Identification of sialylation-related genes as a new diagnostic biomarker of atherosclerosis by machine learning

Wei Song
Tai'an Central Hospital (Tai'an Central Hospital affiliated to Qingdao University, Taishan Medical Center)

Jian-feng Xue
Tai'an Central Hospital (Tai'an Central Hospital affiliated to Qingdao University, Taishan Medical Center)

Ji-fang Liu
Tai'an Central Hospital (Tai'an Central Hospital affiliated to Qingdao University, Taishan Medical Center)

Yue Feng
Zaozhuang Shizhong District People's Hospital

Na Yao
Tai'an Central Hospital (Tai'an Central Hospital affiliated to Qingdao University, Taishan Medical Center)

Lu-hua Yin
Tai'an Central Hospital (Tai'an Central Hospital affiliated to Qingdao University, Taishan Medical Center)

Hong-Peng Yang
Tai'an Central Hospital (Tai'an Central Hospital affiliated to Qingdao University, Taishan Medical Center)

Hai-peng Sun (shp20230313@163.com)
Tai'an Central Hospital (Tai'an Central Hospital affiliated to Qingdao University, Taishan Medical Center)

Pu Zhang
Tai'an Central Hospital (Tai'an Central Hospital affiliated to Qingdao University, Taishan Medical Center)

Research Article

Keywords: Atherosclerosis, Sialylation, differentially expressed genes, machine learning, Immune cell infiltration

Posted Date: August 29th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3265278/v1

License: ☺️ This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License
Abstract

Background:

Atherosclerosis (AS) is a pathological change based on the disorder of lipid metabolism, which is related to the inflammatory process of vascular wall and the high level of low-density lipoprotein. Sialylation is a post-translational modification controlled by sialyltransferase, transporter and neuraminidase family. This process plays a key role in a variety of biological functions. Abnormal sialylation is related to a variety of diseases, including cancer, pathogen infection and cardiovascular disease (CVD). Therefore, this study aims to explore the role of sialylation related genes in AS.

Methods:

Two AS data sets were obtained from the gene expression comprehensive database (GEO). Based on the differentially expressed genes (DEGs) and the sialylation gene set, the differentially expressed sialylation-related genes (De-SRGs) were found. Then, machine learning method is used to find the core gene. The immune cell infiltration method was established to study the immune cell imbalance in AS. Subsequently, we explored two different subtypes based on core genes using 158 AS samples. Gene ontology (GO), Kyoto Encyclopedia of Genes and Genome (KEGG) pathway enrichment, gene set variation analysis (GSVA) and immunoinfiltration analysis are also used to evaluate the different roles of subtypes.

Results:

A total of 36 De-SRGs were identified. Through machine learning algorithm, 5 core genes were identified and 2 subtypes related to core genes were defined. The results of GSVA showed that type A inflammatory response related pathways were significantly enriched, while type B inflammatory response related pathways were significantly enriched.

Conclusion:

Through this study, we have revealed the relationship between Sialylation-related genes and AS, as well as the heterogeneity of AS patients with different Sialylation subtypes. Selecting a Sialylation-signature based on five genes as the best machine learning model can accurately evaluate the diagnosis of AS and control patients. Our research results reveal the progress of Sialylation in AS for the first time, and provide new insights for its potential pathogenesis and potential treatment strategies.

INTRODUCTION

Atherosclerosis (AS) is a pathological change based on the disorder of lipid metabolism. Lipid deposition, fibrous tissue hyperplasia and calcification can be seen in the intima of the diseased artery, and then the degeneration and calcification gradually appear in the middle layer of the artery wall, which leads to the thickening of the artery wall, the reduction of elasticity, and the narrowing of blood vessels, often leading to plaque rupture, bleeding, thrombosis and other complications. AS usually involves large and middle
muscular arteries. Once it develops enough to block the arterial cavity, the relevant organs and tissues will suffer from ischemia and necrosis, which is the main cause of coronary heart disease, cerebral infarction and peripheral vascular disease (1, 2). AS is related to the inflammatory process of blood vessel wall and the high level of low density lipoprotein(3), but the pathogenic gene or molecular mechanism of AS is still unclear, which needs further study to provide basis for targeted treatment and early gene therapy of AS.

