Identification and analysis of spinal cord injury subtypes using weighted gene co-expression network analysis

Qi Chen[1], Ziru Zhao[2], Guoyong Yin[3], Chuanjun Yang[4], Danfeng Wang[4], Zhi Feng[4], Na Ta[5]

[1]Department of Orthopedics, The Second Affiliated Hospital of Nanjing Medical University, Nanjing, China; [2]Department of Orthopedics, The Fourth Affiliated Hospital of Nanjing Medical University, Nanjing, China; [3]Department of Orthopedics, The First Affiliated Hospital of Nanjing Medical University, Nanjing, China; [4]Department of Orthopedics, Anting Hospital, Shanghai, China; [5]Department of Nursing Management, Anting Hospital, Shanghai, China

Contributions: (I) Conception and design: Z Zhao, N Ta, Q Chen; (II) Administrative support: None; (III) Provision of study materials or patients: G Yin, D Wang; (IV) Collection and assembly of data: C Yang; (V) Data analysis and interpretation: Z Feng; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Correspondence to: Guoyong Yin. Department of Orthopedics, The First Affiliated Hospital of Nanjing Medical University, 300 Guangzhou Road, Gulou District, Nanjing 210003, China. Email: guoyongyin888@sina.com.

Background: Spinal cord injury (SCI) has an immediate and devastating impact on the control over various movements and sensations. However, no effective therapies for SCI currently exist.

Methods: To identify and analyze SCI subtypes, we obtained the expression profile data of the 1,057 genes (889 intersection genes) in GSE45550 using weighted gene co-expression network analysis (WGCNA), and 14 co-expression gene modules were identified. Next, we filtered out the network degree top 10 (degree >80) genes, considered the final key SCI genes. A multifactor regulatory network (105 interaction pairs), consisting of messenger RNAs (mRNAs), long non-coding RNAs (lncRNAs), and transcription factors (TFs) was constructed. This network was involved in the co-expression of key genes. We selected the top 10 regulatory factors (degree >4) as core regulators in the multifactor regulatory network.

Results: The results of functional enrichment analysis of the target gene expressing the core regulatory factor [1,059] showed that these target genes were enriched in pathways for human cytomegalovirus infection, chronic myeloid leukemia, and pancreatic cancer. Further, we used the key genes in the co-expression network to categorize the SCI samples in GSE45550. The expression levels of the top 6 genes (*CCNB2*, *CCNB1*, *CKS2*, *COL5A1*, *KIF20A*, and *RACGAP1*) may act as potential marker genes for different SCI subtypes. On the basis of these different subtypes, 8 SCI core gene CDK1-associated drugs were also found to provide potential therapeutic options for SCI.

Conclusions: These results may provide a novel therapeutic strategy for the treatment of SCI.

Keywords: Weighted gene co-expression network analysis (WGCNA); differential gene expression analysis; multi-factor regulatory network; disease subtypes

Submitted Dec 15, 2020. Accepted for publication Mar 03, 2021.
doi: 10.21037/atm-21-340
View this article at: http://dx.doi.org/10.21037/atm-21-340

Introduction

Spinal cord injury (SCI) is a serious disease that can cause loss of motor, sensory, and autonomic nervous system function. However, no effective therapies for SCI currently exist. At the same time, the related biomarker and molecular mechanism still requires elucidation.

With the development of sequencing technology, studies have increasingly focused on gene expression and transcriptome profiles to understand the pathological process of SCI. Some of these studies are related to the repair of spinal cord injuries, such as the change in the transcriptome profile of biomaterials, to cure SCI (1). Several pathophysiological changes occur during SCI development, and different treatments are used depending on its stage. Previous studies have reported that individuals with SCI have a primary injury, and a secondary injury causes motor and sensory disorders (2,3). A primary injury mainly induces blood spinal cord barrier (BSCB) disruption, hemorrhage, surrounding edema, neuronal and glial apoptosis, and necrosis. A secondary injury is marked by inflammation, leukocyte infiltration, and axon growth cone dieback. This indicates that SCI is closely associated with gene functions and signaling pathway changes. As observed previously, using bioinformatic and computational approaches to determine the site of manipulation of the transcription factor (TF) may facilitate the discovery of an effective treatment for SCI (4). However, the key mechanisms regulating the response of cells and the body to SCI are largely unknown (5). Employing bioinformatic and computational approaches more effectively to understand the molecular mechanisms of SCI could enable us to develop new strategies promoting the recovery of neural functions.

Current research has shown that long non-coding RNAs (lncRNAs) act as important regulatory factors during SCI development (6). In addition, the relationship between lncRNAs, TFs, and mRNAs in multicellular organisms shows high tissue specificity (7-9). Subsequently, the relationship between the lncRNAs, TFs, and mRNAs regulating SCI progression has been extensively predicted and analyzed, and is fundamental for developing effective therapeutic strategies. However, the regulating role of lncRNAs and TFs following SCI, in addition to their underlying functional mechanisms during development, have not yet been sufficiently and systematically investigated.

Weighted gene co-expression network analysis (WGCNA) has previously been proven as a powerful method for identifying co-expressed modules and hub genes, TFs, and lncRNAs (1,10-13). The use of WGCNA can classify genes into a model based on pairwise correlations between genes due to their similar expression pattern, and these models can be correlated to different SCI subtypes. Key gene modules as candidate biomarkers or potential therapeutic targets for SCI can be identified through the use of WGCNA (14). It has been used to study diseases which affect the neurological system, such as neuropathic pain (15), Parkinson's disease (16), and Alzheimer's disease (17). This method has also been successfully utilized in previous studies to identify pathway-related modules in SCI patients (18). Hence, the co-expression network for models and hubs of SCI genes is accessible; however, the number of studies addressing this issue remains insufficient.

The development of microarray and RNA sequencing (RNA-seq) technologies can provide an excellent opportunity to further understand the genetic and molecular variations in individuals with SCI. The lncRNA expression profile can also be determined through gene microarray analysis (19). Furthermore, microarray analysis can be used to determine the lncRNA expression level with a higher sensitivity than that observed with RNA-seq (20). Previous studies have shown the accuracy and consistency of using reannotated probes of gene microarray data (21). The above studies have indicated that some of the microarray probes could be utilized to identify lncRNA expression via probe reannotation, although lncRNA expression profiles were not directly examined in these studies.

Here, we obtained the expression profile data of the 1,057 genes (889 intersection genes) in GSE45550 using WGCNA. We concluded that turquoise and brown modules were most relevant to SCI in 14 modules of WGCNA. Then, we constructed a co-expression network of modules related to SCI, which contained genes of 2 modules. We found that these target genes were involved in pancreatic cancer and angiogenesis human cytomegalovirus infection pathways. Furthermore, we concluded that *CCNB2*, *CCNB1*, *CKS2*, *COL5A1*, *KIF20A*, and *RACGAP1* were extremely important for the classification of SCI subtypes. Additionally, we found 8 drugs (AT7519, Alsterpaullone, seliciclib, Indirubin-3'-Monoxime, hymenialdisine, SU9516, olomoucine, and alvocidib) to be associated with the core gene *CDK1* of SCI, as a consequence of drug analysis. These drugs represent potential therapeutic options for SCI.

We present the following article in accordance with the MDAR reporting checklist (available at http://dx.doi.org/10.21037/atm-21-340).

Methods

Gene Expression Omnibus (GEO) data download and preprocessing

Two sets of large sample expression profiles (GSE45550 and GSE45006) were downloaded from the GEO database. The sample details are shown in *Table 1*. For data preprocessing, the probe was mapped to the gene, and the no-load probe

Table 1 The mRNA expression profile dataset for spinal cord injury (SCI) in GEO

Dataset ID	Platform	SCI	Control
GSE45550	GPL1355	18	6
GSE45006	GPL1355	19	5

The samples in GSE45006 are all SCI data. According to principal component and cluster analysis, data obtained post-day 1 and other time points (post-day 3, post-week 1, 2, 8, and 5 duplicates of each group) are significantly different. The sample obtained post-week 1 was mixed with that obtained post-day 1, and was therefore removed. Since the post-day 1 mouse SCI sample was obtained immediately after the occurrence of SCI, it is considered to be similar to the control, and thus the group is shown.

was deleted. If multiple probes corresponded to the same gene, we selected the median value as the expression value of the gene. Then, we used the biomaRt package for R (https://cran.r-project.org/) to re-annotate the mouse gene symbol to a human gene symbol. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013).

Screening differentially expressed genes (DEGs)

The DEGs between the GSE45550 and GSE45006 data sets were analyzed using the limma package for R. The DEGs of the 2 types of samples were screened according to the threshold (P<0.01, |log2FC| >1), and 375 DEGs were obtained in GSE45550. Among these, 188 genes were highly expressed, while the reduced expression of 187 genes was observed in individuals with SCI. Using GSE45006, 930 DEGs were obtained, of which 591 genes were highly expressed. Reduced expression was observed for 339 genes in individuals with SCI.

Candidate gene collection and expression data

We combined the above DEGs with the SCI-related genes obtained from the Online Mendelian Inheritance in Man (OMIM) and Gene databases to obtain a collection of genes. Genes that appeared at least twice during gene collection were selected as research genes. Using data regarding human protein interactions in the Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) database (https://string-db.org), the genes that interacted with these research genes were extracted to further expand the scope of the research genes. Finally, the candidate gene set (n=1,057) was constructed. The expression profile data of genes (n=889) in the GSE108756 candidate gene set was selected for performing WCNA again.

WGCNA

The systematic biological method that is WGCNA uses gene expression data to construct a scale-free network. First, a similarity matrix was constructed using gene expression data by calculating the absolute value of the Pearson correlation coefficient between the 2 genes. Eq. [1] was used to calculate the Pearson correlation coefficient between genes i and j, where x_i and y_j represented the expression values of the genes i and j, respectively.

$$S_{ij} = \left| \frac{1+cor(x_i+y_j)}{2} \right| \qquad [1]$$

Using Eq. [2], the gene expression similarity matrix was converted into an adjacency matrix, and the network type was signed. Here, β was a soft threshold, which was actually the β power of the Pearson correlation coefficient of each pair of genes. This step could strengthen a strong correlation and weaken a weak correlation in an exponential manner.

$$a_{ij} = \left| \frac{1+cor(x_i+y_j)}{2} \right|^{\beta} \qquad [2]$$

Next, the adjacency matrix was converted into a topological matrix using Eq. [3]. The topological overlap measure (TOM) was used to describe the degree of correlation between genes.

$$TOM = \frac{\sum_{u \neq ij} a_{iu} a_{uj} + a_{ij}}{\min\left(\sum_u a_{iu} + \sum_u a_{ju}\right) + 1 - a_{ij}} \qquad [3]$$

The 1-TOM value indicated the degree of dissimilarity between genes i and j. We constructed a hierarchical cluster of genes using 1-TOM as the distance, and used the dynamic cut tree method for module identification. The most representative gene in each module, called the eigenvector gene (ME), represented the overall gene expression level within the module. It was the first principal component in each module. Eq. [4] was used to calculate the ME; i represents the gene in module q and l represented the microarray sample in module q.

$$ME = princomp\,(x_{il}^{(q)}) \qquad [4]$$

We used the Pearson correlation between the expression profile of a gene in all samples and the expression profile of the eigenvector gene to measure the identity of a gene in the module. This was designated the module membership (MM). We determined the MM using Eq. [5], in which MM represented the expression profile of the i gene and the eigenvector gene (ME) of module q. This indicated the identity of gene i in module q when MM =0, and showed that gene i was not in module q. The closer MM was to +1 or −1, the higher was the correlation between gene i and module q. The sign (positive or negative) indicated whether the gene i was positively or negatively related to the module q.

$$MM_i^q = cor\,(x_i, ME^q) \qquad [5]$$

Gene significance (GS) was used to measure the degree of correlation between genes and external information. The higher the GS, the more biologically significant the gene was. A GS =0 indicated that the gene was not involved in the biological problem being studied.

