Identification of Key Ischemic Stroke Genes by Computational Systems Biology

Rongting Yue (✉ rongting.yue@uconn.edu)

University of Connecticut

Abhishek Dutta

University of Connecticut

Research Article

Keywords: systems biology , ischemic stroke , hub genes , co-expression , microarray

Posted Date: November 18th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-1064869/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Identification of Key Ischemic Stroke Genes by Computational Systems Biology

Rongting Yue^{1,*} and Abhishek Dutta¹

¹Department of Electrical and Computer Engineering, University of Connecticut, 371 Fairfield Way, Storrs, CT 06269 USA

*Correspondence: rongting.yue@uconn.edu

ABSTRACT

Stroke is one of the leading causes of death in humans. Even if patients survive from stroke, they may suffer sequelae such as disability. Treatment for strokes remains unsatisfactory due to an incomplete understanding of its mechanisms. This study investigates Ischemic Stroke (IS), a primary subtype of stroke, through analyses based on microarray data. Limma (in R) derives differentially expressed genes, and the protein-protein interaction (PPI) network is mapped from the database. Gene co-expression patterns are obtained for clustering gene modules by the Weighted Correlation Network Analysis (WGCNA), and genes with high connectivity in the significantly co-expressed modules are selected as key regulators. Common hubs are identified as *Cdkn1a*, *Nes* and *Anxa2*. Based on our analyses, we hypothesize that these hubs might play a key role in the onset and progression of IS. Result suggests the potential of identifying unexplored key regulators by the systemic method used in this work. Further analyses aim at expanding candidate genes for screening biomarkers for IS, and experimental validation is required on identified potential hubs.

Keywords: systems biology | ischemic stroke | hub genes | co-expression | microarray.

- Abnormal regulations and co-expression patterns of genes are utilized to systemically detect pathogenesis in stroke process.
- Protein-protein interaction network is mapped based on differentially expressed genes derived by the moderated t-statistics (by Limma).
- An approximately scale-free network is constructed to cluster gene modules based on the correlations between gene expression profiles (by WGCNA).
- Common genes identified are hypothesized to account for IS mechanisms.

1 INTRODUCTION

Stroke, as a cerebrovascular disease, takes the third most significant fraction of deaths in the world while also being a leading cause of disability¹. Moreover, patients who have survived a stroke may still lose independence and require long-term healthcare² with great economic burden³. As a heritable disease caused by blood blockage in brain artery, IS is a common subtype that makes up 87% of all strokes⁴. However, diagnosis of IS based on clinical brain imaging is not always accessible, and biomarkers based on gene expression profiles are more efficient in identifying strokes⁵. Therefore, the continuous efforts to find out the risk factors of IS are significant, so are the efforts for active prevention, accurate diagnosis and points for therapeutic interventions⁶.

Among several stroke models of rodent animals that have been investigated to explore treatments for IS, proximal occlusion of the middle cerebral artery (MCA) via the intraluminal suture technique (so-called filament or suture model) is not only the most frequently used but also the most clinically relevant model that mimics the mechanism of human IS^{7,8}. This model has been introduced as mouse injury model for experimental stroke research⁹.

Over the past decade, investigators have utilized gene expression profiles based on data from microarrays to detect key regulators of disease mechanisms, since the profiles have the potential to serve as disease-related biomarkers. Probes in microarrays are capable of measuring thousands of gene expression levels or intensities simultaneously. These high-throughput technologies identify physical and functional relationships between proteins and genes on a large scale. However, mechanisms of IS's causes and progression relevant to abnormal gene expressions remain largely unclear. Therefore, more comprehensive information about abnormal gene expressions on IS needs to be implemented to intervene stroke process more efficiently. One way to

perform gene expression analyses is to find out differentially expressed genes (DEG), which tell abnormal gene up- and down-regulations that are responsible for diseases of interest^{10,11}. The resulting DEGs can be further investigated through enrichment analysis based on different databases. Gene co-expression is another approach that finds out biomarkers for prediagnosis from the view of systems biology, and it has uncovered significantly novel signaling pathways for stroke^{12–14}. In the present study, an analysis based on these two methods was performed to compare gene expressions of mice with Middle Cerebral Artery Occlusion (MCAO)-induced stroke and mice with sham operations. A small group of significantly differentially expressed genes were identified as key regulators. The result has been further filtered by co-expressed gene clusters in the network based on the co-expression information from the correlation between genes and gene expression similarity from network topology level. The core genes in the module were excavated to ascertain the biomarkers highly correlated with the diagnosis of IS. The resulting protein-coding genes *Cdkn1a*, *Nes* and *Anxa2* were predicted to account for the pathogenesis in stroke process. The next step of this study will focus on expanding hub candidates in geneset and validation through the combination of functional enrichment analysis and other bioinformatics methods to potentially provide new therapeutic targets for the diagnosis and intervention of IS.

2 RESULTS

2.1 Differentially Expressed Genes from Microarray

High throughput microarray provides statistics of gene expressions. Gene expression data of GSE35338 and the description file of GPL1261 were downloaded from the NCBI GEO database. After microarray normalization, the differentially expressed genes between the stroke samples and the controls were analyzed. A total of 21746 gene profiles were extracted and converted from GSE35338. Gene expression data was analyzed for three different time points according to description file.

