(3): 556-563.

51. K Kanemaru, B Meckelein, D C Marshall, et al. Synthesis and secretion of active alpha 1-antichymotrypsin by murine primary astrocytes. *Neurobiol Aging*. 1996;17(5): 767-771.

52. Bagoly Z, Koncz Z, H Jolán, et al. Factor XIII, clot structure, thrombosis. *Thromb Res*. 2012. 129(3): 382-7. D Martinj OP, Arjen JCS, Frits RR, et al. Coagulation factor XIII gene variation, oral contraceptives, and risk of ischemic stroke. *Blood*. 2008. 111 (3): 1282-1286.

53. Benjamin Br, Kiely NJ, Shimul C, et al. Novel Factor XIII variant identified through whole-genome sequencing in a child with intracranial hemorrhage. *Cold Spring Harb Mol Case Stud*. 2018. 4(6): a003525.

54. Ewan MS, Nour El-HB, Katherine M, et al. The mTOR regulated RNA-binding protein LARP1 requires PABPC1 for guided mRNA interaction. *Nucleic Acids Res*. 2021. 49(1): 458-478.

55. Eric H, Joerg EB, Susanne H, et al. Two PABPC1-binding sites in GW182 proteins promote miRNA-mediated gene silencing. *EMBO J*. 2010. 29(24): 4146-60.

56. Peng YZ, Yuan JJ, Zhang ZC, et al. Cytoplasmic poly(A)-binding protein 1 (PABPC1) interacts with the RNA-binding protein hnRNPLL and thereby regulates immunoglobulin secretion in plasma cells. *J Biol Chem*. 2017. 292(29): 12285-12295.

57. Du JH, Li ZX, Li QZ, et al. Enoyl coenzyme a hydratase domain-containing 2, a potential novel regulator of myocardial ischemia injury. *J Am Heart Assoc*. 2013. 2(5): e00233.

58. Huang LG, Li JP, Pang XM, et al. MicroRNA-29c Correlates with Neuroprotection Induced by FNS by Targeting Both Birc2 and Bak1 in Rat Brain after Stroke. *CNS Neurosci Ther*. 2015. 21(6): 496-503.

59. Wang YQi, Qin YX, Guo TW, et al. High Concentration of Glial Cell Line-Derived Neurotrophic Factor Protects Primary Astrocytes from Apoptosis. *Dev Neurosci*. 2018;40(2): 134-144.

60. Jian ZH, Liu R, Zhu X, et al. The Involvement and Therapy Target of Immune Cells After Ischemic Stroke. *Front Immunology*. 2019.10: 2167.

61. Liesz A, Hu X, Kleinschnitz C, et al. Functional role of regulatory lymphocytes in stroke: facts and controversies. *Stroke*. 2015 (46): 1422–30.

62. Yan G, Qiang L, Wei W, et al. Ischemic neurons recruit natural killer cells that accelerate brain infarction.
63. Lin Y, Zhang JC, Yao CY, et al. Critical role of astrocytic interleukin-17A in post-stroke survival and neuronal differentiation of neural precursor cells in adult mice. *Cell Death Dis.* 2016(7): e2273.
PCA diagram before and after batch effect removal and volcano map of DEGs. (A) PCA cluster diagram of the GSE16561 and GSE22255 datasets before batch effect removal. (B) PCA diagram of the GSE16561 and GSE22255 datasets after batch effect removal. (C) Volcano plot of DEGs; red, grey and blue respectively represents up-regulated DEGs, no significant difference genes and down-regulated DEGs.
Figure 2

Results of GO enrichment analysis. (A) Bubble plots where horizontal axis represents generation of each pathways included in three modules (BP, CC and MF). (B) Bar plots where horizontal axis represents numbers of genes under each pathways that can be classified into three modules (BP, CC and MF). (C) Circle diagram which shows the general situation of enrichment processes.
Figure 3

Results of KEGG enrichment analysis. (A) Bubble plots where horizontal axis represents gene ratio of each pathways. (B) Bar plots where abscissa displays the number of genes under each pathways. (C) Circle diagram which shows the general situation of enrichment processes.
Figure 4

Screening process of potential diagnostic markers. (A) The LASSO coefficient profile of mRNAs. (B) The tuning parameters ($\log \lambda$) of IS related mRNAs were used to cross-validate the error curve. Perpendicular imaginary lines were plotted at the optimal value according to the minimal criterion and 1-se criterion. (C) Process of screening potential diagnostic markers via SVM-RFE algorithm. (D) The common part of diagnostic markers screened by two algorithms is shown by venn diagram.
Figure 5

Analysis of characteristics of immune infiltration. (A) Three-dimensional PCA cluster diagram of immune infiltration between IS group and normal group. (B) Three-dimensional PCA diagram of two groups that each one has a distribution ellipse. (C) Correlation plot of immune cells. The size of the eccentricity of each ellipse represents the strength of the correlation; red represents a negative correlation, blue represents a positive correlation. The lighter the color, the weaker the correlation. (D) Violin plot of the quantities of 22 types of immune cells between two groups.
Figure 6

Calculation of the best multi-variate fitted prognostic factor. (A) Asymmetrical diagram of correlations between 27 potential biomarkers (5 genes and 22 kinds of immune cells). (B) ROC curve of the efficacy of 5 diagnostic genes and fitted factor in train set. (C) Evaluation of efficacy of 4 diagnostic prognostic types of cells and fitted factor in train set. (D) Evaluation of efficacy of 5 diagnostic genes and fitted factor in test set. (E) Evaluation of efficacy of 4 diagnostic types of cells and fitted factor in test set.
Figure 7

Correlation between 5 potential diagnostic genes and 22 types of immune cells. (A) Correlative relationship between BIRC2 and 22 type of immune cells. (B) Correlative relationship between CTSG and 22 type of immune cells. (C) Correlative relationship between ECHDC2 and 22 type of immune cells. (D) Correlative relationship between F13A1 and 22 type of immune cells. (E) Correlative relationship between PABPC1 and 22 type of immune cells. The color of dot as well as the color of dot represents the
correlative relationship. The color is closer to deep green, the correlative relationship is stronger and the color is closer to yellow, the correlation is weaker. The bigger the size of dot, the stronger the correlation, the smaller the size, the weaker the correlation.

**Supplementary Files**

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryFigures.docx