Using the WGCNA package for R, a weighted co-expression network was constructed to obtain the expression profile data for the candidate gene set. The modules that were closely related to SCI were screened. Inter-subnets were constructed for the co-expressed gene sets in the module, and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis was performed. Core differential genes of SCI were selected according to the network node degree of SCI-related pathways.

Pivot analysis

The pivot node referred to: (I) at least 2 interaction pairs within the module gene; and (II) the P value, which indicated the significance of the interaction between the node and each module, and was expected to be less than or equal to 0.05. This statistical method was hypergeometric.

Non-coding RNA (ncRNA) pivot analysis: Through the interactions between ncRNA and mRNA in the RNA Association Interaction Database (RAID) 2.0 database, seen as the interaction background, we counted the interaction pairs between the ncRNA and the module gene, each ncRNA and the gene in the module, and each ncRNA and the gene outside the module. The pivots were screened for significant P values for the hypergeometric test.

TF pivot analysis

The human TF-mRNA regulatory relationship observed in the Transcription Regulatory Relationships Unraveled Sentence-based Text mining (TRRUST) v2 database acted as an interaction background. All interaction pairs between the TF and module gene were counted. The interaction pairs between each TF and the gene in the module, and those between each TF and the extra-module gene, were also counted. The pivots were screened for significant P values for the hypergeometric test.

Extraction of miRNA, lncRNA, TF, and mRNA to construct a multifactor regulatory network

The sRNA target Base (starBase) v2.0 database (http://starbase.sysu.edu.cn/targetSite.php) shows the interactive relationship between miRNA-lncRNA, miRNA-circRNA, and miRNA-mRNA. We screened out the lncRNA and miRNA by analyzing the interaction pairs between ncRNA, miRNA, and core differential genes in the starBase v2.0 database. The human TF-mRNA regulatory relationship is included in the TRRUST v2 database, and TFs that regulate core differential genes are screened by the TRRUST v2 database. By combining the interaction pairs screened in the previous step, a multifactor regulatory network was constructed, which included mRNA, TF, lncRNA, and miRNA. To analyze their degree of networking, we screened the regulatory factors affecting the network based on the degree of the node, obtained the target gene of the regulatory factor, and performed functional enrichment analysis on the target gene to further verify the influence of the regulatory factor on the disease.

Key co-expression genes mediate SCI typing

We further used the key co-expression genes to classify SCI samples for disease typing. We determined the best K-value (number of categories) according to the method for determining the best sum of the squared error (SSE) inflection point, and used the unsupervised clustering method involving K-means and t-distributed stochastic neighbor embedding (T-SNE) dimensionality reduction to identify the SCI subtype. The random forest model was used to examine the importance of these key co-expression genes in different subtypes. Significant differential genes were analyzed in different subtypes. These differential genes may act as potential marker genes for the SCI subtype.

Drug discovery of SCI

Using the core genes analyzed via SCI typing, and all drugs in DrugBank, we identified a drug that could have a therapeutic effect on SCI.

Statistical Analysis

Two sets of large sample expression profiles (GSE45550 and GSE45006) were downloaded from the GEO database. Weighted gene co-expression network analysis (WGCNA) was used to obtain expression profile data.

Results

Screening differential genes and gene sets

After downloading the SCI expression profile data set GSE45550 from GEO, in which SCI *vs.* Control =18:6, we used the limma package in R to analyze differential genes. According to the standard for selecting different expressed genes (P<0.01, |log2FC| >1), the DEGs of the 2 types of samples were selected, and 375 DEGs were obtained. Totals of 188 and 187 genes were upregulated and downregulated in SCI patients, respectively (*Figure 1*).

We downloaded the SCI expression profile data set GSE45006 from GEO, in which SCI *vs.* Control =19:5. Principal component analysis (PCA) was used on all samples. The samples could be clearly distinguished, and the second sample was removed post-week 1 based on the PCA results. The sample group was marked as early (control) and late (SCI), and differential genes were analyzed using the limma package in R. According to the standard for selecting different expressed genes (P<0.01, |log2FC| >1), 930 DEGs were selected and obtained from the 2 types of samples. In SCI, 591 and 339 genes were upregulated and downregulated, respectively (*Figure 2A,B,C*).

We combined 2 sets of DEGs with SCI-related genes in the GENE and OMIM databases. The genes from at least 2 of the 4 gene sets were selected as research objects to obtain 56 genes. As shown in *Figure 2D*, 111,276 protein interaction pairs were obtained from the STRING database, and the genes that interacted with these 56 genes were further expanded to obtain a candidate gene set (1,057 genes). The expression profile data of the 1,057 genes (889 intersection genes) in GSE45550 was selected for the next WGCNA.

WGCN construction of candidate gene sets

We constructed WGCNs for candidate gene sets using the WGCNA software package in R. The study showed that the co-expression network conformed to the scale-free network: the log(k) value of the node with the degree of connection k was negatively correlated with the log(P(k)) value of the probability of the node, and the correlation coefficient was greater than 0.85. To ensure that the network was scale-free, we chose the optimal value of β=7 (sft$powerEstimate=7). The next step was to convert the expression matrix into an adjacency matrix, and then convert the adjacency matrix into a topological matrix. Based on TOM, we used the average-linkage hierarchical clustering method to cluster genes according to the criteria of the hybrid dynamic cut tree, and set the minimum number of genes in a single gene network module to 10. After using the dynamic shear method to determine the gene module, we determined the eigengenes of each module. We then clustered the module, merged the closer modules into a new module, and set the height to 0.25. A total of 14 modules were obtained (*Figure 3*). The number of genes in each module is shown in *Table 2*, wherein 889 genes were assigned to 14 modules.

We calculated the Pearson correlation coefficient of the ME for each module and sample features (if the Pearson correlation coefficient was the higher, the module was more important). We could conclude from *Figure 3* that the turquoise and brown modules were most relevant to SCI.

SCI co-expression network module, gene function, and enrichment analysis

Based on the relationship of the genes expressed in the 2 turquoise and brown co-expression modules, we chose co-expression pairs with a co-expression weight greater than 0.1 as the side of the final co-expression network. We constructed a co-expression network of modules related to SCI (*Figure 4A,B*), containing genes from 2 modules.

We performed KEGG pathway enrichment analysis on the genes in the turquoise and brown modules using the clusterProfiler R package, as shown in *Figure 4*. The turquoise module is mainly involved in the cell cycle, cellular senescence and PI3K-Akt signaling pathways; the

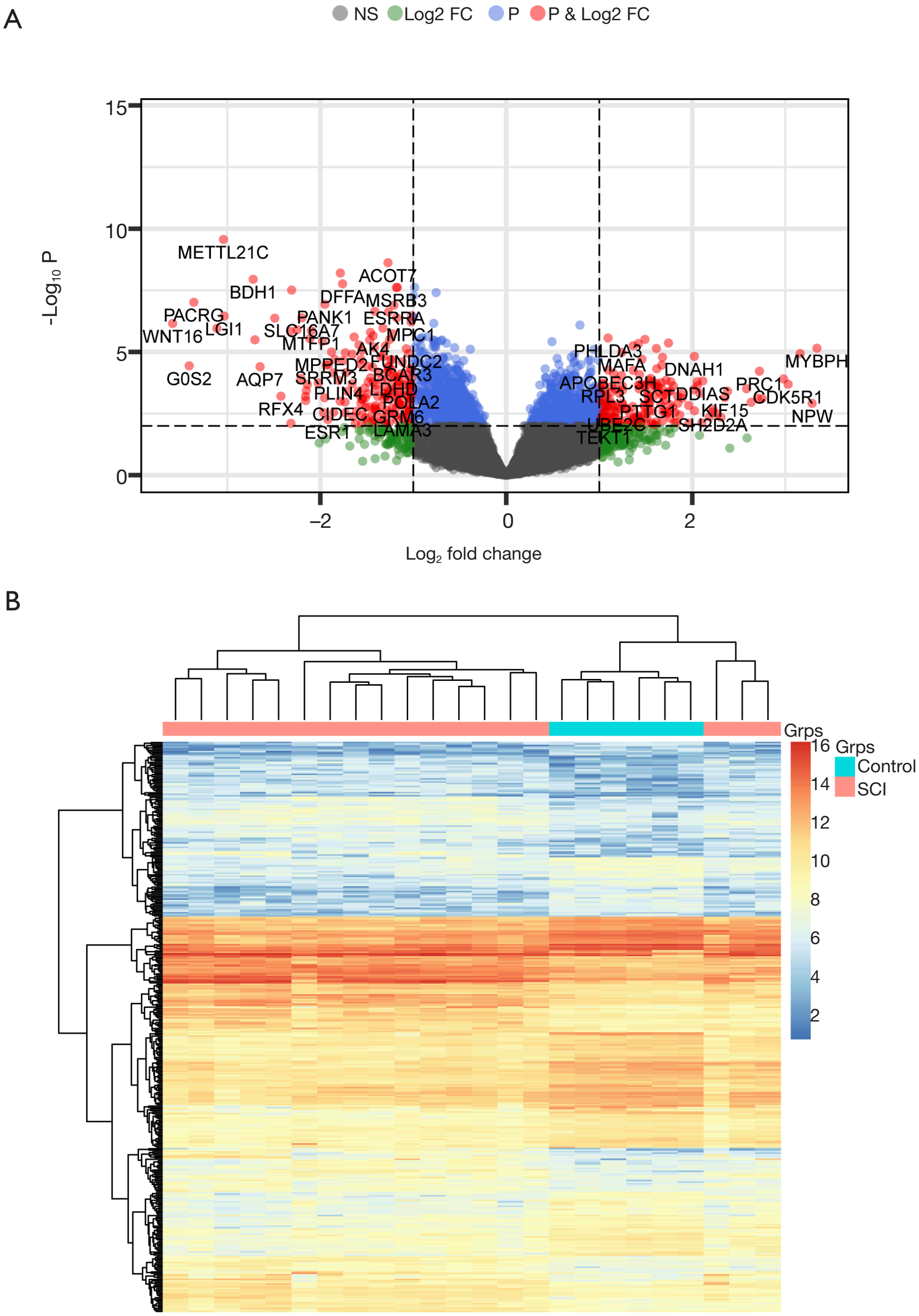


Figure 1 The differential genes in GSE45550 were obtained from GEO (SCI *vs.* Control =18:6). According to the standard for the selection of different expressed genes (P<0.01, |log2FC| >1), the DEGs of 2 types of samples were selected, and 375 DEGs were obtained. Of these, 188 genes were upregulated and 187 genes were downregulated in SCI. (A) Volcano map of the differential genes in GSE45550. (B) Unsupervised hierarchical clustering of contained genes (rows) and samples (columns) was performed, and a heat map was generated. On the right side, the green and orange colored regions represent the Control and SCI, respectively. GEO, Gene Expression Omnibus; SCI, spinal cord injury; DEGs, differentially expressed genes.