Cut-off criteria of fold-change greater than 1.5 and *P*-values lower than 0.05 was chosen to filter significantly differentially expressed genes in mice with the stroke than the sham surgery. As a result, for up-regulation, 1130 genes, 466 genes and 414 genes were identified for the first day, third day and seventh day, respectively, after treatments. For down-regulated genes, 919 genes, 277 genes and 25 genes were identified, respectively. Venn diagrams indicate that for these three groups after treatment, 139 genes were detected in common for up-regulation, and 3 genes were for down-regulation (Fig. 1a and Fig. 1b). The common genes were mapped into the STRING database to get PPI network. The interaction confidence of PPI was set to be higher than 0.7. The network was imported to Cytoscape (Fig. 2a) and all the disconnected genes were removed for topological analysis. Proteins with degrees more than 10 and stress centrality greater than 1000 were selected as hub proteins (Fig. 2b). As a result, 27 proteins in the network were chosen as key regulators for MCAO-induced stroke.

Figure 1. Venn diagrams of up- and down-regulated genes when detecting DEGs with "Limma" package in R. To identify abnormal gene expressions in all three sampling time points, the genes intersected by all time points are chosen.

2.2 Gene Co-expression

All gene probes (after matching gene expression data with gene symbols from GEO Platform) were re-arranged and filtered according to their mean expression levels. All 21746 genes were ranked in descending order. The first 6,000 genes were used in our work for the subsequent analysis. Since metabolic networks in all organisms have been suggested to be scale-free networks and scale-free network phenomena has been observed in many empirical studies⁶, a co-expression network approximated to the scale-free structure was obtained by balancing connectivity and scale-free topology fit index. Gene correlation matrices were calculated and raised to a soft-thresholding power to establish a co-expression network. The scale-free topology fit index

(a) Visualization of the mapped PPI network by 142 commonly differentially expressed genes through STRING database. All the disconnected nodes were removed. Nodes are proteins mapped from genes, and edges are interactions mined from text or experimental reports.

(b) Resulting network after being screened by node stress centrality and degrees. The 27 nodes that remained were highly connected hub proteins from DEG analysis.

Figure 2. Construction of PPI network by mapping DEGs into STRING database.

fails to go above 0.8 with suitable mean connectivity (Fig.3) for the 21 samples from the dataset. In this case, we go with power 8 that provides mean connectivity of around 165 and median connectivity of around 131, which is reasonable for the trade-off between the scale-free topological index and mean connectivity. This power value reduces the noise when computing correlations of gene expressions.

Figure 3. Choice of a soft threshold for gene co-expression clustering based on WGCNA. Scale independence and mean connectivity for different soft-thresholding powers. Higher soft thresholds correspond to lower mean connectivity. The power that makes the scale independence higher than 0.8 can not provide a suitable mean connectivity as shown in the plots.

A hierarchical clustering tree was constructed to identify the gene co-expression network, following the procedures of WGCNA.

As shown in Fig.4a, the threshold of cut height (refers to the maximum dissimilarity that qualifies modules for merging) was set to 0.2, such that the modules higher than this height were dissimilar. Genes in the grey module were the ones that have not been assigned to any other modules. A total of 21 modules were identified except for the grey module, and 9 modules were obtained after merging (Fig.4b).

(a) Module clustering based on the eigengenes of the modules. In the dendrogram, heights are the dissimilarities between the detected modules. Branches with heights less than 0.2 (which corresponds to modular similarity less than 80%) were merged as shown by the red line.

(b) Module clustering after merging. Tree branches are the clusters. Colors refer to different modules. In the dendrogram, height corresponds to dissimilarity. Clusters are defined by cutting branches off the dendrogram. Large height leads to a big module, while small height leads to a small but tight module. Grey module contains genes that do not belong to any of the co-expressed clusters.

Figure 4. Detection and clustering of the co-expressed modules based on dendrogram.

With mean gene significance across modules (Fig.5a), two modules labeled with "lightyellow" (containing 584 genes) and "greenyellow" (containing 1033 genes) were identified to ascertain modules that were more likely to account for MCAO-induced stroke. Intramodular connectivities were calculated, and the genes with the highest 30% from each module were used (175 genes from the "lightyellow" module and 310 genes from the "greenyellow" module) to select hubs inside these modules. These genes were chosen as the hubs in stroke-related gene modules. As a result, Venn diagram in Fig.5b shows 3 genes *Cdkn1a*, *Nes* and *Anxa2* in common for gene sets from DEG and co-expression. The resulting genes are biologically and statistically meaningful, since their expressions are varied significantly, according to the microarray data, and they are drawn from the hub candidates in two significant gene co-expression modules in organisms with stroke. These 3 common hub genes are hypothesized to play an important role in the onset and progression of IS.

Figure 5. Analyses of the selected modules based on gene co-expression.