Sialic acid is a negatively charged carbohydrate that acts as the terminal residue of oligosaccharide chain of glycoprotein or glycolipid in eukaryotic cells. Sialylation is a post-translational modification controlled by sialyltransferase, transporter and neuraminidase family. This process plays a key role in a variety of biological functions, including regulating the interaction between leukocytes and endothelial cells, signal transduction, maintaining normal protein conformation and intracellular transport (4–11). Abnormal sialylation is associated with a variety of diseases, including cancer, pathogen infection and cardiovascular disease (CVD)(12, 13). However, the mechanism of sialylation in AS is not clear.

In this study, we aim to identify new AS diagnostic genes based on bioinformatics and machine learning. We analyzed the data sets GSE21545 and GSE43292 to determine the differentially expressed sialylation-related genes (De-SRGs) between AS and healthy samples, screened the core gene based on machine learning, constructed the miRNA-TF-mRNA regulatory network of hub gene, and the correlation between hub gene and immune infiltration, providing a new idea for the diagnosis and treatment of this disease.

**MATERIALS AND METHODS**

**Data resources**

Two sets of transcriptome data in this study are from gene expression database (GEO) (http://www.ncbi.nlm.nih.gov/geo). The GSE21545 microarray data sample includes the gene expression of peripheral blood mononuclear cells and carotid plaque in 223 patients undergoing endarterectomy. For the convenience of research, we selected 126 of them for the gene expression of carotid plaque. The GSE21545 microarray data samples include the gene expression of carotid plaque in 32 patients with hypertension and the gene expression of normal carotid tissue adjacent to the plaque. The two sets of data sets are respectively based on the platform of Affymetrix Human Gene 1.0 ST Array and Affymetrix Human Genome U133 Plus 2.0 Array. Sialylation-related genes (SRGs) are identified through the Molecular Signatures Database (MSigDB) (14–16). Marker genes of different immune cells are from TISIDB database(http://cis.hku.hk/TISIDB/download.php). Use regnetwork database(https://regnetworkweb.org/). Predict upstream miRNA and transcription factor (TF).

**Data processing**

Combine the data in GSE21545 and GSE43292, and use R packets "lima" and "sva" to remove the batch effect, reduce the differences between data sets due to experimental errors, and recombine multiple data sets, so that downstream analysis can only consider biological differences.

**Screening and analysis of De-SRGs**
Limma package of R language was used to analyze the gene expression of AS groups and control groups in the combined data set. In order to expand the screening conditions, only false discovery rate (FDR) was used to screen DEGs, limiting the PDR value to < 0.05. Cluster thermal map and volcanic map display of DEGs.

The ClusterProfiler package of R language was used to analyze the GO and KEGG enrichment pathway.

Intersect DEGs and SRGs to obtain the differential expression De-SRGs related to AS, and perform the GO and KEGG enrichment pathway analysis. Using STRING(http://string-db.org) to carry out protein-protein interaction network analysis, build the protein-protein interaction (PPI) network of De-SRGs, select the main PPI network with a comprehensive score greater than 0.4, and send the obtained PPI data to the Cytascape software for further analysis.

Using machine learning to identify core genes related to AS

In order to reduce the chance of De-SRGs, we use LASSO, SVM-RFE and Random forest methods to create De-SRGs models in turn, and select the cross genes of these three methods as core genes. The diagnostic ability of core gene was evaluated by receiver operating characteristic (ROC) curve of subjects and calculating the area under curve (AUC), accuracy, sensitivity and selectivity. Similarly, the ROC curve was used to evaluate the diagnostic ability of the regression model.