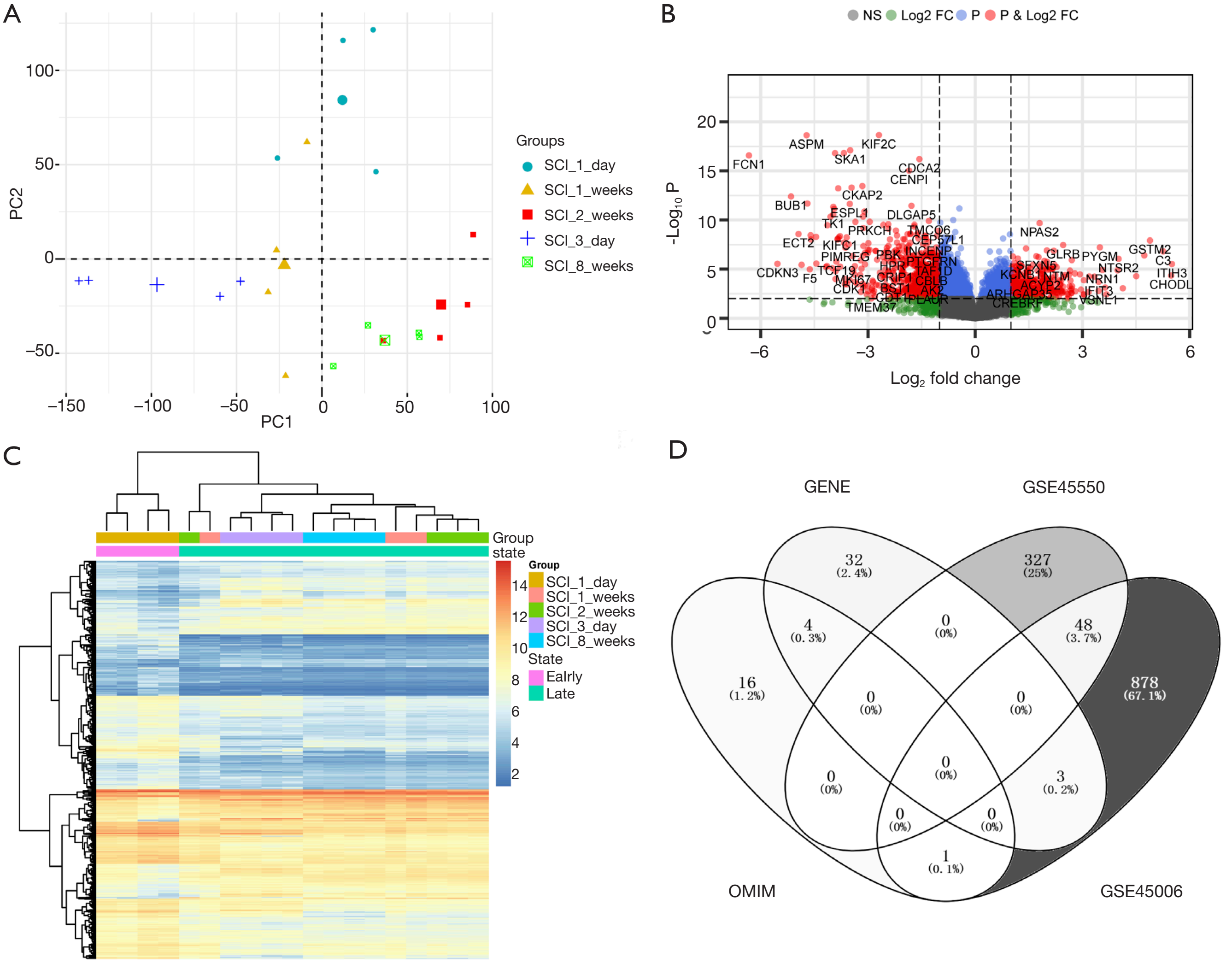


Figure 2 The DEGs of GSE45006 (SCI *vs.* Control =19:5) and the research genes from at least 2 of the 4 gene sets were selected. (A) PCA has been used for all samples from GSE45006. (B) Volcano map of the DEGs (P<0.01, ǀlog2FCǀ >1). (C) Heatmap of the DEGs. The sample group is marked as early (control) and late (SCI). (D) The Venn diagram of the DEGs from GSE45006 and GSE45006 were combined with SCI-related genes in the Gene and OMIM databases. The genes from at least 2 of the 4 gene sets were selected as research objects to obtain 56 genes. DEGs, differentially expressed genes; SCI, spinal cord injury; PCA, principal component analysis; OMIM, Online Mendelian Inheritance in Man.

brown module is mainly involved in Epstein-Barr virus infection, the insulin signaling pathway, and the proteasome pathway.

The KEGG pathway annotation of the 2-module co-expression networks using Cytoscape's ClueGO plug-in to screen for significant enrichment (P<0.05) pathways further confirmed the strong correlation between the turquoise and brown modules and SCI (Figure S1).

The degree distribution of the subnetwork is shown in Figure S2. As the node degree became higher, the number of nodes reduced. It was suggested that most of the genes in the network tended to be independent, and only few genes formed obvious clusters. The expression levels of these genes were changed, and this affected their interactions with neighboring genes. Then, these genes and their neighboring genes were co-expressed, which affected downstream biological functions. Hence, genes with higher node degrees are likely to be key genes in subsequent SCI

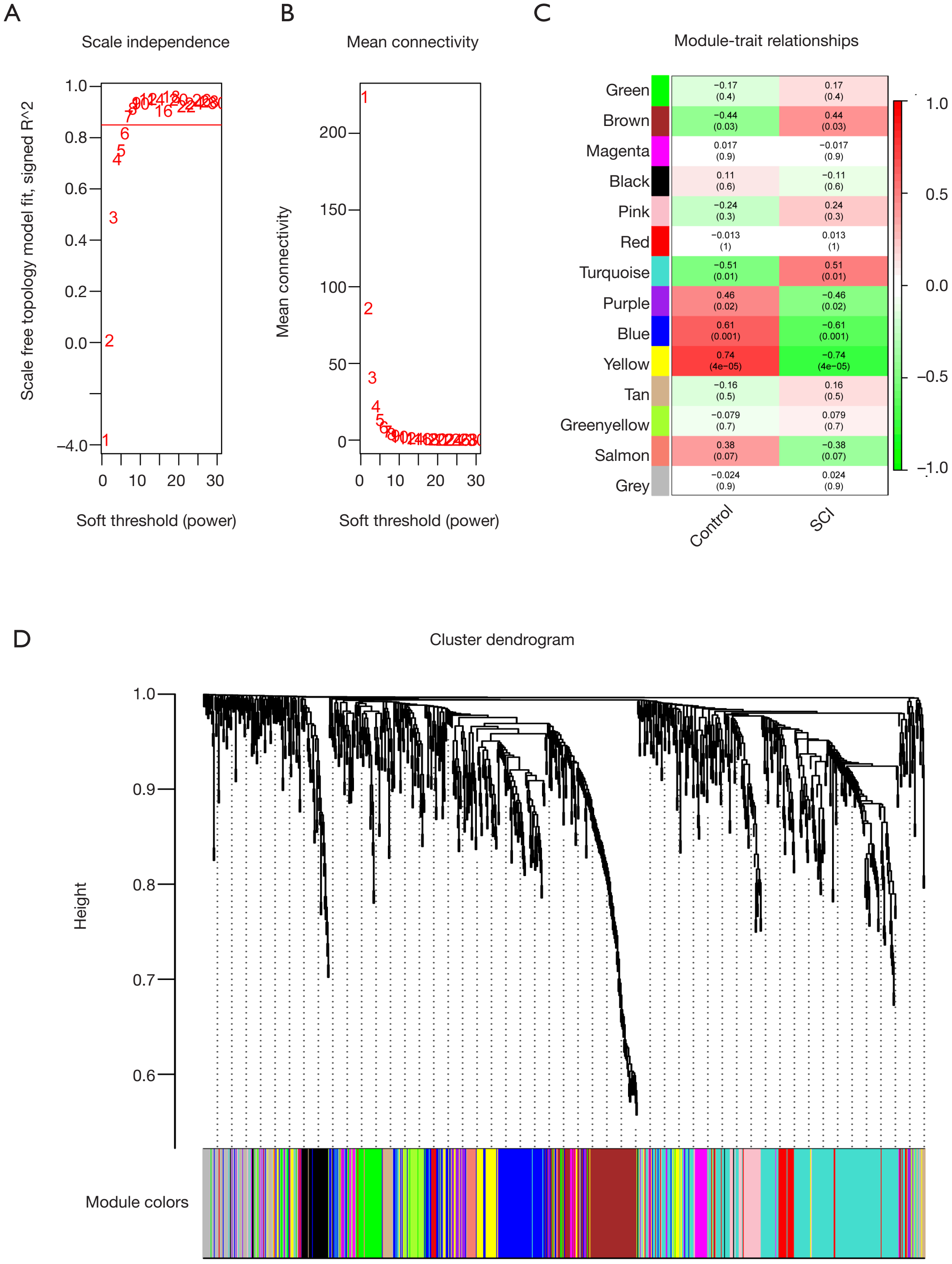


Figure 3 A WGCN was constructed for candidate gene sets using WGCNA. (A) and (B) show the analysis of network topology for various soft-threshold powers, and (C) is a module-feature correlation. The rows represent the eigenvector genes for each module, and the columns represent the sample classification information. The number in each grid represents the correlation coefficient between the gene module and the corresponding sample feature, and the number in the brackets is the P value. (D) indicates the gene dendrogram and shows the module colors. WGCN, weighted gene co-expression network; WGCNA, weighted gene co-expression network analysis.

Table 2 Gene number of each module

Module	No. of genes
Black	38
Blue	117
Brown	86
Green	51
Green yellow	36
Grey	74
Magenta	37
Pink	37
Purple	36
Red	44
Salmon	18
Tan	28
Turquoise	231
Yellow	56

development. Therefore, the gene from the top 10 node degree (degree >80, *Table 3*) in the network was selected as a key gene in the co-expression network.

ncRNA and TFs regulate the module genes

With regard to the ncRNA-mRNA interactions in the RAID 2.0 database for the interaction background, the pivot nodes (ncRNA) regulating the turquoise and brown functional modules were searched. We detected 6 and 15 pivot-TFs and 63 and 79 pivot-ncRNAs ($P<0.05$) for the turquoise and brown modules, respectively. *Table 4* shows 2 modules of pivot TFs. *Table 5* also shows 2 modules of partial pivot-lnc RNAs.

Construction of a multifactor regulatory network

We performed screening of miRNA-lncRNA interaction pairs from the starBase v2.0 database when the number of supporting experiments ≥3 and the number of cancer types ≥1. We obtained 376 miRNA-lncRNA interaction pairs.

The miRNA-mRNA interaction pairs from starBase v3.0 were screened when the number of supporting experiments ≥3 and the number of cancer types ≥1. Then, the interaction pair needed to be identified by any 3 software, such as targetScan, picTar, RNA22, PITA, or miRanda/mirSVR. We obtained 39,219 miRNA-mRNA interaction pairs.

The human TF-mRNA regulatory relationship was determined from the TRRUST v2 database, with a total of 9,396 TF-mRNA interaction pairs.

To construct a multifactor regulatory network, the miRNAs and TFs that interacted with the co-expressing key genes (n=10) were first selected, and the lncRNAs that interacted with these miRNAs were screened. Finally, a multifactor regulatory network with 105 interacting pairs was obtained (Figure S3). The degree of each regulatory factor (miRNA, lncRNA, TF) was calculated, and the top 10 regulatory factors were screened as core regulators (degree >4, *Table 6*).

Functional enrichment analysis of core regulatory factor target genes

To identify target genes for these core regulatory factors, 1,059 target genes interacting with core regulators were screened from miRNA-lncRNA, miRNA-mRNA, and mRNA-TF interaction pairs. To explore the functions and pathways regulated by these core regulatory factors, pathway enrichment analysis of their target genes was performed (*Figure 5*). These target genes were involved in pathways of diseases such as pancreatic cancer and angiogenesis human cytomegalovirus infection.