3 METHODS

3.1 Data and Pre-processing

Microarray collects the expression levels of different genes in the form of fluorescent intensity. Gene expression profiles of young adult male mice in GSE35338 (Series matrix of *Mus musculus* data) come from platform GPL1261 in Gene Expression Omnibus (GEO) database. Samples with treatment labels "sham surgery" (controls) and "MCAO-induced stroke" (cases) are selected for analyses. Data is sampled from the purified reactive astrocytes after these treatments.

The probe-level data in Affymetrix probe result file were converted into expression measures and normalized to account for technical variation by Robust Multichip Average algorithm¹⁵, since this technique does not utilize mismatched spots in the microarray. This algorithm is made up of three steps: background correction, normalization and summarization¹⁵. More specifically, background correction is performed based on the probe signal S and background noise BG , and the model observed perfect match PM of RNA probes

$$PM_{ij} = BG_{ij} + S_{ij}, \quad BG \sim N(\mu, \sigma), \quad S \sim Exp(r), \quad (BG, S > 0, BG \text{ and } S \text{ are independent}) \quad (1)$$

where PM_{ij} refers to the observed expression profiles from the j th probe on array i . The purpose of this step is to derive a transformation function B , such that

$$B(PM) = E(S|PM) > 0 \quad (2)$$

When substituting $BG = PM - S$, the joint probability density function of S and BG is

$$f_{S,BG} = f_{S,BG}(S, PM - S) \quad (3)$$

With total probability theorem, the conditional probability density function of signal intensity S given perfect match PM is¹⁶

$$f_{S|PM} = \frac{f_{S,PM(s,pm)}}{\int_0^{PM} f_{S,PM(s,pm)} dx} = \frac{\frac{1}{\sigma} \phi\left(\frac{S-f_1}{\sigma}\right)}{\Phi\left(\frac{f_1}{\sigma}\right) + \Phi\left(\frac{PM-f_1}{\sigma}\right) - 1} \quad (4)$$

where $f_1 = pm - \mu - r\sigma^2$, ϕ is the probability density function (PDF) and Φ is the cumulative distribution function of standard normal distribution. Then the background adjusted PM given measurement is¹⁶

$$E(S|PM) = f_1 + \sigma \frac{\phi\left(\frac{f_1}{\sigma}\right) - \phi\left(\frac{pm-f_1}{\sigma}\right)}{\Phi\left(\frac{f_1}{\sigma}\right) + \Phi\left(\frac{pm-f_1}{\sigma}\right) - 1} \quad (5)$$

The unknown parameters μ , σ and C for cumulative density function of BG and S are estimated through non-parametric kernel estimation¹⁵.

Assume n probe values for gene x are $x_1, x_2, \dots, x_n$ that are *i.i.d*, note their PDF as f , then kernel density estimation is

$$\hat{f}_h(x) = \frac{1}{n} \sum_{k=1}^n K_h(x - X_k) \quad (6)$$

where $K_h(\cdot) = \frac{1}{h} K(\frac{\cdot}{h})$ is a scaled kernel, positive h is bandwidth used as a smoothing factor, and $K(\cdot)$ is the kernel function that is non-negative and $E[K(\cdot)] = 0$ and $\int K(\cdot) dx = 1$ (e.g., Gaussian kernel for microarray data¹⁷ is $\frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}u^2)$).

Then the quantile normalization is performed to make each gene expression distribution identical in statistical properties, such that expression profiles in each sample have identical quartiles and the order of gene expressions within each sample is reserved. The last step, summarization, uses Tukey Median Polish method to reduce noise in an array. The smoothing operations are needed to eliminate the column and row effect in the gene expression matrix. Assume the rows and columns in expression matrix are n samples and m genes, respectively, note the expression profile of sample i and gene j as X_{ij} , then procedures of Tukey Median Polish method in each iteration are listed in Algorithm 1. Data after normalization are almost within the same range. The mean expression levels are used for the repeated probes. (When performing dendrogram clustering based on Euclidean distance and Complete-linkage, only a few outliers are present.)

Algorithm 1: Tukey Median Polish

while *Change of row or column medians* < *threshold* **do**

- 1) Compute row median $M_j = \text{median}(X_{ij})$ ($i = 0, 1, \dots, n$) for gene j . Do this for all genes, and find overall median M_{all} within all row medians M_j s;
- 2) Subtract X_{ij} of gene j by its row median M_j and overall median M_{all} to get the residual values in expression matrix;
- 3) Compute column medians $M_i = \text{median}(X_{ij})$ ($j = 0, 1, \dots, m$) for sample i . Do this for all samples;
- 4) Subtract X_{ij} of sample i by its column median M_i to get the residual values in expression matrix, and add the median of the column medians to the common median M_{all} .

end

3.2 Identify the Differentially Expressed Genes

The statistics of gene expression profiles are analyzed by "Limma" package¹¹ (in R) to find out the Differentially Expressed Genes (DEG). Data is grouped according to the first day, third day and seventh day, respectively, after treatment. The empirical Bayes moderation is utilized in the algorithm, since an inverse Chi-square prior is estimated from data¹¹. Then data is fitted into multiple linear models, and the moderated t-statistics are computed¹¹. The following steps are adapted from the DEG analysis based on "Limma".