Evaluation of immune cell infiltration

Immunology cells, inflammatory cells, mesenchymal tissues, fibroblasts, and various cytokines and chemokines serve as the immunological microenvironment composition. Analysis of immune cell infiltration has a crucial driving function in understanding disease progression and therapeutic response. An extension of the GSEA methodology, the single sample gene set enrichment analysis (ssGSEA), was built using 23 immune gene sets. The immunologic features of all samples were evaluated using the ssGSEA technique using the “GSVA” R package.

Unsupervised cluster analysis of AS group samples for core genes

The relationship between core genes and all genes was analyzed. Based on the above core gene analysis, the R package “ConsensusClusterPlus” is used to perform unsupervised cluster typing of AS samples and calculate the expression difference of core genes between the two types. Differential gene analysis and GO and KEGG enrichment analysis were performed among subtypes. Download HALLMARK, KEGG pathway and Reactome pathway from Msigdb database, and use R packet GSVA to score the pathway. Compare the differences of pathways between subtypes and draw the heat map using R packet pheatmap.

Construction of miRNA-TF-mRNA regulatory network
The regnetwork database was used to predict the miRNA and transcription factors in the upstream of the gene, and the screening standard p value was \( \leq 0.05 \). After obtaining the miRNA-TF-mRNA regulatory relationship, use Cytoscape to visualize the miRNA-TF-mRNA regulatory network.

RESULT

Data processing and screening of DEGs

The data in the data set of GSE21545 and GSE21545 chips are converted into probes and gene names according to the platform annotation information. First, the R language preprocessCore package is used to homogenize the two data sets to eliminate the adverse effects caused by singular sample data (Fig. 1A,B). The two groups of data were combined according to the gene name. The combined data set included 158 AS samples and 32 control group samples. The R package "limma" and "sva" are used to remove the batch effect and reduce the differences between data sets due to experimental errors. The PCA results show that the two data sets have good fusion after merging (Fig. 1C,D). The difference analysis was carried out using the R language limma package, and the difference genes were screened according to the standard of adj.P.Val < 0.05, and 1821 genes were up-regulated and 1996 genes were down-regulated (Fig. 1E,F).

Functional and pathway enrichment analysis of DEGs

We conducted a GO and KEGG enrichment pathway analysis in R to investigate the possible role of these genes. The GO enrichment analysis of DEGs shows that the biological process (BP) of DEGs is mainly concentrated in the positive regulation of cytotoxic production, positive regulation of response to external stimulus, positive regulation of response to external positive regulation of response to external stimulus, positive regulation of cell adhesion, and molecular biological function (MF) is mainly concentrated in the active binding, GTPase regulator activity, Nucleoside-triphosphatase regulator activity, cellular components (CC) are mainly concentrated in cell-substrate junction, focal adhesion, cell leading edge (Fig. 2A). KEGG pathway analysis shows that Lipid and spherochlysis, Chemokine signaling pathway, Cell adhesion molecules, etc. are significantly enriched (Fig. 2B).

Screening De-SRGs and conducting enrichment analysis and PPI network

Intersect DEGs and SRGs to obtain AS-related differentially expressed SRGs (De-SRGs), as shown in Fig. 3A. A total of 36 De-SRGs were selected, including 21 up-regulated genes and 15 down-regulated genes. The GO enrichment analysis of 36 De-SRGs shows that BP is mainly enriched in protein and macroporous glycolation, glycoprotein metabolism and biosynthesis process, MF is mainly enriched in glycosyltransfer activity, carbohydrate binding, UDP-glycosyltransfer activity, CC is mainly enriched in vacuolar lumen, lysosomal lumen, Golgi apparatus subcompartment (Fig. 3B). KEGG pathway analysis showed that Mucin type O-glycan and other types of O-glyca biosynthesis, Lysosome, Amino sugar and Nucleotide sugar metabolism were significantly enriched (Fig. 3C). Input 36 De-SRGs into the online
STRING database, build the PPI network of De-SRGs, and import the PPI network data obtained from the main PPI network with a comprehensive score greater than 0.4 into the Cytascape software for visualization (Fig. 3D).