Analysis of SCI classification

According to the results of the WGCNA co-expression module analysis, the turquoise and brown modules were further analyzed to obtain the top 10 (degree >80) genes and regulatory factors that were strongly related to the top 10 genes in the multifactor regulatory network. The GSE45550 data set was validated, and 18 SCI samples were identified. All SCI samples were classified using the K-means unsupervised clustering method. First, the optimal K-value was selected by finding the inflection point of SSE (the sum of the squares of the distances of all points to the center of the cluster to which it belongs). As seen in *Figure 6A*, the decrease slowed after K=3; hence, K=3 was chosen. We used the R tsne of the R-package to reduce the dimensions of the gene expression data (*Figure 6B*). All SCI samples could be clearly divided into 3 categories. In combination with the expression of key genes in all SCI patients (*Figure 6C*), highly consistent clustering results were observed, as shown in *Figure 6B*. Therefore, we can speculate that the 10 key genes for screening have great

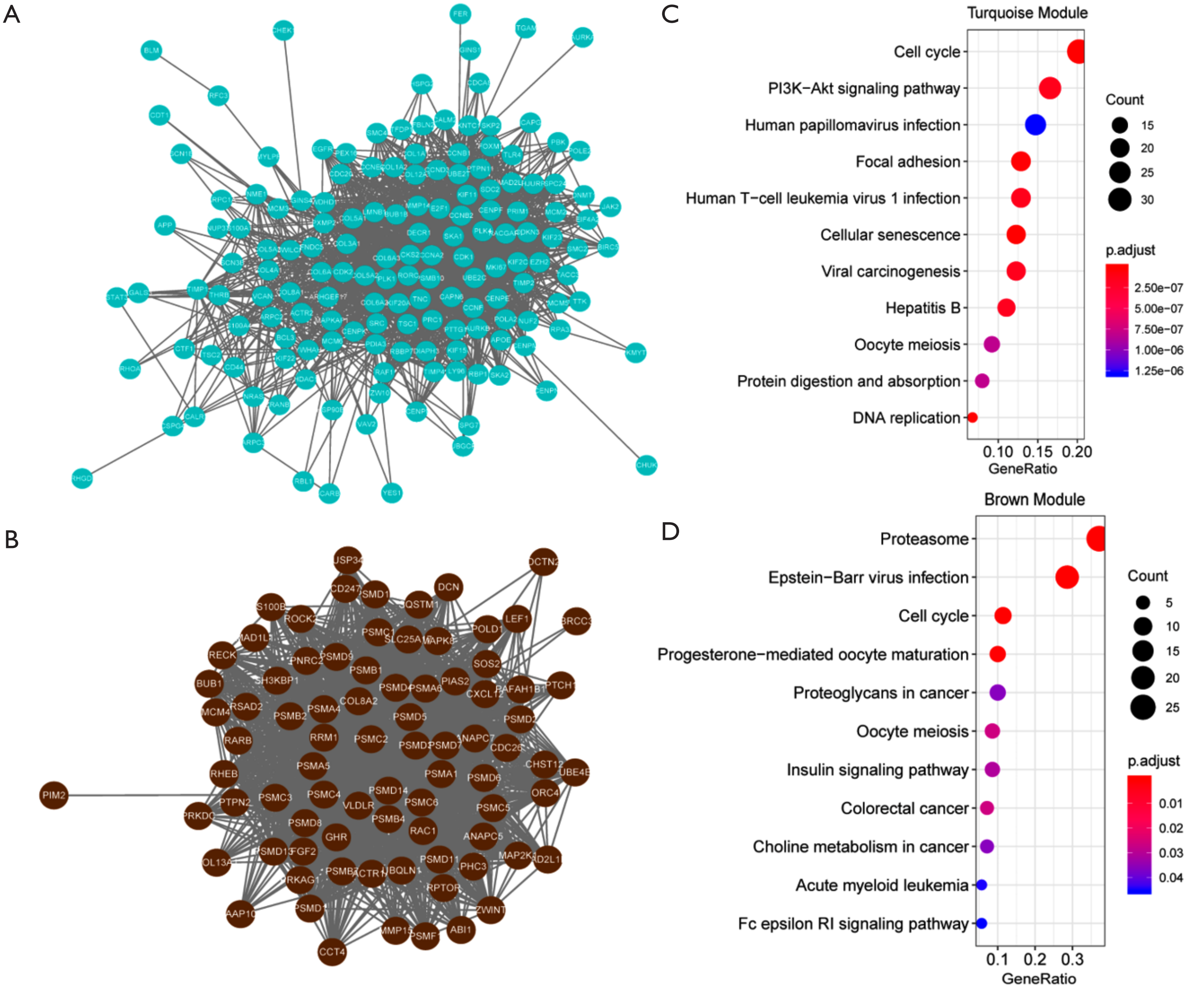


Figure 4 Co-expression network of modules and enriched pathways. (A) Co-expression network of turquoise modules associated with SCI. (B) KEGG pathway enrichment analysis of genes in the turquoise modules. (C) Co-expression network of brown modules related to SCI. (D) KEGG pathway enrichment analysis of the genes in the turquoise modules. KEGG, Kyoto Encyclopedia of Genes and Genomes; SCI, spinal cord injury.

significance for the classification of SCI patients.

Marker gene mining in different subtypes

To further analyze the differences in key genes in different clusters, the random forest model was used to characterize the importance of these 10 genes for different SCI subtypes (Figure S4). We concluded that *CCNB2*, *CCNB1*, *CKS2*, *COL5A1*, *KIF20A*, and *RACGAP1* (top 6) were extremely important for SCI classification. Therefore, the expression of the top 6 genes was evaluated in 3 subclasses. As shown (*Figure* 7), the expression of these 6 genes was significantly different in different clusters. Therefore, the expression of these 6 genes might imply the presence of different SCI subclasses, which is of great significance for selection of the correct drug for patients.

Table 3 The key genes in co-expression networks

Degree	Gene
116	*CDK1*
110	*CKS2*
108	*COL5A1*
105	*COL5A2*
92	*RACGAP1*
89	*KIF20A*
88	*UBE2C*
86	*CCNB2*
83	*TNC*
81	*CCNB1*

Table 4 The pivot transcription factors of the brown and turquoise modules

Symbol	P value	Connection	Module
BRCA2	0.000504	2	Brown
LMO2	0.002935	2	Brown
RARA	0.012915	2	Brown
HDAC1	0.016491	4	Brown
PTTG1	0.033447	2	Brown
HIF1A	0.034374	3	Brown
YBX1	0.000132	10	Turquoise
TP53	0.00024	31	Turquoise
E2F1	0.000375	24	Turquoise
E2F4	0.012355	7	Turquoise
MYCN	0.012853	9	Turquoise
E2F3	0.014587	5	Turquoise
ARID3A	0.022408	3	Turquoise
ETV4	0.024753	5	Turquoise
E4F1	0.034916	2	Turquoise
IRF3	0.034916	2	Turquoise
HCFC1	0.034916	2	Turquoise
NR3C2	0.034916	2	Turquoise
MEF2C	0.048278	3	Turquoise
NR4A1	0.048278	3	Turquoise
SP1	0.048952	36	Turquoise

Table 5 The pivot ncRNAs of the brown and turquoise modules (P<0.01)

Symbol	P value	Connection	Module
hsa-miR-6733-5p	0.000474	11	Brown
hsa-miR-6739-5p	0.000687	11	Brown
hsa-miR-4720-3p	0.001612	6	Brown
hsa-miR-633	0.002974	7	Brown
hsa-miR-3153	0.005435	11	Brown
kshv-miR-K12-6-3p	0.005871	6	Brown
hsa-miR-4524b-5p	0.006406	9	Brown
hsa-miR-5189-3p	0.008139	4	Brown
hsa-miR-331-3p	0.008626	11	Brown
hsa-miR-7161-5p	0.009273	7	Brown
hsa-miR-4524a-5p	0.009585	9	Brown
hsa-miR-4772-5p	0.004152	5	Turquoise
hsa-miR-4721	0.005072	10	Turquoise
hsa-miR-193b-3p	0.005424	60	Turquoise
hsa-miR-6516-5p	0.005561	14	Turquoise
hsa-miR-208a-3p	0.005783	24	Turquoise
hsa-miR-7113-5p	0.006199	15	Turquoise
hsa-miR-215-5p	0.007331	46	Turquoise
hsa-miR-4783-5p	0.007409	3	Turquoise
hsa-miR-7112-5p	0.007409	3	Turquoise
hsa-miR-23b-3p	0.008178	75	Turquoise
hsa-miR-7977	0.008327	24	Turquoise
hsa-miR-6839-3p	0.009031	11	Turquoise

Table 6 The core regulator candidates

Degree	Factor
17	CCNB1
14	XIST
11	CDK1
10	TNC
9	RACGAP1
9	MALAT1
6	H19
5	hsa-miR-29b-3p
5	hsa-miR-29a-3p
5	hsa-miR-106a-5p

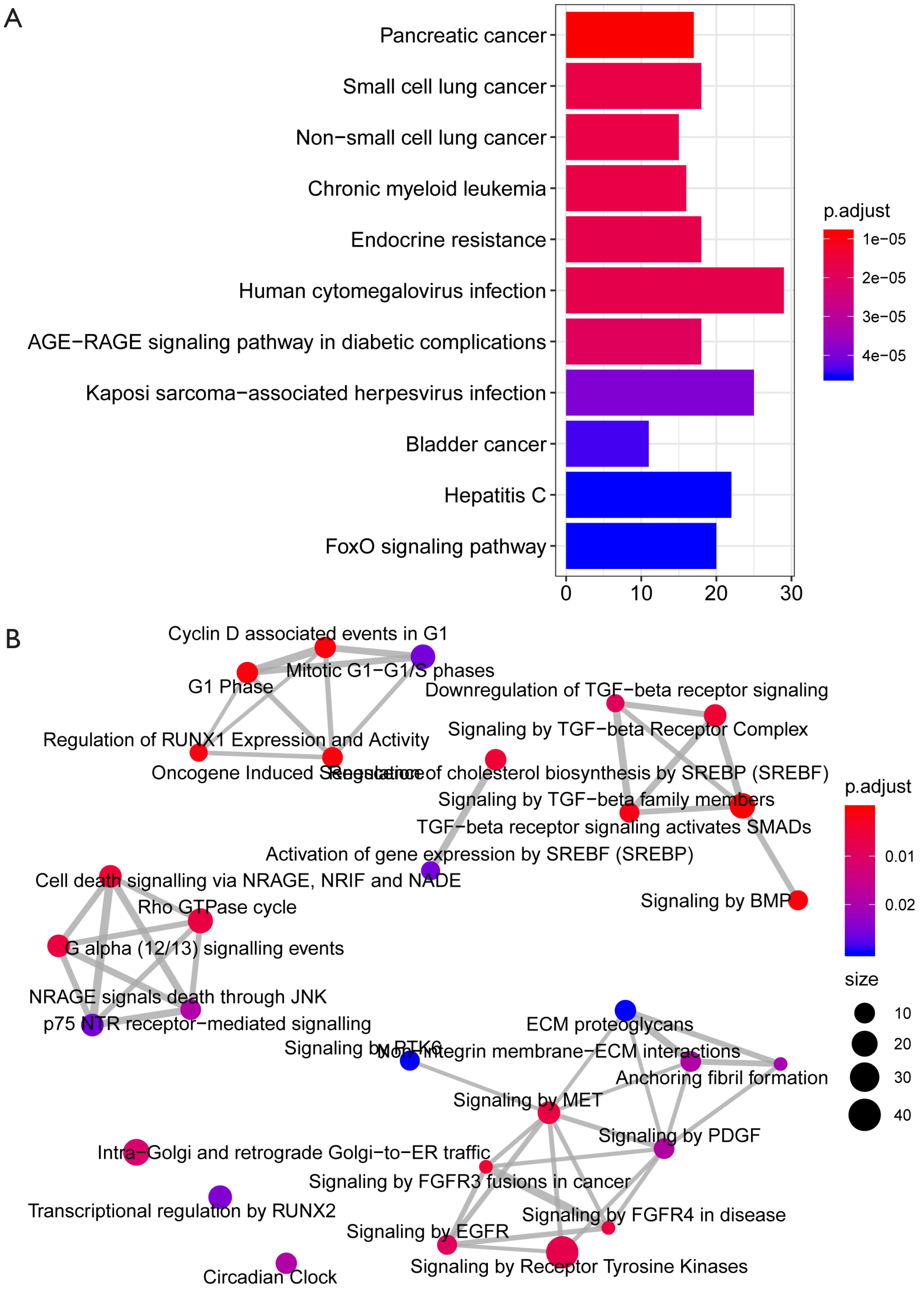


Figure 5 The pathway enrichment analysis of target genes interacting with core regulators from miRNA-lncRNA, miRNA-mRNA, and mRNA-TF interaction pairs. miRNA, micro RNA; lncRNA, long non-coding RNA; mRNA, messenger RNA; TF, transcription factor.