A linear model from several microarray for particular gene k is assumed to have mean $E[y_k] = XC_k$ and variance $\text{Var}[y_k] = W_k \sigma_k^2$, where y_k is the obtained expression profile of gene k (i.e., y_k is the k th row of expression matrix), X is the design matrix that explores differences between target samples (2 columns are sample traits with MCAO-induced stroke and sham surgery, and rows are different samples), C_k is coefficient vector, W_k is weight matrix, and σ_k^2 is the variance of expression profiles of a specific gene k ¹¹. Nonzero values on the diagonal refer that the measurement data is present. The contrast matrix F makes the pairwise comparisons and specifies the hypothesis needed to be tested, and it measures the change between cases and controls. Contrasts of coefficients is defined as $\epsilon_k = F^T C_k$.

Linear models are fitted to gene expression data to obtain coefficient estimator $\hat{C}_k$, sample variance estimator γ_k^2 and estimated covariance

$$\text{Var}(\hat{C}_k) = \text{Var}_k \gamma_k^2 \quad (7)$$

where Var_k is a positive definite matrix independent of γ_k^2 . The contrast estimator $\hat{\epsilon} = F^T \hat{C}_k$ has estimated covariance

$$\text{Var}(\hat{\epsilon}_k) = F^T \text{Var}_k F \gamma_k^2 \quad (8)$$

The contrast estimator is nearly normal with mean ϵ_k and covariance matrix $F^T \text{Var}_k F \sigma_k^2$. Note d_{kj} as the j th diagonal element of $F^T \text{Var}_k F$, then given ϵ_{kj} and σ_k^2 , the individual contrast estimator $\hat{\epsilon}_{kj}$ of gene g follows normal distribution:

$$\hat{\epsilon}_{kj} | \epsilon_{kj}, \sigma_k^2 \sim N(\epsilon_{kj}, d_{kj} \sigma_k^2) \quad (9)$$

Given σ_k^2 , sample variance estimator γ_k^2 follows Chi-square distribution with a scalar $\frac{\sigma_k^2}{g_k}$, where g_k is residual degrees of freedom for a linear model of the specific gene k .

$$\gamma_k^2 | \sigma_k^2 \sim \frac{\sigma_k^2}{g_k} \chi_{g_k}^2 \quad (10)$$

The inverse Chi-square prior that derives variances σ_k^2 among genes is estimated with prior estimator γ_a^2 and g_a degrees of freedom. (g_a and γ_a^2 are estimated from observed/residual sample variances γ_k^2 that follows nearly Chi-square distribution. These two parameters are obtained through setting the empirical to the expected values for mean and variance of the finite and nearly normal $\log \gamma_k^2$.)

$$\frac{1}{\sigma_k^2} \sim \frac{1}{g_a \gamma_a^2} \chi_{g_a}^2 \quad (11)$$

Given prior observation with variance d_{0j} at j th time point, log-fold-change of DEGs follows distribution $N(0, d_{0j} \sigma_k^2)$, where d_{0j} is estimated from moderated t-statistics $\tilde{t}_{kj}$ with knowledge of g_a and p_j (i.e., probability of non-zero ε_{kj} or valid portion of DEGs).

The ordinary t -statistic is

$$t_{kj} = \frac{\hat{\varepsilon}_{kj}}{\gamma_k \sqrt{d_{kj}}} \quad (12)$$

where d_{kj} is the j th diagonal element of $F^T \text{Var}_k F$ and it is the unscaled standard deviation, and the moderated t-statistics $\tilde{t}_{kj}$ with added degrees of freedom $g_k + g_a$ has the form

$$\tilde{t}_{kj} = \frac{\hat{\varepsilon}_{kj}}{\tilde{\gamma}_k \sqrt{d_{kj}}} \quad (13)$$

that sample variance γ_k^2 is replaced by posterior variance $\tilde{\gamma}_k^2$, which is

$$\tilde{\gamma}_k^2 = \frac{g_a \gamma_a^2 + g_k \gamma_k^2}{g_a + g_k} \quad (14)$$

From the statistical result, 1.5 fold-change of genes expression (corresponds to 0.58 in logarithmic scale) in a Student's t-test across replicas helps select the significantly variant genes¹⁸. Note that positive fold-change refers to up-regulation of genes and negative value refers to down-regulation. P -value that describes the statistical certainty of genome-wide significance often uses 0.05 as the threshold such that the test is believed to be not by chance with 95% certainty¹⁹. Gene expression profiles on microarray are sampled at first day, third day and seventh day, respectively, after treatment, according to the description file. Therefore, the up-regulated and down-regulated genes should be analyzed separately. Enrichment analyses can further map these DEGs into KEGG database to detect abnormal gene expression according to signaling pathways.