Using machine learning to identify core genes

To reduce the chance of De-SRGs, we use three different machine learning algorithms, LASSO, SVM-RFE and Random forest, to build prediction models. 36 De-SRGs selected 9 genes using LASSO algorithm, 22 genes were determined based on SVM-RFE, and 10 genes were selected by random forest according to importance (Fig. 4A-C). Five core genes (CD22, GALNT2, GCNT4, SIGLEC6, SELE) are obtained by intersection of candidate genes obtained from three different machine learning algorithm models (Fig. 4D). As shown in Fig. 4E, there are positive and negative correlations between core genes. Then, we use the "glm" software package to create a regression model based on five hub genes and draw the ROC curve. According to the ROC curve, the AUC of the regression model used to distinguish AS from the control groups was 0.797. In addition, we drew the ROC curve of five core genes for distinguishing AS from control groups, and the five core genes have high predictive value for AS (Fig. 4F).

Evaluation of immune cell infiltration

We used the ssGSEA function of the R language GSVA package to evaluate the proportion of immune cell infiltration, and compared the differences in immune cell infiltration between the AS samples and the control samples. The results showed that compared with the control group, all 21 types of immune cells in the AS group except for the CD 56 dim natural killer cell and the Type 2 T helper cell increased, with a statistically significant difference (P < 0.05) (Fig. 5A). And the overall expression of De-SRGs between the AS samples and the control samples is shown in Fig. 5B. At the same time, we compared the correlation between five core genes and immune cell infiltration, and analysis showed that CD22, SELE, and SIGLEC6 were positively correlated with most immune cells, while GALNT2 and GCNT4 were negatively correlated with all immune cells (Fig. 5C-G).

Consensus clustering analysis of sialoylation-related genes clusters

In order to clarify the expression pattern of sialylation related in AS patients, based on the five core genes, the R package "ConsensusClusterPlus" was used to perform unsupervised cluster typing of AS samples, and k = 2 was selected as the best value. 158 samples in AS group were divided into two different subtypes, A and B, of which 100 were type A and 58 were type B (Fig. 6A). Principal component analysis (PCA) showed that there were significant differences between the two types (Fig. 6B). The expression level of core genes in the two subtypes was visualized by thermogram and boxplot. Compared with type A, GALNT2 and GCNT4 in type B were significantly higher, while SELE in type A was higher than that in type B (Fig. 6C,D).

GSVA of biological pathways between subclusters of cuproptosis
Through GSVA analysis, several pathways with differential expression were enriched, and they were shown in a heatmap. Compared with cluster B, the Hallmark activities of apoptosis, IL6 jak stat3 signaling, P53 pathway, and IL2 stat5 signaling in type A increased, while the activities of kars signaling DN, myogenesis and apical junction were lower (Fig. 7A). In type A, the expression of KEGG pathways such as tprosine metabolism, vascular smooth muscle connection was significantly reduced, while the expression of related pathways such as toll like receptor signaling pathway, cell cycle and P53 signaling pathway was significantly higher (Fig. 7B). Based on the Reactome pathway, the results of GSVA showed that homologus DNA pairing and strand exchange, DNA strand elongation, tryptophan catabolism and other related pathways were significantly enriched in type A. Neurexins and neuroligins, neuronal system, physiological factors, etc. are enriched in type B (Fig. 7C).

**Analysis of functional differences between subtypes**

In order to further understand the functional differences between these two subtypes, we conducted differential expression analysis and used adj.P.Val < 0.05 to screen for differential genes. The results showed that: 287 downregulated and 116 upregulated DEGs were found between subtypes A and B (Fig. 8A). GO enrichment analysis of DEGs showed that BP, CC, and MF were enriched with response to molecular of bacterial origin and Leukocyte migration, collagen – containing extracellular matrix, and external side of plasma membrane, active binding, and cyclokine receptor binding, respectively (Fig. 8B). KEGG enrichment analysis revealed that these genes were mainly enriched in Cytokine – Cytokine receptor interaction, Viral protein interaction with Cytokine and Cytokine receptor, and Chemokine signaling pathway (Fig. 8C).