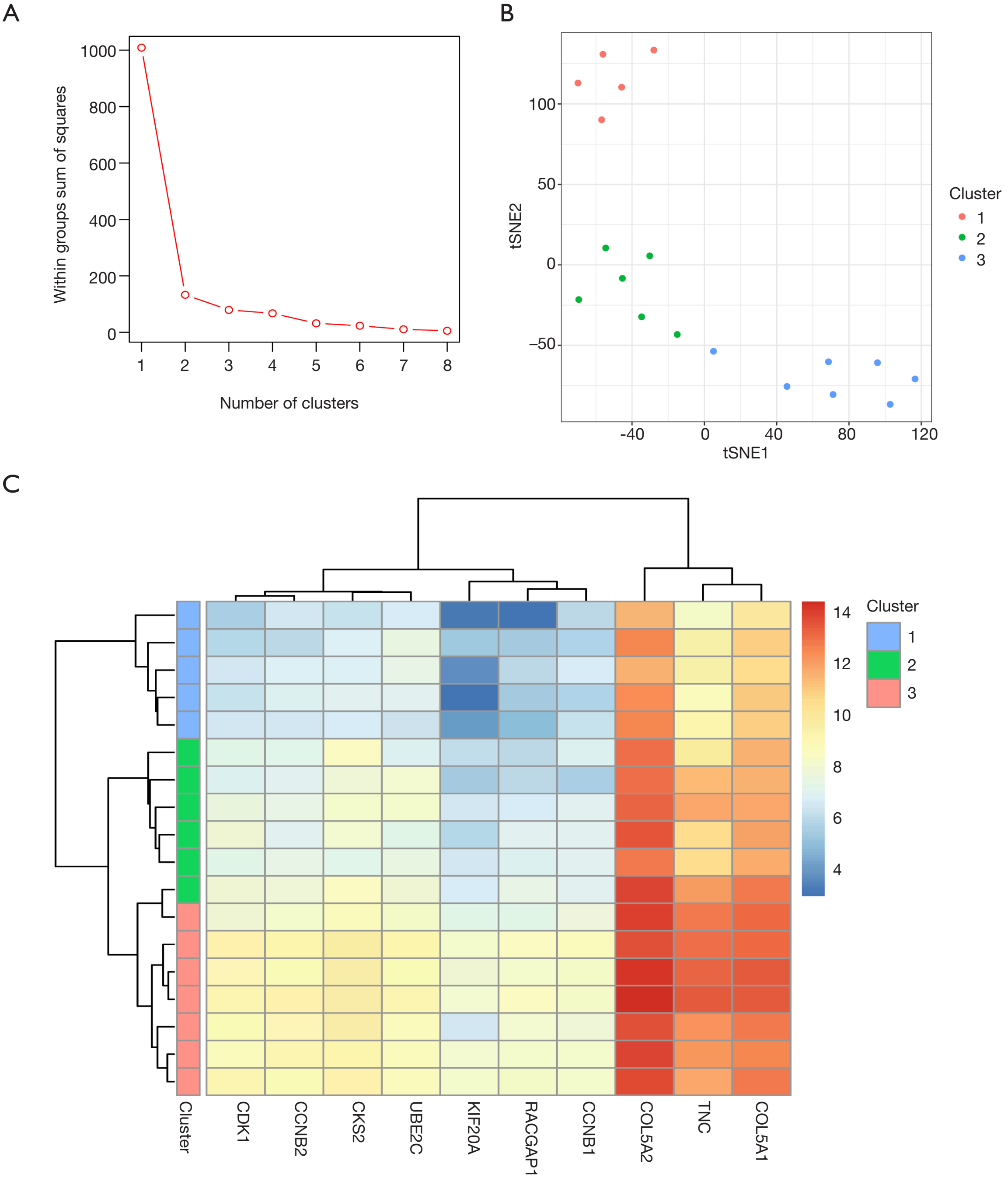


Figure 6 Analysis of SCI classification in GSE45550. (A) The optimal K-value was found on the basis of SSE. (B) All SCI samples were clearly divided into three categories in GSE45550. (C) The expression of 10 key genes occurred during the classification of SCI samples in GSE45550. SCI, spinal cord injury; SSE, sum of the squared error.

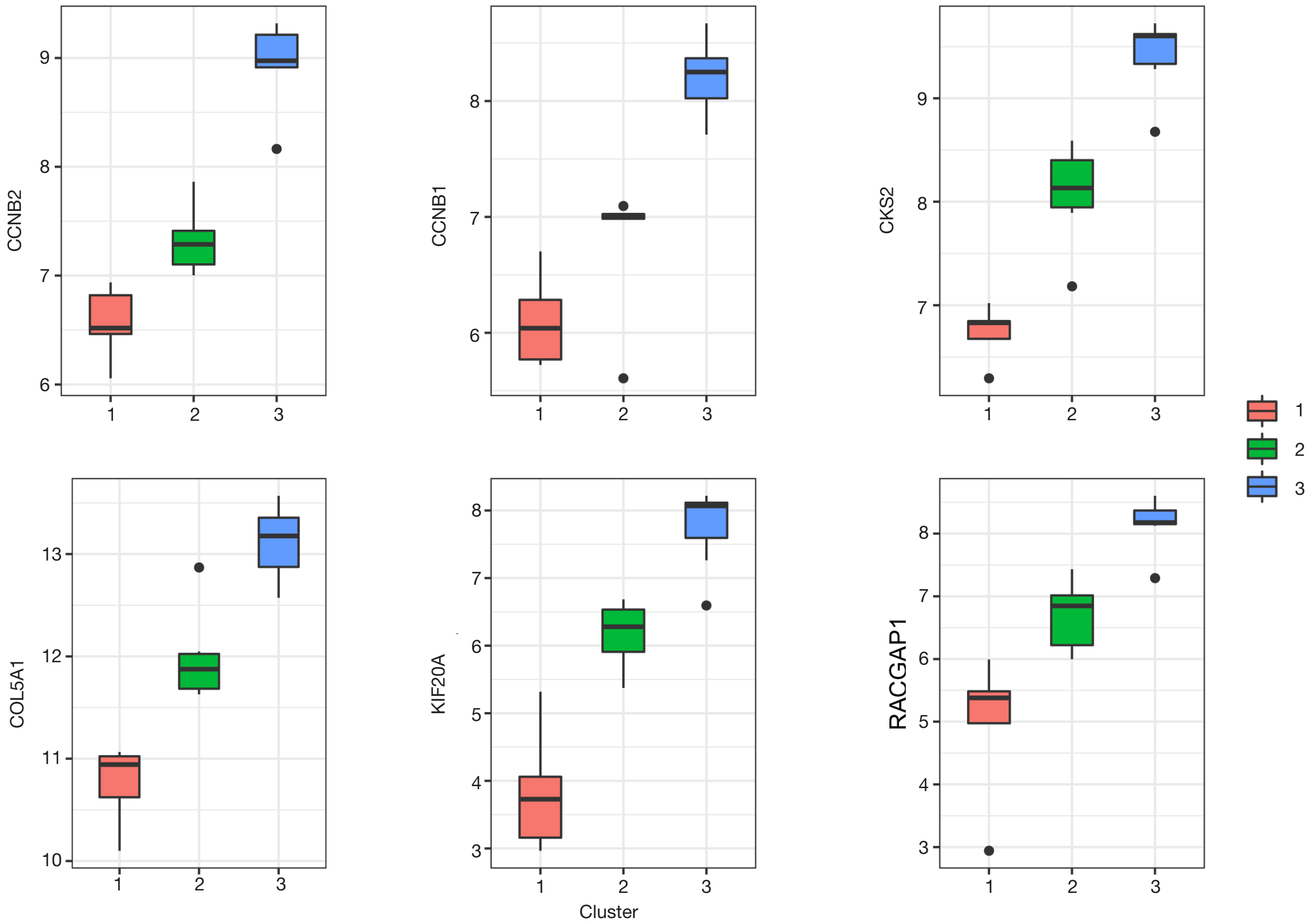


Figure 7 The expression of the top 6 significant genes was evaluated in 3 subclasses.

Table 7 SCI-related drug-gene relationship pairs

Drug	Gene
Etidronic acid	*PTPRS*
Etidronic acid	*ATP6V1A*
Pregabalin	*CACNA1A*
Pregabalin	*SLC1A1*

DrugBank drug analysis

Information regarding drug-gene relationship pairs collected in DrugBank (https://www.drugbank.ca/), SCI treatment drugs etidronic acid and pregalalin, and corresponding drug-related genes were obtained (*Table 7*). Among these, *ATP6V1A* was the target gene of the top 10 regulatory factors hsa-miR-29a-3p and hsa-miR-29b-3p, which further confirmed the reliability of the above analysis.

A total of 8 drugs (AT7519, Alsterpaullone, seliciclib, Indirubin-3'-Monoxime, hymenialdisine, SU9516, olomoucine and alvocidib) were associated with the core gene *CDK1* of SCI, which provided potential therapeutic options for SCI.

Distribution of immune infiltration

The distribution of immune infiltration in mouse cells has not yet been fully uncovered. We first explored immune infiltration in tissues with 71 immune cell subpopulations using the CIBERSORT algorithm. *Figure 8* shows the proportion of immune cells in different colors in each sample, and the length of the bars indicates the level of the immune cell population in the bar graph. From the graph, we found that the tissues with a high percentage of myeloid

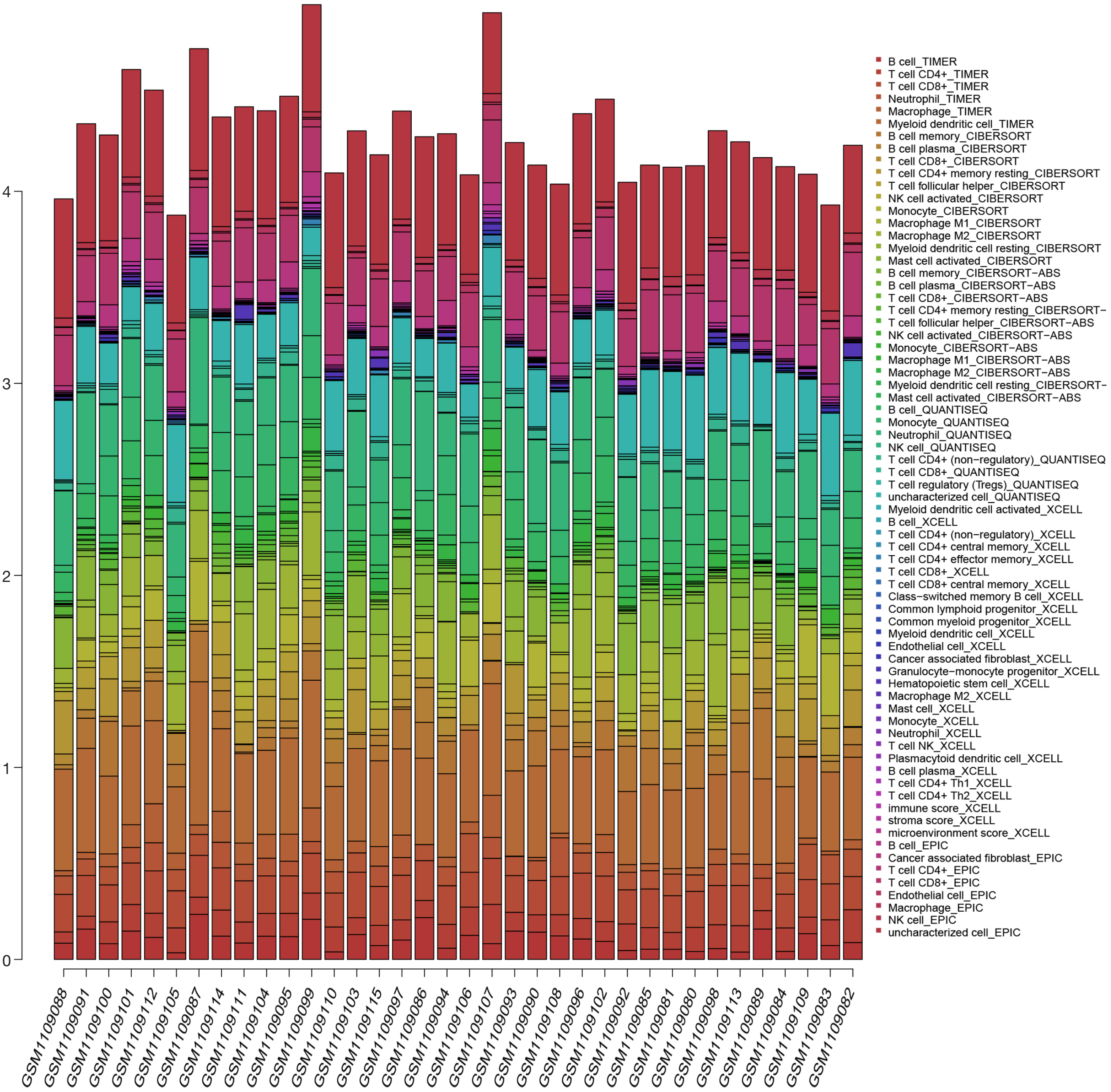


Figure 8 The proportion of immune cells in different colors in each sample, and the length of the bars indicates the level of the immune cell population.

dendritic cells, neutrophil_QUANTISEQ, T cell CD4+ memory restin, and T cell CD8 occupied a very important fraction of immune cells. In contrast, B cell memory, natural killer (NK) cell activation, etc. were relatively low.