The resulting DEGs are mapped into the STRING database. After setting the confidence level of the interactions, a set of Protein to Protein Interactions (PPIs) are downloaded from the database directly. Only a few genes are responsible for regulation. Since gene regulations are many-to-many, network structure offers an excellent visualization to describe their relations. PPI network is visualized in "Cytoscape" and analyzed with toolbox "NetworkAnalyzer". Hub proteins that act as key regulators can account for most of the disease mechanisms and can be identified with topological parameters²⁰. Stress centrality can be a criterion to select the "hub", since it describes the capability of a protein to connect nodes as a communication center in a protein network²¹, which is computed as follows:

$$C(v) = \sum_{s \neq v} \sum_{t \neq v} \sigma_{st}(v) \quad (15)$$

where σ denotes the number of shortest paths containing edge v , s and t are nodes that are different from the given node v . Another criterion used to get hub proteins is the degree. Degree records the number of nodes connected directly to a given node. High degree value indicates that the nodes are likely to be hubs in a certain network²¹. The resulting DEGs are too many, although the burden can be partially eased by adjusting thresholds, e.g., larger fold-change. Also, in DEG analysis, comparisons between samples are performed for all time points, which means the complexity and the computational load increases when data is sampled from several time points. And no information is available on edges in the PPI network.

3.3 Clustering Based on Gene Co-expression

To not analyze gene expressions according to timestamps, WGCNA²² (in R) can be incorporated to screen real hubs²³. It provides an unsupervised analysis method that clusters genes based on their expression profiles. A co-expression network is established that approximates the scale-free structure by properly raising gene expression correlations to a soft-thresholding power, enlarging the differences between correlations. It identifies the highly co-expressed gene sets in which those genes have identical expression patterns. For example, if the samples are all control (or case), all the co-expressed genes should be up-regulated or down-regulated within those samples. Weights in the network annotate correlations. A biological scale-free network has few hub nodes with plenty number of connections, and the remaining nodes have fewer connections²⁴. Two advantages of the scale-free network are that a new node will be more likely to be connected to the hubs, and the network is robust to accidental failures. Besides, the clustering approach based on gene co-expression outperforms principal component analysis when dealing with high-dimensional gene expression data, since it does not focus on the dimensions of data with the highest variation⁶.

With gene expression profiles, Pearson's correlation between a pair of gene expression profiles x_i, x_j can be calculated:

$$cor(x_i, x_j) = \frac{\sum (x_i - \bar{x}_i)(x_j - \bar{x}_j)}{\sqrt{\sum (x_i - \bar{x}_i)^2 \sum (x_j - \bar{x}_j)^2}} \quad (16)$$

According to WGCNA, the co-expression similarity is derived as the absolute value of this correlation:

$$s_{ij} = |cor(x_i, x_j)| \quad (17)$$

Adjacency matrix describes connection strength between gene expression profiles in the proposed weighted network and it is calculated by raising the co-expression similarity to a power:

$$a_{ij} = (s_{ij})^\lambda \quad (18)$$

where soft-thresholding power λ is used to reduce the noise of the correlations in the adjacency matrix and used in co-expression modules. Key concept in WGCNA is connectivity that describes the relative significance of a gene in network⁶. Connectivity measurement for each gene is the sum of the connection strengths between that gene and all the other genes in the network, and the mathematical representation is $\sum_{j \neq i} a_{ij}$ for the i -th gene, a_{ij} is the adjacency of gene i and j . Mean connectivity C_m , also refers to network density²⁵, is then calculated as follows:

$$C_m = \frac{\sum_i \sum_{j \neq i} a_{ij}}{n(n-1)} \quad (19)$$

The higher power makes the network closer to a scale-free structure, but it results in lower mean connectivity. Generally, a kink appears in the scale independence as soft threshold changes to reserve mean connectivity, since the scale-free topology fit will not improve much after the kink²⁶. A larger scale-free topology fit related to connectivity refers to a high signed R^2 , indicates that the model explains most of the variability of the response data around its mean. A proper λ can then be chosen to balance the scale-free topological index and mean connectivity. Since Low gene expressions are usually regarded as noise, the normalized gene expression profiles are filtered by mean expression (or variance) for robustness. A topological overlap matrix (TOM) is then calculated to measure the topological overlap of two genes i and j as follows:

$$TOM_{ij} = \frac{a_{ij} + \sum_{k \neq i, j} a_{ik} a_{kj}}{\min(k_i, k_j) + 1 - a_{ij}} \quad (20)$$

where the similarity measures in TOM take the range [0,1] and the dissimilarity is then derived by subtracting similarity from 1. Then, an agglomerative hierarchical clustering strategy is applied using the TOM-based dissimilarities as the merging distances of the branches, and these branches are merged with a bottom-up strategy through dynamic hybrid cut algorithm²⁷. Member quantity M within a given cluster core is as follows:

$$M = \min\{\text{int}(\frac{N_0}{2} + \sqrt{N - \frac{N_0}{2}}), N\} |_{N_0=100} \quad (21)$$

where N_0 is the minimum cluster size (chosen to be $N_0 = 100$) and N is the total number of objects in the cluster. Small clusters are merged into the nearest larger clusters within the same branches in dendrogram based on average dissimilarities, and tiny clusters that fail to meet the size requirement can be assigned to the closest basic cluster even if they are not on the same branch. The other unlabelled objects remain to be individuals.