**Construction of miRNA-TF-mRNA regulatory network**

Using the regnetwork database to predict miRNAs and transcription factors upstream of genes, a total of 67 miRNAs and 21 TFs were predicted. The above results were imported into the Cytoscape to construct a miRNA-TF-mRNA regulatory network. As shown in the Fig. 9, there are one-to-many and many-to-one corresponding relationships between mRNA, miRNA, and TF, indicating that there may be more complex regulatory relationships between them.

**Discuss**

Cardiovascular disease (CVD) is the main disease endangering human health, and atherosclerosis (AS) is the basis of many cardiovascular diseases, which is very common in middle-aged and elderly people. AS has always been the main cause of death in the global aging population (17). Coronary heart disease, stroke, etc. caused by AS seriously affect people's quality of life. AS is associated with the inflammatory process of vascular disease and higher levels of low density lipoprotein (18). In the past decade, with the progress of laboratory research, people's understanding of the etiology and pathology of atherosclerosis has improved significantly, but many aspects of the disease development are still unclear (19). However, the pathogenic genes or molecular mechanisms of AS are still unclear, and in-depth research is needed to provide evidence for targeted therapy and early gene therapy for AS.
Sialoylation is a post translational modification controlled by the sialic acyltransferase, transporter, and neuraminidase family. This process plays a key biological role in maintaining cell cell interaction and is involved in many pathological conditions, such as cancer, embryo death, and immune system abnormalities (20). In previous large cohort studies, it was found that high serum sialic acid levels, both in protein binding and free forms, are independent risk factors for CVD (21, 22). However, there are few studies on the mechanism of sialoylation in the occurrence and development of AS. The results of this study provide a more comprehensive perspective for understanding molecular mechanisms and discovering new therapeutic targets in this field.

Using two datasets from the GEO database, we conducted a combined analysis. Through analysis of gene expression levels in the AS and control groups, we identified 3817 DEGs, including 1821 up-regulated genes and 1996 down-regulated genes. GO enrichment analysis shows that positive regulation of cytokine production, leukocyte cell-cell adhesion, positive regulation of cell adhesion, etc. are highly enriched, which is consistent with the mechanism of AS previously studied. In our study, through KEGG analysis, we found that DEGs participate in the Lipid and Atherosclerosis, Chemokine Signaling Pathway, and Cell Adhesion Molecules pathways, reflecting a partial mechanism of AS occurrence.

To further study the mechanism of sialylation in AS, we intersected DEGs and SRGs to obtain differential expression SRGs (De-SRGs) related to AS. A total of 36 De-SRGs were selected. According to GO enrichment, De-SRGs are highly enriched in lysosomes, glycosyltransferase activity, and glycosylation processes at all levels, consistent with previous studies on sialylation. KEGG pathway shows that Mucin type O-glycan biosynthesis, Lysosome, Amino sugar, and Nucleotide sugar metabolism play an important role.