As seen in *Figure 9*, there was also a clear correlation between the percentages of the different subsets. Immune cells with significant negative correlations included monocyte and mast cells (−0.75), cancer associated fibroblast and mast cells (−0.72), and T cell regulatory and mast cells (−0.58). Significantly positively correlated immune

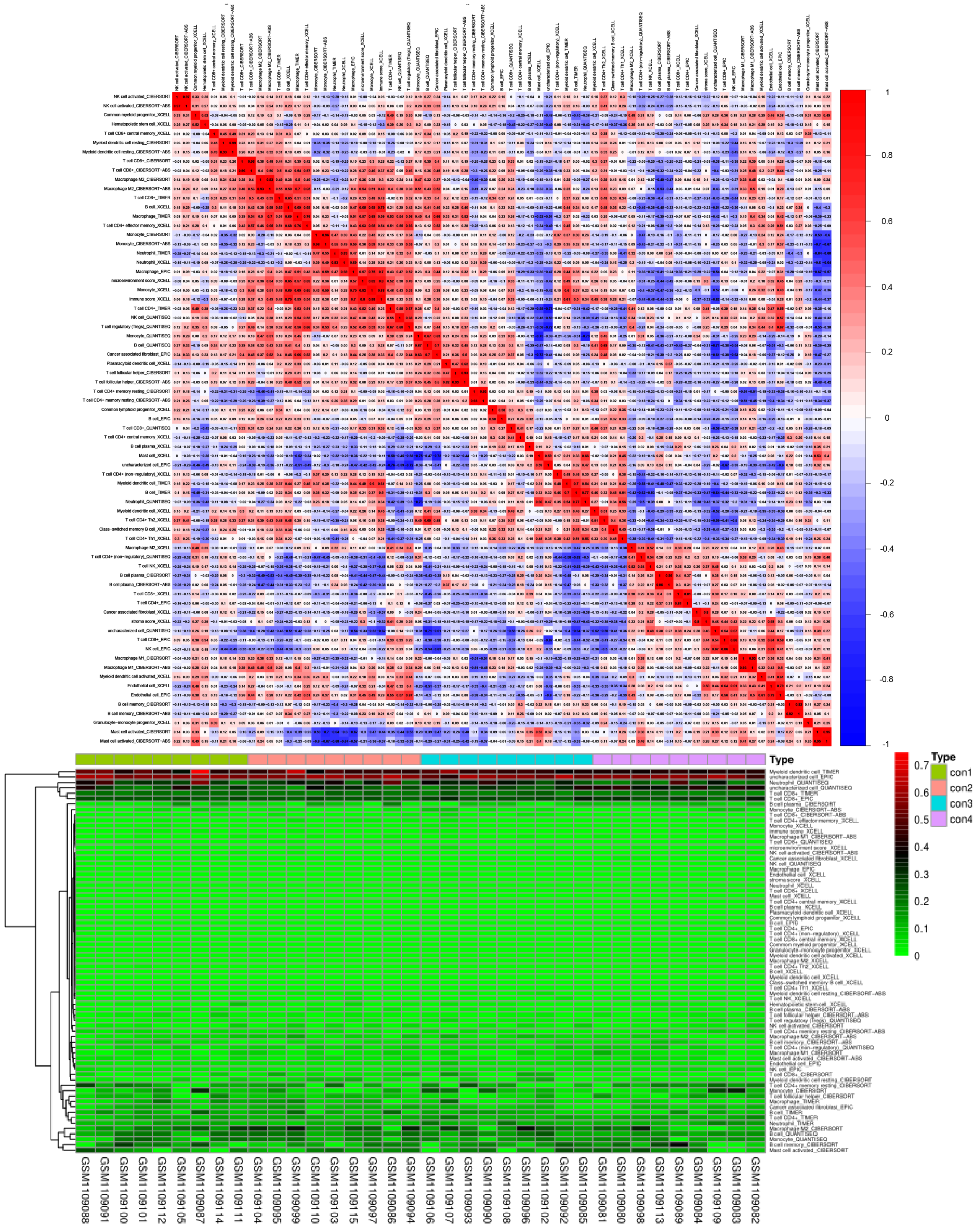


Figure 9 The correlation between the percentages of the different subsets.

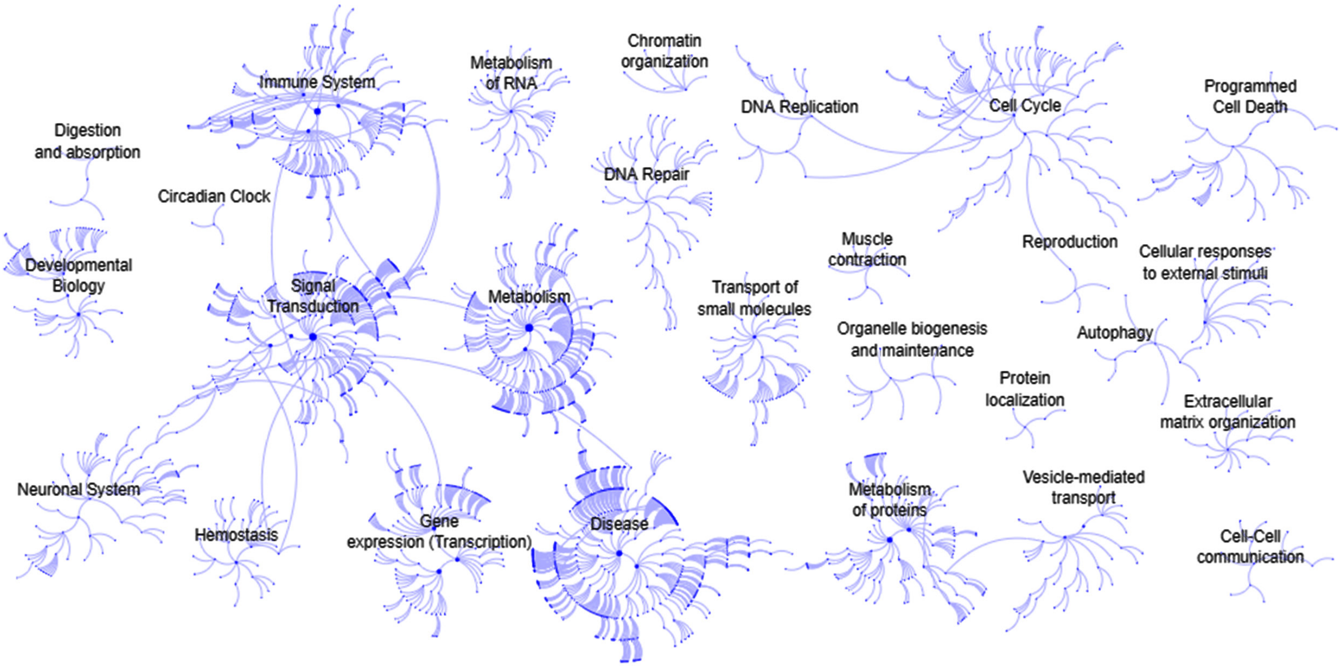


Figure 10 Major human biological research fields covered in the reactome database.

cells included macrophage M2 and T cell CD8+ (0.56), macrophage M2_CIBERSORT and T cell CD4+ (0.65) or microenvironment (0.54), and T cell CD4+ and B cell (0.68). According to the heat map of the above cells (*Figure 10*), the levels of myeloid dendritic cell, neutrophil, and T cell CD8 and B cell plasma in the samples included in the heat map were relatively high. In conclusion, as a regulated process, immune cell infiltration and its heterogeneity in SCI may be of particular clinical relevance.

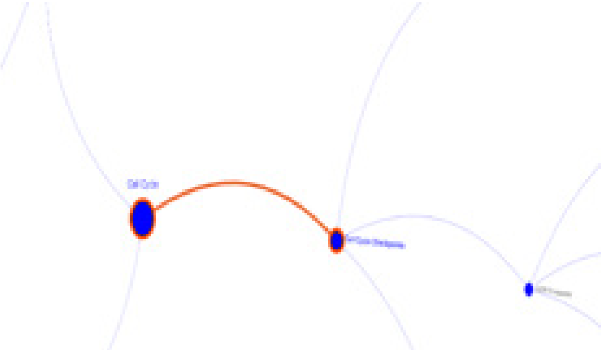
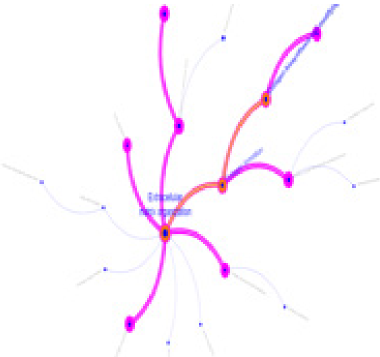
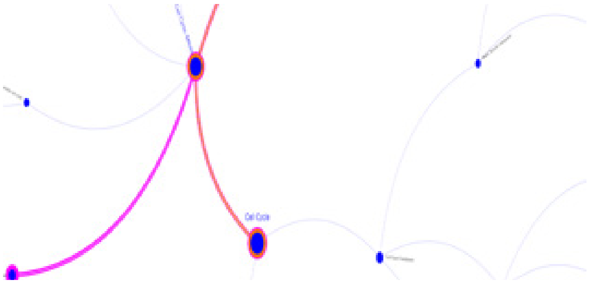
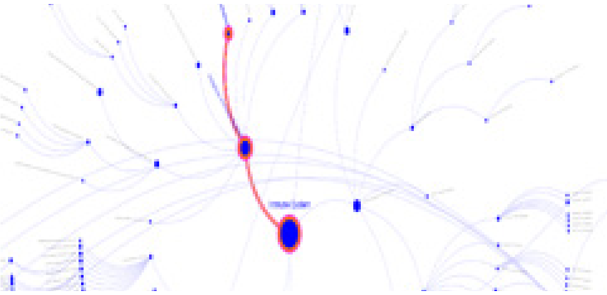
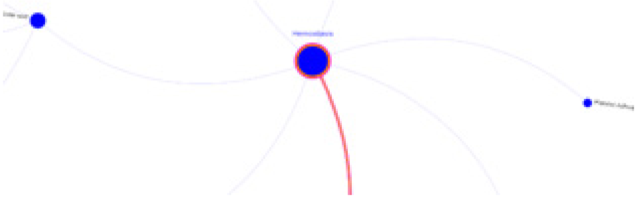
Construction of network reactomics

Reactome is a database of expert-authored, peer-reviewed articles on various responses and biological pathways in the human body. It currently covers more than 70% of the 20,000 manually annotated human proteins in the UniProt database, and it is used for 46 major human biological research fields, such as apoptosis, HIV, influenza virus life cycle, DNA replication, transcription, physiological hemostatic mechanism, and carbohydrate metabolism pathway, among others. The normal function of 1005 proteins related to OMIM disease phenotype was recorded (*Figure 2*). The above-mentioned 6 genes were entered into the reactome database for analysis, and the biological functions of 4 genes were eventually discovered (*Table 8*).