A cut height (maximum joining height) of 0.2 is set to fuse modules with 80% similarity and reverse number of clusters as much as possible. Scatters of cluster cores describe the average of all pairwise dissimilarities of the elements between each two of the clusters. A maximum core scatter, based on the 5th percentile of joining heights, is used to exclude objects which are far from a cluster, even if they belong to the same branch of the dendrogram. A minimum cluster gap, based on cut height and the 5th percentile of joining heights, is used to separate clusters from their surroundings.

Eigen-genes E , which are the unique orthonormal superpositions of the genes²⁸, are used as the representatives to tell differences between one gene module and another by taking the first principal component as a summary of that module, after performing singular value decomposition on the expression matrix. The computation is as follows:

$$X = USE^T \quad (22)$$

where X comprises normalized gene expression profiles, U is the matrix with columns as eigenvectors of XX^T , and S is a matrix with singular values. Modular similarity is calculated by the correlation between the module eigengenes. Modules whose eigengenes are highly correlated are merged. Gene significance is measured by the absolute correlation between the trait (stroke) and the expression profile measures. The modular significance is measured by mean gene significance across modules. A set of significant genes, with the highest intramodular connectivities in the module, are selected as hubs in that module. Note hub DEGs as set A , and the co-expressed genes identified above as B , then the intersection geneset C can be derived

$$C = A \cap B, \quad s.t. \quad C \subseteq A \text{ and } C \subseteq B \quad (23)$$

where the intersection of this co-expressed gene set and DEGs are the target genes that are not only biologically meaningful²⁷, but also statistically meaningful to stroke. The modules that are close to stroke are identified till this step. However, a single module contains hundreds of genes, and the most important genes have to be detected. The correlations between genes and module and between phenotype (stroke) and genes are analyzed jointly to solve the aforementioned issue. These genes are not only highly related to their clustering modules but also highly related to IS.

4 DISCUSSION

IS is a major subtype of stroke with an unclear predominant etiology. The differentially expressed gene products may be the biomarker in the processes of brain damage after stroke, including axon guidance and apoptotic response to cerebral ischemia^{10,29}. More risk factors of IS remain to be fully elaborated for therapeutic interventions⁶.

Analyses based on microarray data have gained increasing attention because they acquire high throughput gene expression data simultaneously and provide us comprehensive knowledge of variation in gene expression to increase our understanding of normal and diseased states³⁰. In this study, we use microarray data of mice model with MCAO-induced stroke since it is one of the most frequently used ones for the capability to mimic the mechanism of human IS^{7,8}. We analyzed the genetic profiles of reactive and quiescent astrocytes in young adult mouse brains with MCAO-induced stroke from GEO database. Since the methodologies of identifying genetic differential expressions and co-expression are highly efficient for targeting disease-related genes and proteins³¹, they provide potential clinical diagnoses, prognostic biomarkers and therapeutic targets for acute ischemic stroke via the identification of DEGs and co-expressed genes.

We first detected altered gene expressions from microarray. DEG analysis showed there were 139 genes and 3 genes in common for up- and down-regulation, respectively. These genes were detected significantly differentially expressed in mice with MCAO-induced stroke. PPI network was constructed by mapping DEGs into STRING database, and a total of 27 hub proteins were identified through topological parameters. The hubs were key regulators that might play an important role in the process of IS. The next phase of this study involved detecting gene co-expression patterns by WGCNA based on an unsupervised manner. Under the assumption that all genes are connected, WGCNA quantifies gene connection strength with the expression correlations. 9 modules were detected (besides the grey module), and 2 of them had higher mean gene significance, which indicated that genes within them were more likely to account for IS. Further in this study, the resulting protein-coding genes Cdkn1a (Cyclin Dependent Kinase Inhibitor 1A), Nes (Nestin) and Anxa2 (Annexin A2) intersected by two methods above were identified as the common hubs, which were predicted to be biologically and statistically meaningful. We hypothesized that the common hubs are most likely to play an important role in the onset and progression of IS based on our analyses.

Cdkn1a encodes cyclin-dependent kinase inhibitor that regulates the cell cycle progression at G1. This gene has been widely investigated as a tumor suppressor for tumor progression and prevention. Though there are not many studies that reveal the direct correlation between Cdkn1a and IS or brain injury, indirect evidences can be found that Cdkn1a is the target gene of p53³², and inhibition of p53 pathway intervenes the process of neurogenesis after IS³³. Single-nucleotide polymorphisms (SNPs) in the human Anxa2 gene have been associated with the increased risk of stroke³⁴. Also, Anxa2 has been studied to connect stroke with several SNPs in the Endothelin 1 gene (on chromosome 6)³⁵. In the rat thrombosis models, the potential clinical utility of recombinant Anxa2 protein in IS has caught the attention of researchers³⁶. Nes is a protein-coding gene in Homo sapiens and

it encodes Nestin protein. Nestin is originally introduced in neuroepithelial stem cells and defines intermediate filament (IF) protein in brains³⁷. Recent evidence suggests that inflammation is involved in the pathogenic progression of ischemic stroke^{3,38}. Nestin is predicted to serve as a context-dependent marker of activated microglia/macrophages, and microglial expression of Nestin increases in response to neuroinflammation³⁹. In atherosclerosis-related research, Nestin cells have been proven to direct inflammatory cell migration during chronic inflammation⁴⁰. Based on experiments, recent studies have validated that Nes expression is closely relevant to ischemic brain^{41,42}, confirming our results obtained by more efficient, comprehensive and generalizable computational systems biology methods grounded in rigorous statistical analyses of microarray differential gene expression data. The key to developing effective and clinically applicable treatment methodologies is a better understanding of the pathophysiology of the disease. Structural hubs in molecular interaction networks potentially provide the therapeutic targets that intervene in the causes or progressions of diseases. With the interaction of gene expression products on the molecular level, more potential biomarkers and therapeutic targets can be identified.