To reduce the chance of De-SRGs, we used three different machine learning algorithms, LASSO, SVM-RFE, and Random forest, to build prediction models. Finally, five genes, CD22, GALNT2, GCNT4, SIGLEC6, and SELE, were defined as core genes. Currently, the recognized molecular mechanisms of AS mainly include abnormal accumulation of lipids, immune and inflammatory reactions, the transformation of monocytes into macrophages, and the proliferation and migration of smooth muscle cells. Immune response plays an important role in AS. Therefore, we compared the differences in immune cell infiltration between AS samples and control group samples. The results showed that compared with the control group, all 21 immune cells in AS group samples, except for CD56 dim natural killer cell and Type 2 T helper cell, increased, with a statistically significant difference (P < 0.05). Macrophages help maintain local inflammatory responses and can spread the development of plaque and promote thrombosis (23). The polarization of macrophages is crucial for the inflammatory response, as M1 macrophages initiate and maintain inflammation, while activated M2 macrophages resolve inflammation (24). A study shows that local and systemic B cells respond to a cholesterol rich diet, which indicates that the innate immune response in atherosclerosis (25) is activated. During the process of AS, the initial B cells will further differentiate into memory B cells and plasma cells, consistent with the increasing trend in the proportion of memory B cells in this study. The potential mechanism of mast cells on AS process can be attributed to: stimulating leukocyte migration, increasing vascular permeability, promoting lipoprotein accumulation,
foam cell formation and foam cell apoptosis, endothelial cell apoptosis leading to vascular wall erosion and thrombosis, promoting plaque instability and rupture, such as smooth muscle cell apoptosis, collagen decomposition, stimulating neovascularization and vulnerable microvascular plaque bleeding (26). The core gene CD22 in this study, also known as SIGLEC2, belongs to the sialic acid binding Ig-like lectin family (Siglec) with another core gene SIGLEC6. This study shows that CD22 is only negatively correlated with Type 2 T helper cells and positively correlated with other immune cells, while SIGLEC6 is positively correlated with all infiltrating immune cells. Siglec is a lectin subfamily that can specifically bind to the structure of sialic acid terminal glycans and is expressed on immune cells. It plays a role in promoting cell to cell interactions, mediating signal transduction between cells, and regulating immune cell responses. It can also mediate the recognition of pathogens and participate in the process of immune cell activation, proliferation, and apoptosis. The current research on the relationship between the Siglec family and arterial lesions mainly focuses on the role of Siglec-1, and it has been confirmed that inhibiting the expression of Siglec-1 using siRNA can significantly reduce the severity of AS lesions by reducing monocyte adhesion and macrophage recruitment (27).

GALNT2 is an n-acetylgalactosamine acyltransferase 2 that participates in the initial step of regulating glycosylation of mucoprotein. In various studies in human and animal models, this enzyme has been found to affect serum HDL-C and TG levels (28), and it also plays a regulatory role in adipogenesis and related cell phenotypes (29). Previous studies have shown that GALNT2 performs biological functions by activating the PI3K–Akt signaling pathway, which is consistent with our KEGG analysis of GALNT2. Our analysis shows that GCNT4 is mainly enriched in pathways such as cGMP–PKG signaling pathway, Calcium signaling pathway, and Regulation of actin cytoskeleton. GCNT is a member of the glucosamine transferase family and a key mediator for the synthesis, branching, and oligomerization of the core structure of mucins (30).

Many studies have confirmed that GCNT4 is closely related to the occurrence, development, and prognosis of tumors, but its relationship with AS and cardiovascular diseases has not been reported. Selectin, or CD62, is a glycoprotein responsible for heterogeneous interactions between cells under fluid dynamics. The selectin family consists of three members. The core gene SELE in this study is E-selectin, which is a cell adhesion molecule that is only expressed on various activated stimuli, and its expression is up-regulated in atherosclerosis (31–33). SELE mediates the rolling of leukocytes on the surface of EC, which is the first step to recruit these cells into atherosclerotic plaque (31). Inflammatory monocytes infiltrate the vascular wall, differentiate into macrophages, and aggravate the progress of inflammation and atherosclerosis (34). This study shows that SELE mainly plays a biological role through pathways such as Chemokine signaling pathway, Lipid, and Atherosclerosis.

Based on these five core genes, we used unsupervised cluster analysis to divide the AS samples into two subtypes, A and B. GSVA analysis showed that type A was mainly enriched in inflammatory response pathways, while type B was mostly enriched in metabolic pathways, which also showed that Sialylation was involved in the complex mechanism of AS occurrence and development.

This study also has some limitations. Firstly, due to the lack of experimental and clinical confirmation, our research results need further confirmation. The above results are based on comprehensive
bioinformatics analysis. More importantly, our data comes from a public database, and we lack raw sequencing data, so there may be selection bias. A carefully designed prospective study is needed to further validate our findings.