Discussion

The gene expression profile for SCI has been extensively investigated in recent decades. Furthermore, a combination of RNA sequencing and bioinformatics analysis allows for a comprehensive analysis of changes in gene and RNA expression in SCI patients (6). No effective therapies have yet been developed for restoring neural circuits following SCI (22). Previous studies have indicated that chondroitinase ABC and anti-NogoA treatment could promote axonal sprouting and functional recovery by reducing inhibitory signaling in the environment after SCI (23,24). Neural stem cell grafts containing growth factors elicit long-distance axon regeneration (25,26). Additionally, electrical epidural stimulation and brain–machine interfaces have reportedly improved motor functions after SCI (27-30). However, SCI is a complicated pathological condition. Although several therapeutic strategies have shown potential in preclinical studies, few have moved to the clinical trial stage (30-32). Although surgical decompression and high-dose methylprednisolone are the primary clinical treatments for SCI, their efficacy remains unclear (33).

Table 8 The biological functions of four genes obtained from the reactome database.

Target protein	Main composition diagram	Function
Ccnb1		Extracellular matrix organization (Rattus norvegicus)
Col5a1		Extracellular matrix organization (Rattus norvegicus)
Kif20a		Cell Cycle (Rattus norvegicus)
Racgap1	 	Hemostasis (Rattus norvegicus), Immune System (Rattus norvegicus)

Understanding the spatial and temporal difference in the transcription profiles of chosen subpopulations after SCI could aid discovery of the key proteins with potential uses in clinical application.

As SCI is a result of multiple pathological processes, different therapeutic targets are identified in different patients. Hence, SCI could be further classified for the provision of personalized treatment to patients. It is clinically important to identify the SCI subtype, construct the regulatory network involved with SCI according to the pathological process, and discover potentially effective drugs for SCI. Here, co-expression patterns in SCI were identified using WGCNA, a powerful bioinformatics method (1). Among the 14 identified co-expression modules, the turquoise and brown modules were considered to be closely correlated with SCI. Among them, the turquoise module mainly participated in the cell cycle, cell senescence, and PI3K-Akt signaling pathway, and the brown module mainly participated in Epstein-Barr virus infection, insulin signaling, and the proteasome pathway. These results were consistent with those of previous studies. Previous studies have shown that cell cycle modulation and PI3K-Akt signaling are important therapeutic strategies in SCI (34-37). Using the co-expression relationship of the 2 modules in combination with the results after validation of spinal cord samples (GSE45550), 6 key genes (*CCNB2*, *CCNB1*, *CKS2*, *COL5A1*, *KIF20A*, and *RACGAP1*) were known to have an important role in the classification of SCI. The genes *CCBN1* and *COL5A1* are reportedly associated with SCI (38,39). Our results were in agreement with those of previous studies.

Our results indicated that *CCNB2*, *CCNB1*, *CKS2*, *COL5A1*, *KIF20A*, and *RACGAP1* could be used to categorize SCI. Currently, SCI research focuses on neuroplasticity and regeneration (22), and the SCI subtypes have not yet been identified on a genetic level. Our results may provide significant information regarding SCI subtypes.

Based on SCI classification, we explored the use of 2 drugs for treating SCI via DrugBank. Etidronic acid is an osteoporosis drug used in clinics (40), and pregabalin is used to manage chronic pain (41). According to our results, another 8 drugs (AT7519, Alsterpaullone, seliciclib, Indirubin-3'-Monoxime, hymenialdisine, SU9516, olomoucine and alvocidib) may have therapeutic effects on SCI. These 8 drugs were involved in the cell cycle. Based on the SCI subtype, we need to examine whether these drugs could effectively be used for SCI treatment.

There were several limitations to this study. We did not have enough data to construct an SCI expression profile. The SCI subtype and use of expected drugs may also appear inaccurate as sample numbers were insufficient, which affected the evaluation of the WGCNA co-expression module of DEGs, miRNA, lncRNA, and TFs.

With the rapid growth of microarray data and RNA sequencing, we believe that our method could have potential applications for SCI patients. In addition, our future research would aim to verify the potential drugs that might play an important therapeutic role in SCI.

Acknowledgments

We thank all the research staff members for their contributions to this study.

Funding: This work was supported by Agricultural and Social Research Projects in the Jiading District of Shanghai. This study was funded by the Zhejiang Medical and Health Science and Technology Project (2019KY342).

Footnote

Reporting Checklist: The authors have completed the MDAR reporting checklist. Available at http://dx.doi.org/10.21037/atm-21-340

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at http://dx.doi.org/10.21037/atm-21-340). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013).

References

1. Duan H, Ge W, Zhang A, et al. Transcriptome analyses reveal molecular mechanisms underlying functional recovery after spinal cord injury. Proc Nat Acad Sci USA 2015;112:13360-5.
2. Tran AP, Warren PM, Silver J. The biology of regeneration failure and success after spinal cord injury. Physiol Rev 2018;98:881-917.
3. Bloom O, Herman PE, Spungen AM. Systemic inflammation in traumatic spinal cord injury. Exp Neurol 2020;325:113143.
4. Venkatesh I, Blackmore MG. Selecting optimal combinations of transcription factors to promote axon regeneration: Why mechanisms matter. Neurosci Lett 2017;652:64-73.
5. Courtine G, Sofroniew MV. Spinal cord repair: advances in biology and technology. Nat Med 2019;25:898-908.
6. Wang W, Su Y, Tang S, et al. Identification of noncoding RNA expression profiles and regulatory interaction networks following traumatic spinal cord injury by sequence analysis. Aging (Albany NY) 2019;11:2352-68.
7. Ye S, Yang L, Zhao X, et al. Bioinformatics method to predict two regulation mechanism: TF–miRNA–mRNA and lncRNA–miRNA–mRNA in pancreatic cancer. Cell Biochem Biophys 2014;70:1849-58.
8. Dong M, Wang X, Zhao HL, et al. Integrated analysis of transcription factor, microRNA and LncRNA in an animal model of obliterative bronchiolitis. Int. J Clin Exp Pathol 2015;8:7050.
9. Takemata N, Ohta K. Role of non-coding RNA transcription around gene regulatory elements in transcription factor recruitment. RNA Biol 2017;14: 1-5.
10. Yin L, Cai Z, Zhu B, et al. Identification of key pathways and genes in the dynamic progression of HCC based on WGCNA. Genes 2018;9:92.
11. Liu X, Hu AX, Zhao JL, et al. Identification of key gene modules in human osteosarcoma by co-expression analysis weighted gene co-expression network analysis (WGCNA). J Cell Biochem 2017;118:3953-9.
12. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 2008;9:559.
13. Liu Z, Li M, Fang X, et al. Identification of surrogate prognostic biomarkers for allergic asthma in nasal epithelial brushing samples by WGCNA. J Cell Biochem 2019;120:5137-50.
14. Squair JW, Tigchelaar S, Moon KM, et al. Integrated systems analysis reveals conserved gene networks underlying response to spinal cord injury. Elife 2018;7:e39188.
15. Cobos EJ, Nickerson CA, Gao F, et al. Mechanistic differences in neuropathic pain modalities revealed by correlating behavior with global expression profiling. Cell Rep 2018;22:1301-12.
16. Chen JA. Gene co-expression network analysis implicates microRNA processing in Parkinson's disease pathogenesis. Neurodegener Dis 2018;18:191-9.
17. Liang JW, Fang ZY, Huang Y, et al. Application of weighted gene co-expression network analysis to explore the key genes in alzheimer's disease. J Alzheimers Dis 2018;65:1353-64.
18. Wang T, Wu B, Zhang X, et al. Identification of gene coexpression modules, hub genes, and pathways related to spinal cord injury using integrated bioinformatics methods. J Cell Biochem2019;120:6988-97.
19. Guo G, Ren S, Kang Y, et al. Microarray analyses of lncRNAs and mRNAs expression profiling associated with diabetic peripheral neuropathy in rats. J Cell Biochem 2019;120:15347-59.
20. Cabili MN, Trapnell C, Goff L, et al. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev 2011;25:1915-27.
21. Liao Q, Liu C, Yuan X, et al. Large-scale prediction of long non-coding RNA functions in a coding–non-coding gene co-expression network. Nucleic Acids Res 2011;39:3864-78.
22. Hutson TH, Giovanni SD. The translational landscape in spinal cord injury: focus on neuroplasticity and regeneration. Nat Rev Neurol 2019;15:732-45.
23. Pakulska MM, Tator CH, Shoichet MS. Local delivery of chondroitinase ABC with or without stromal cell-derived factor 1 alpha promotes functional repair in the injured rat spinal cord. Biomaterials 2017;134:13-21.
24. Freund P, Schmidlin E, Wannier T, et al. Nogo-A–specific antibody treatment enhances sprouting and functional recovery after cervical lesion in adult primates. Nat Med 2006;12:790-2.
25. Lu P, Wang Y, Graham L, et al. Long-distance growth and connectivity of neural stem cells after severe spinal cord injury. Cell 2012;150:1264-73.
26. Lu P, Woodruff G, Wang Y, et al. Long-distance axonal growth from human induced pluripotent stem cells after spinal cord injury. Neuron 2014;83:789-96.
27. Capogrosso M, Milekovic T, Borton D, et al. A brain–

spine interface alleviating gait deficits after spinal cord injury in primates. Nature 2016;539:284-8.
28. Moraud EM, Capogrosso M, Formento E, et al. Mechanisms underlying the neuromodulation of spinal circuits for correcting gait and balance deficits after spinal cord injury. Neuron 2016;89:814-28.
29. Alam M, Rodrigues W, Pham BN, et al. Brain-machine interface facilitated neurorehabilitation via spinal stimulation after spinal cord injury: Recent progress and future perspectives. Brain Res 2016;1646:25-33.
30. Bonizzato M, Pidpruzhnykova G, DiGiovanna J, et al. Brain-controlled modulation of spinal circuits improves recovery from spinal cord injury. Nat Commun 2018;9:3015.
31. Rao JS, Zhao C, Zhang A, et al. NT3-chitosan enables de novo regeneration and functional recovery in monkeys after spinal cord injury. Proc Nat Acad Sci USA 2018;115:E5595-604.
32. Dulin JN, Adler AF, Kumamaru H, et al. Injured adult motor and sensory axons regenerate into appropriate organotypic domains of neural progenitor grafts. Nat Commun 2018;9:84.
33. Ahuja CS, Martin AR, Fehlings M. Recent advances in managing a spinal cord injury secondary to trauma. F1000Res 2016;5:F1000.
34. Wu J, Zhao Z, Zhu X, et al. Cell cycle inhibition limits development and maintenance of neuropathic pain following spinal cord injury. Pain 2016;157:488-503.
35. Li H, Zhang X, Qi X, et al. Icariin inhibits endoplasmic reticulum stress-induced neuronal apoptosis after spinal cord injury through modulating the PI3K/AKT signaling pathway. Int J Biol Sci 2019;15:277-86.
36. Chen J, Wang Z, Zheng Z, et al. Neuron and microglia/macrophage-derived FGF10 activate neuronal FGFR2/PI3K/Akt signaling and inhibit microglia/macrophages TLR4/NF-κB-dependent neuroinflammation to improve functional recovery after spinal cord injury. Cell Death Dis 2017;8:e3090.
37. Chen CH, Sung CS, Huang SY, et al. The role of the PI3K/Akt/mTOR pathway in glial scar formation following spinal cord injury. Exp Neurol 2016;278:27-41.
38. Shi Z, Ning G, Zhang B, et al. Signatures of altered long noncoding RNAs and messenger RNAs expression in the early acute phase of spinal cord injury. J Cell Physiol 2019;234:8918-27.
39. Didangelos A, Puglia M, Iberl M, et al. High-throughput proteomics reveal alarmins as amplifiers of tissue pathology and inflammation after spinal cord injury. Sci Rep 2016;6:21607.
40. Papathanasiou KE, Turhanen P, Brückner SI, et al. Smart, programmable and responsive injectable hydrogels for controlled release of cargo osteoporosis drugs. Sci Rep 2017;7:4743.
41. Derry S, Bell RF, Straube S, et al. Pregabalin for neuropathic pain in adults. Cochrane Database Syst Rev 2019;1:CD007076.