Further studies include the expansion of geneset that account for IS by the approaches used in this work. Missing data occurs in microarray during experiments, and it results in an adverse effect on downstream analysis, which requires imputation algorithms to make use of available information to impute missing values⁴³. Our study used the intersection of DEGs from all time points, and this might be inaccurate when the measurement of gene expression data is missing, which leads to fewer gene candidates for the screening process. In a co-expression network, important modules for the disease are selected through module significance. The unselected modules may also contain genes that contribute to IS, but they are ignored. The genes with the highest connectivity in those less significant modules may be drawn to expand the co-expressed gene candidates. Also, the co-expression network does not offer causality of gene regulation since the undirected edges lose some regulatory information (e.g., the direction of suppression/activation). Besides, experimental validation of functions of unexplored genes is required to test if they contribute to IS mechanism. Also, drug molecules that bind with these disease-related omics molecules should be detected for therapeutic purposes, since specific anti-inflammatory interventions are beneficial either in prevention or acute treatment. Through drug-drug interactions, drugs have the potential to be re-purposed or combined to improve current therapeutic performance for IS.

5 CONCLUSION

Gene expression profiles from mice models with MCAO-induced stroke were analyzed. Gene differential expression screened by co-expression patterns has identified the *Cdkn1a* gene, the *Nes* gene and the *Anxa2* gene that contribute to the pathogenesis of IS from a systemic view. Further studies are needed to expand gene candidates and find drug modules that combine target gene expression products to intervene the process of IS.

References

1. Lozano, R. *et al.* Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the global burden of disease study 2010. *The lancet* **380**, 2095–2128 (2012).
2. Chen, R.-L., Balami, J. S., Esiri, M. M., Chen, L.-K. & Buchan, A. M. Ischemic stroke in the elderly: an overview of evidence. *Nat. Rev. Neurol.* **6**, 256–265 (2010).
3. Moskowitz, M. A., Lo, E. H. & Iadecola, C. The science of stroke: mechanisms in search of treatments. *Neuron* **67**, 181–198 (2010).
4. Virani, S. S. *et al.* Heart disease and stroke statistics—2020 update: a report from the american heart association. *Circulation* **141**, e139–e596 (2020).
5. Stamova, B. *et al.* Gene expression profiling of blood for the prediction of ischemic stroke. *Stroke* **41**, 2171–2177 (2010).
6. Zhao, W. *et al.* Weighted gene coexpression network analysis: state of the art. *J. biopharmaceutical statistics* **20**, 281–300 (2010).
7. Engel, O., Kolodziej, S., Dirnagl, U. & Prinz, V. Modeling stroke in mice-middle cerebral artery occlusion with the filament model. *J. visualized experiments: JoVE* (2011).
8. Smith, H. K., Russell, J. M., Granger, D. N. & Gavins, F. N. Critical differences between two classical surgical approaches for middle cerebral artery occlusion-induced stroke in mice. *J. Neurosci. Methods* **249**, 99–105 (2015).
9. Chan, P. H. *et al.* Brain infarction is not reduced in sod-1 transgenic mice after a permanent focal cerebral ischemia. *Neuroreport* **5**, 293–296 (1993).
10. Li, P. *et al.* Identification of circulating micrornas as potential biomarkers for detecting acute ischemic stroke. *Cell. molecular neurobiology* **35**, 433–447 (2015).