### Conclusion

Through a series of bioinformatics techniques, our work has revealed the relationship between Sialylation related genes and AS, as well as the heterogeneity of AS patients with different Sialylation subtypes. Choosing a Sialylation signature based on five genes as the best machine learning model can accurately evaluate the diagnosis of AS and control groups. Our findings reveal for the first time the involvement of Sialylation in the progression of AS and provide new insights into its potential pathogenic processes and treatment strategies.

### Declarations

**Ethics approval and consent to participate**

This study is an analysis of public database data and does not require ethical research in order to involve human and animal experiments. All participants in this study agreed to publish this article in the journal.

**Consent for publication**

The research data is sourced from public databases and does not involve any personal data.

**Competing interests**

All authors disclosed no relevant relationships.

**Funding**

This work was supported by the Shandong Natural Science Foundation General Project(NO. ZR202102180826) and Foundation Research Project of Tai’an Science and Technology Guidance Plan (NO. 2021NS159).

**Authors' contributions**

As the first author of the article, Wei Song participated in the entire process of designing, writing, and publishing the article. As co first authors, Jian-feng Xue and Ji-fang Liu are mainly responsible for extracting article data, selecting statistical methods, and writing articles. Yue Feng is mainly responsible for processing and analyzing differences between two sets of chip data. Na Yao and Lu-hua Yin participate in the immune infiltration analysis of Atherosclerosis tissue. Hong-Peng Yang, Hai-peng Sun and Pu Zhang, as co corresponding authors, coordinated the design of the entire experiment and reviewed the article.
Acknowledgements

This work was supported by grant from the Shandong Natural Science Foundation General Project(NO. ZR202102180826).

Availability of data and material

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

References


Figures
Figure 1

Data processing and DEGs screening. Use the R language preprocessCore package to homogenize the two data sets, before (A) and after (B). GEO dataset combination and differentially expressed genes. (C, D) PCA showed the batch effect between integrated datasets before and after
Figure 2

Enriched items in GO and KEGG analyses using DEGs. (A) Enriched items in GO analysis. (B) Enriched items in KEGG pathway analysis.
Figure 3

Screening De-SRGs and conducting enrichment analysis and PPI. (A) The overlap of genes between DEGs and SRGs. (B) Enriched items in GO analysis. (C) Enriched items in KEGG pathway analysis. (D) PPI network.
**Figure 4**

Using machine learning to identify core genes. (A–C) Construction of De-SRGs using LASSO regression, SVM, and RF algorithm. (D) The Venn diagram shows the overlap of candidate genes between the above three algorithms. (E) Circos plot displaying the relationship between the core
Figure 5

Evaluation of immune cell infiltration. (A) Correlation matrix of all 23 types of immune cell subtype
compositions. (B) Comparison of 23 immune cell subtypes between patients in AS and controls; The size and color of the circle represent the Pearson correlation coefficients. ns, no significance; * represents $p < 0.05$, ** represents $p < 0.01$; *** represents $p < 0.001$.(C-G) Correlation between immune infiltrating cells with CD22, GALNT2, GCNT4, SELE and SIGLEC6.
Figure 6

Identification of AS subtypes in sialylation. (A) Subclusters were performed with differential genes. (B) PCA diagram showing the distribution of different subclusters. Heatmap (C) and boxplot (D) show differential expression of core genes between subtypes. ns, no significance; ** represents \( p < 0.01 \); *** represents \( p < 0.001 \).
Figure 7

GSVA of key pathways between subtypes. (A) Enriched pathways based on the HALLMARK pathway. (B) Enriched pathways based on the KEGG pathway. (C) Enriched pathways based on the Reactome pathway.
Figure 8

Analysis of functional differences between subtypes. (A) The volcano plot of DEGs. (B) Enriched items in GO analysis. (C) Enriched items in KEGG pathway analysis.
Figure 9

MiRNA-TF-mRNA regulatory network