(English Language Editor: J. Jones)

Cite this article as: Chen Q, Zhao Z, Yin G, Yang C, Wang D, Feng Z, Ta N. Identification and analysis of spinal cord injury subtypes using weighted gene co-expression network analysis. Ann Transl Med 2021;9(6):466. doi: 10.21037/atm-21-340

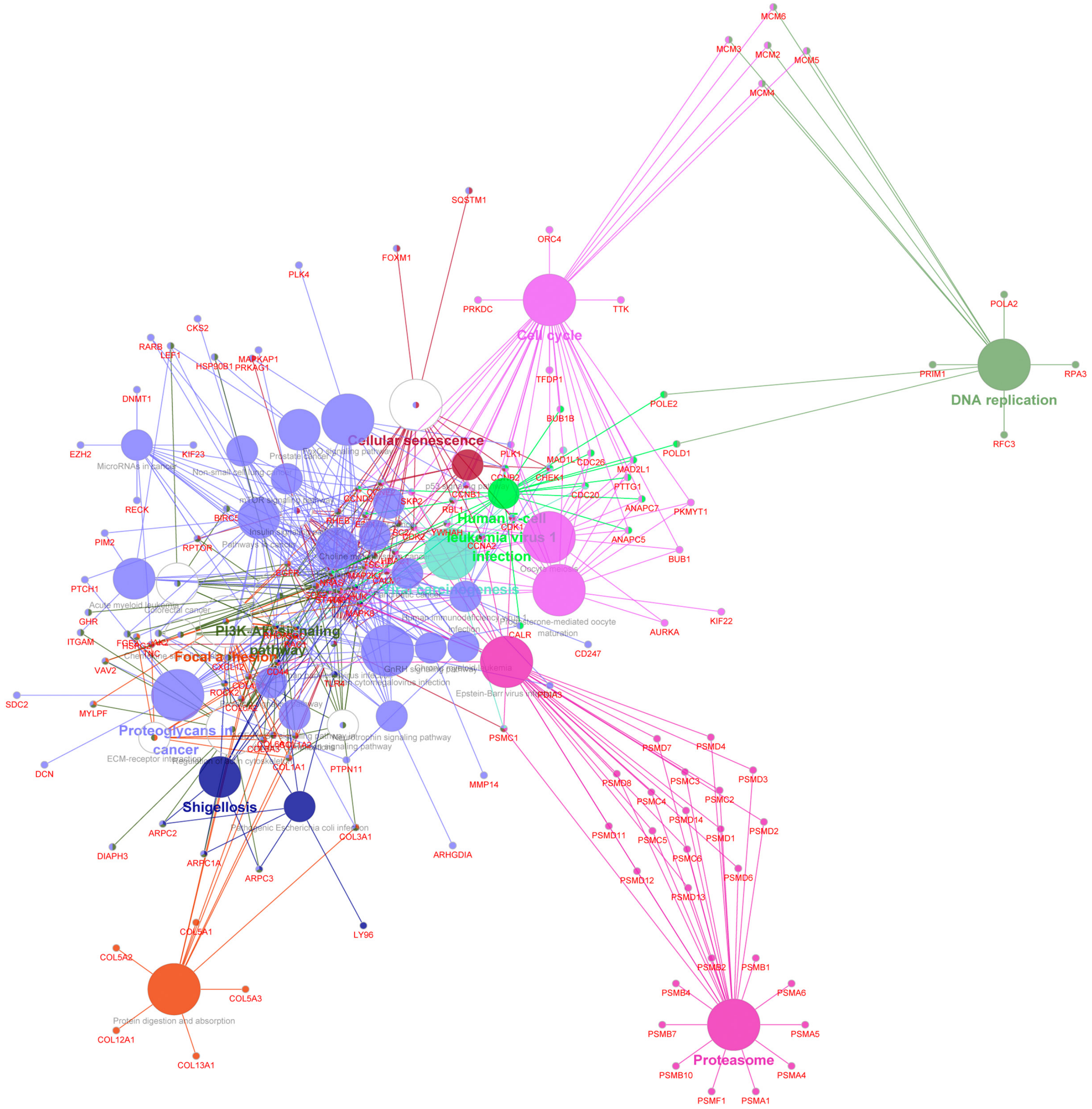


Figure S1 The annotation of the turquoise and brown module co-expression networks in the KEGG Pathway.

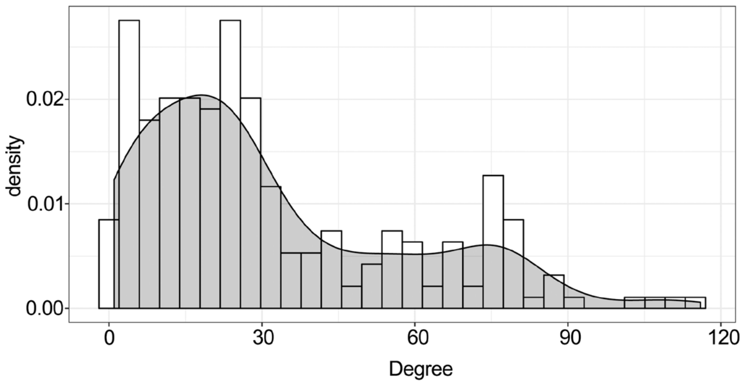


Figure S2 The degree distribution of the sub-network.

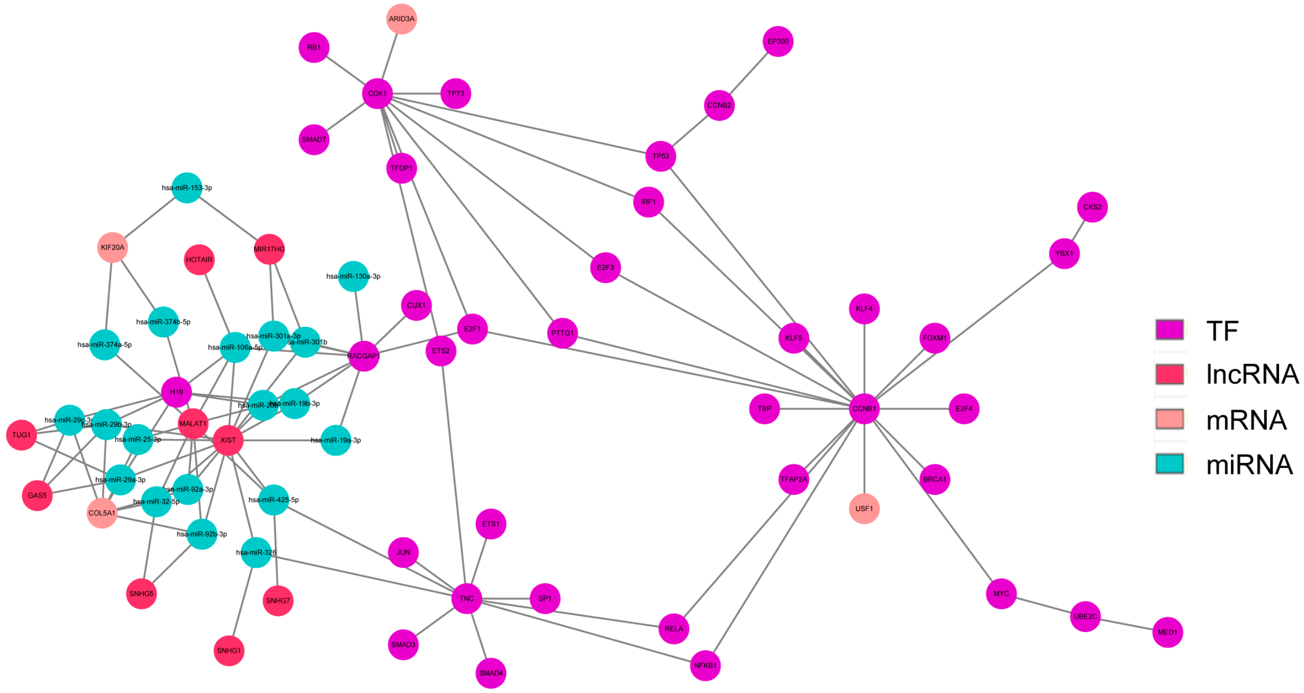


Figure S3 The multi-factor regulatory network of the miRNAs, lncRNAs, and transcription factors. The miRNAs and transcription factors that interacted with co-expressing key genes (N=10) were selected and the lncRNAs that interacted with these miRNAs were screened out.

http://dx.doi.org/10.21037/atm-21-340

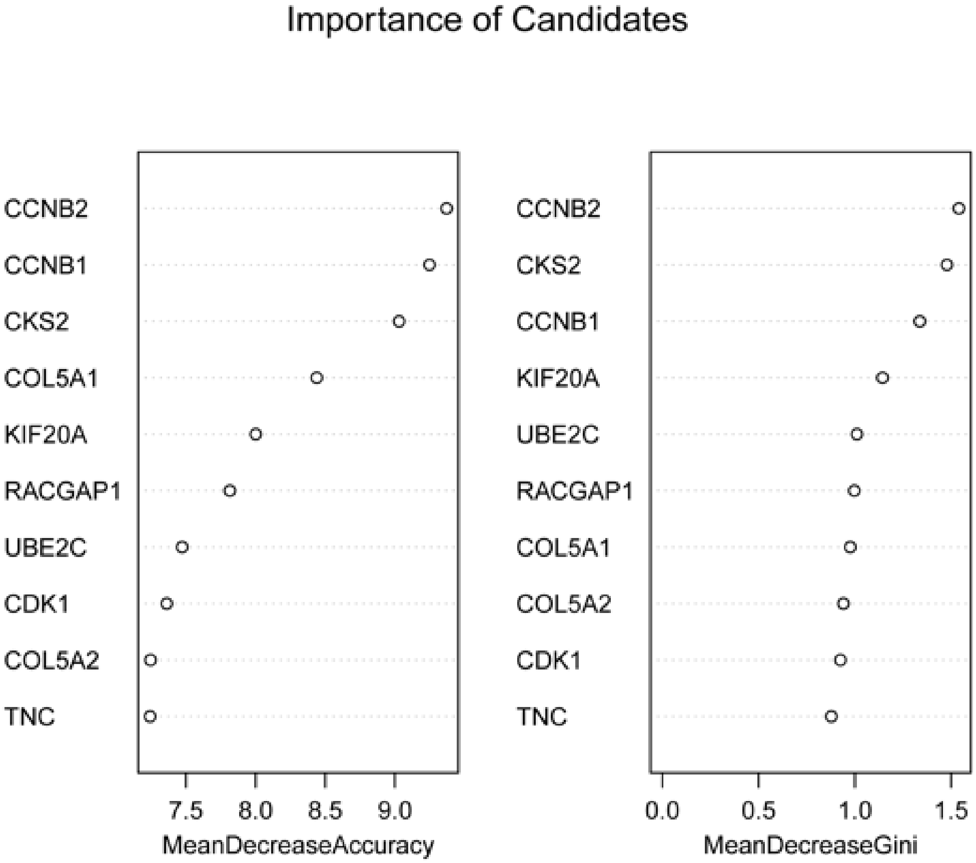


Figure S4 The significance of these 10 genes for different subtypes of SCI based on the random forest model.

http://dx.doi.org/10.21037/atm-21-340