11. Ritchie, M. E. *et al.* limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic acids research* **43**, e47–e47 (2015).
12. Wang, M. *et al.* Lncrnas related key pathways and genes in ischemic stroke by weighted gene co-expression network analysis (wgcn). *Genomics* **112**, 2302–2308 (2020).
13. Wu, Z. *et al.* Lncrna-n1lr enhances neuroprotection against ischemic stroke probably by inhibiting p53 phosphorylation. *Mol. neurobiology* **54**, 7670–7685 (2017).
14. Zhao, F. *et al.* Microarray profiling and co-expression network analysis of lncrnas and mrnas in neonatal rats following hypoxic-ischemic brain damage. *Sci. reports* **5**, 1–11 (2015).
15. Irizarry, R. A. *et al.* Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249–264 (2003).
16. Bolstad, B. M. *Low-level analysis of high-density oligonucleotide array data: background, normalization and summarization* (University of California, Berkeley Berkeley, 2004).
17. Hänzelmann, S., Castelo, R. & Guinney, J. Gsva: gene set variation analysis for microarray and rna-seq data. *BMC bioinformatics* **14**, 1–15 (2013).
18. Raouf, A. *et al.* Transcriptome analysis of the normal human mammary cell commitment and differentiation process. *Cell stem cell* **3**, 109–118 (2008).
19. Qu, H.-Q., Tien, M. & Polychronakos, C. Statistical significance in genetic association studies. *Clin. investigative medicine. Med. clinique et experimentale* **33**, E266 (2010).
20. Sriroopreddy, R., Sajeed, R., Raghuraman, P. & Sudandiradoss, C. Differentially expressed gene (deg) based protein-protein interaction (ppi) network identifies a spectrum of gene interactome, transcriptome and correlated mirna in nondisjunction down syndrome. *Int. journal biological macromolecules* **122**, 1080–1089 (2019).
21. Scardoni, G., TOSADORI, G., LAUDANNA, C., Fabbri, F. & FAIZAAAN, M. Centiscape: Network centralities for cytoscape.
22. Langfelder, P. & Horvath, S. Wgcna: an r package for weighted correlation network analysis. *BMC bioinformatics* **9**, 1–13 (2008).
23. Yuan, L. *et al.* Co-expression network analysis identified six hub genes in association with progression and prognosis in human clear cell renal cell carcinoma (ccrcc). *Genomics data* **14**, 132–140 (2017).
24. Barabási, A.-L. & Bonabeau, E. Scale-free networks. *Sci. american* **288**, 60–69 (2003).
25. Dong, J. & Horvath, S. Understanding network concepts in modules. *BMC systems biology* **1**, 1–20 (2007).
26. Langfelder, P. & Horvath, S. Tutorials for the wgcna package (2014).
27. Langfelder, P., Zhang, B. & Horvath, S. Dynamic tree cut: in-depth description, tests and applications. *November* **22**, 2007 (2007).
28. Alter, O., Brown, P. O. & Botstein, D. Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci.* **97**, 10101–10106 (2000).
29. Sekerdag, E., Solaroglu, I. & Gursoy-Ozdemir, Y. Cell death mechanisms in stroke and novel molecular and cellular treatment options. *Curr. neuropharmacology* **16**, 1396–1415 (2018).
30. Bolón-Canedo, V., Sánchez-Marono, N., Alonso-Betanzos, A., Benítez, J. M. & Herrera, F. A review of microarray datasets and applied feature selection methods. *Inf. Sci.* **282**, 111–135 (2014).
31. Van Dam, S., Vosa, U., van der Graaf, A., Franke, L. & de Magalhaes, J. P. Gene co-expression analysis for functional classification and gene–disease predictions. *Briefings bioinformatics* **19**, 575–592 (2018).
32. Fischer, M. Census and evaluation of p53 target genes. *Oncogene* **36**, 3943–3956 (2017).
33. Wang, J. *et al.* Long noncoding rna h19 prevents neurogenesis in ischemic stroke through p53/notch1 pathway. *Brain research bulletin* **150**, 111–117 (2019).
34. Flanagan, J. M. *et al.* Genetic predictors for stroke in children with sickle cell anemia. *Blood, The J. Am. Soc. Hematol.* **117**, 6681–6684 (2011).
35. Sebastiani, P., Ramoni, M. F., Nolan, V., Baldwin, C. T. & Steinberg, M. H. Genetic dissection and prognostic modeling of overt stroke in sickle cell anemia. *Nat. genetics* **37**, 435–440 (2005).

36. Hedhli, N. *et al.* The annexin a2/s100a10 system in health and disease: emerging paradigms. *J. Biomed. Biotechnol.* **2012** (2012).
37. Lendahl, U., Zimmerman, L. B. & McKay, R. D. Cns stem cells express a new class of intermediate filament protein. *Cell* **60**, 585–595 (1990).
38. Muir, K. W., Tyrrell, P., Sattar, N. & Warburton, E. Inflammation and ischaemic stroke. *Curr. opinion neurology* **20**, 334–342 (2007).
39. Krishnasamy, S. *et al.* Molecular imaging of nestin in neuroinflammatory conditions reveals marked signal induction in activated microglia. *J. neuroinflammation* **14**, 1–14 (2017).
40. Del Toro, R. *et al.* Nestin+ cells direct inflammatory cell migration in atherosclerosis. *Nat. communications* **7**, 1–12 (2016).
41. Shin, Y.-J. *et al.* Characterization of nestin expression and vessel association in the ischemic core following focal cerebral ischemia in rats. *Cell tissue research* **351**, 383–395 (2013).
42. Sakuma, R. *et al.* Brain pericytes serve as microglia-generating multipotent vascular stem cells following ischemic stroke. *J. neuroinflammation* **13**, 1–13 (2016).
43. Liew, A. W.-C., Law, N.-F. & Yan, H. Missing value imputation for gene expression data: computational techniques to recover missing data from available information. *Briefings bioinformatics* **12**, 498–513 (2011).

Acknowledgements

We like to thank Aakanksha Singh, UTC Institute for Advanced Systems Engineering at UConn for proofreading our manuscript.

Author contributions statement

All authors contributed to the discussion of the contents, and reviewed and edited the manuscript.

Competing interests

The authors declare no competing interests.